Geometry Driven Statistics

Geometry Driven Statistics

Edited by

Ian L. Dryden

University of Nottingham, UK

John T. Kent

University of Leeds, UK

WILEY

This edition first published 2015

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data applied for.

A catalogue record for this book is available from the British Library.

ISBN: 9781118866573

Typeset in 10/12pt TimesLTStd by SPi Global, Chennai, India

1 2015

Contents

Preface

Kanti Mardia is celebrates his 80th birthday on 3 April 2015. Kanti has been a dynamic force in statistics for over 50 years and shows no signs of slowing down yet. He has made major contributions to many areas of statistics including multivariate analysis, directional data analysis, frequentist inference, Bayesian inference, spatial and spatial-temporal modelling, shape analysis and more specific contributions to application areas such as geophysics, medicine, biology and more recently bioinformatics. A distinctive feature of Kanti's activities has been the annual series of LASR (Leeds Annual Research Statistics) workshops which he established and organized. These have helped to foster interdisciplinary advances in these research areas and have given rise to a long-standing series of proceedings containing short state-of-the-art papers published by Leeds University Press.

A common theme that unifies much of his work is the importance of geometry in statistics, hence the name of this volume, "Geometry Driven Statistics."

The research areas in which Kanti has worked continue to evolve and attract great interest and activity. It is, therefore, timely to provide a collection of papers from high-profile researchers summarizing the state of the art, giving some new developments and providing a vision for the future. Many of the authors have collaborated with Kanti at some stage in his career or know him personally.

To set the context for the later chapters, the book starts with some historical information on Kanti's life and work, together with a list of his main publications.

The papers have been split into four main topics, though of course there is considerable overlap and cross-fertilization between them:

- directional data analysis
- shape analysis
- spatial, image and multivariate analysis
- bioinformatics

The unifying theme throughout the book is geometry – with the first two topics specifically about statistics on manifolds. Directional data analysis involves the analysis of points on a circle (e.g., wind directions) or points on a sphere (e.g., location on the earth's surface), which are particularly simple non-linear manifolds. Kanti's 1972 book *Statistics of Directional Data* gave great visibility to the topic area and contained many novel developments, with a second edition *Directional Statistics* published in 2000 with Peter Jupp. Shape analysis involves the study of much more complicated manifolds, where the shape of an object involves removing information about location, rotation and scale. The topic has numerous applications including the study of organisms in biology or molecules in chemistry. Kanti's 1998 book *Statistical Shape Analysis*, jointly written with Ian Dryden, summarizes the statistical aspects of the field.

The third topic is particularly broad, involving data collected over geographic regions, image data or other high-dimensional multivariate data. An important classic book that is very relevant here is Kanti's 1979 book *Multivariate Analysis*, jointly written with John Kent and John Bibby. The final topic has been a particular focus for Kanti in the past decade, especially geometric topics such as Bayesian approaches to structural bioinformatics, where the shapes of proteins are key for determining function. Kanti's work in the area has been highlighted by his 2012 edited volume *Bayesian Methods in Structural Bioinformatics* with Jesper Ferkinghoff-Borg and Thomas Hamelryck. All four of the main themes are highly connected. Indeed several of the papers could easily have been placed within a different theme, which emphasizes an underlying unity behind the main ideas of this volume.

Ian L. Dryden and John T. Kent

List of Contributors

Norhashidah Awang
School of Mathematical Sciences, Universiti Sains Malaysia, Penang, Malaysia

Khandoker Shuvo Bakar
Department of Statistics, Yale University, New Haven, CT, USA

Stuart Barber
Department of Statistics, School of Mathematics, University of Leeds, Leeds, UK

Sandra Barragán
Department of Statistics and O.R., Universidad de Valladolid, Valladolid, Spain

Fred L. Bookstein
Department of Statistics, University of Washington, Seattle, WA, USA
Department of Anthropology, University of Vienna, Vienna, Austria

Wouter Boomsma
Department of Biology, University of Copenhagen, Copenhagen, Denmark

Clive E. Bowman
Mathematical Institute, University of Oxford, Oxford, UK

Sandy Burden
National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, New South Wales, Australia

Kai Cao
Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

Yasuko Chikuse
Faculty of Engineering, Kagawa University, Takamatsu, Kagawa, Japan

Noel Cressie
National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, New South Wales, Australia

Jesper Ferkinghoff-Borg
Biotech Research and Innovation Center, University of Copenhagen, Copenhagen, Denmark

Miguel A. Fernández
Department of Statistics and O.R., Universidad de Valladolid, Valladolid, Spain

Jesper Foldager
Department of Biology, University of Copenhagen, Copenhagen, Denmark

Jes Frellsen
Department of Engineering, University of Cambridge, Cambridge, UK

Riccardo Gatto
Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

Walter R. Gilks
Department of Statistics, School of Mathematics, University of Leeds, Leeds, UK

John C. Gower
Department of Mathematics and Statistics, The Open University, Milton Keynes, UK

Peter J. Green
School of Mathematics, University of Bristol, Bristol, UK
University of Technology, Sydney, New South Wales, Australia

Arief Gusnanto
Department of Statistics, School of Mathematics, University of Leeds, Leeds, UK

Thomas Hamelryck
Department of Biology, University of Copenhagen, Copenhagen, Denmark

John Haslett
School of Computer Science and Statistics, Trinity College, Dublin, Ireland

Anil K. Jain
Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

S. Rao Jammalamadaka
Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA

Peter E. Jupp
School of Mathematics and Statistics, University of St Andrews, St Andrews, UK

Wilfrid S. Kendall
Department of Statistics, University of Warwick, Coventry, UK

Nitis Mukhopadhyay
Department of Statistics, University of Connecticut, Storrs, CT, USA

Colleen Nooney
Department of Statistics, School of Mathematics, University of Leeds, Leeds, UK

Daniel E. Osborne
Department of Mathematics, Florida Agricultural and Mechanical University, Tallahassee, FL, USA

Vic Patrangenaru
Department of Statistics, Florida State University, Tallahassee, FL

Shyamal D. Peddada
National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

Orathai Polsen
Department of Applied Statistics, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Mingfei Qiu
Department of Statistics, Florida State University, Tallahassee, FL, USA

Cristina Rueda
Department of Statistics and O.R., Universidad de Valladolid, Valladolid, Spain

Sujit Kumar Sahu
Mathematical Sciences and S^{3}RI, University of Southampton, Southampton, UK

Anuj Srivastava
Department of Statistics, Florida State University, Tallahassee, FL, USA

Charles C. Taylor
Department of Statistics, University of Leeds, Leeds, UK

Douglas Theobald
Biochemistry Department, Brandeis University, Waltham, MA, USA

Hilary W. Thompson
School of Medicine, Division of Biostatistics, Louisiana State University, New Orleans, LA, USA

Qian Xie
Department of Statistics, Florida State University, Tallahassee, FL, USA

Zhengwu Zhang
Department of Statistics, Florida State University, Tallahassee, FL, USA

Part I

KANTI MARDIA

1

A Conversation with Kanti Mardia

Nitis Mukhopadhyay
Department of Statistics, University of Connecticut, Storrs, CT, USA

This paper originally appeared in *Statistical Science* 2002, Vol. 17, No. 1, 113–148.

Kantilal Vardichand Mardia was born on April 3, 1935, in Sirohi, Rajasthan, India. He earned his B.Sc. degree in mathematics from Ismail Yusuf College — University of Bombay, in 1955, M.Sc. degrees in statistics and in pure mathematics from University of Bombay in 1957 and University of Poona in 1961, respectively, and Ph.D. degrees in statistics from the University of Rajasthan and the University of Newcastle, respectively, in 1965 and 1967. For significant contributions in statistics, he was awarded a D.Sc. degree from the University of Newcastle in 1973. He started his career as an Assistant Lecturer in the Institute of Science, Bombay and went to Newcastle as a Commonwealth Scholar. After receiving the Ph.D. degree from Newcastle, he joined the University of Hull as a lecturer in statistics in 1967, later becoming a reader in statistics in 1971. He was appointed a Chair Professor in Applied Statistics at the University of Leeds in 1973 and was the Head of the Department of Statistics during 1976–1993, and again from 1997 to the present. Professor Mardia has made pioneering contributions in many areas of statistics including multivariate analysis, directional data analysis, shape analysis, and spatial statistics. He has been credited for path-breaking contributions in geostatistics, imaging, machine vision, tracking, and spatio-temporal modeling, to name a few. He was instrumental in the founding of the Center of Medical Imaging Research in Leeds and he holds the position of a joint director of this internationally eminent center. He has pushed hard in creating exchange programs between Leeds and other scholarly centers such as the University of Granada, Spain, and the Indian Statistical Institute, Calcutta. He has written several scholarly books and edited conference proceedings and other special volumes. But perhaps he is best known for

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

his books: *Multivariate Analysis* (coauthored with John Kent and John Bibby, 1979, Academic Press), *Statistics of Directional Data* (second edition with Peter Jupp, 1999, Wiley) and *Statistical Shape Analysis* (coauthored with Ian Dryden, 1998, Wiley). The conferences and workshops he has been organizing in Leeds for a number of years have had significant impacts on statistics and its interface with IT (information technology). He is dynamic and his sense of humor is unmistakable. He is a world traveler. Among other places, he has visited Princeton University, the University of Michigan, Harvard University, the University of Granada, Penn State and the University of Connecticut. He has given keynote addresses and invited lectures in international conferences on numerous occasions. He has been on the editorial board of statistical, as well as image related, journals including the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Journal of Environmental and Ecological Statistics*, *Journal of Statistical Planning and Inference* and *Journal of Applied Statistics*. He has been elected a Fellow of the American Statistical Association, a Fellow of the Institute of Mathematical Statistics, and a Fellow of the American Dermatoglyphic Association. He is also an elected member of the International Statistical Institute and a Senior Member of IEEE. Professor Mardia retired on September 30, 2000 to take a full-time post as Senior Research Professor at Leeds — a new position especially created for him.

In April, 1999, Professor Kanti V. Mardia was invited to the University of Connecticut as a short-term guest professor for four weeks. This conversation began on Monday, April 19, 1999 in Nitis Mukhopadhyay's office in the Department of Statistics, University of Connecticut, Storrs.

1.1 Family background

Mukhopadhyay: Kanti, shall we start at the origin, so to speak? Where were you born?

Mardia: I was born in Sirohi on April 3, 1935. Sirohi, was the capital of the Sirohi State about ten thousand square miles in area, in Rajasthan, before India's independence. Subsequently, the Sirohi State became the Sirohi district. Sirohi is situated about four hundred miles east of Bombay. One of the greatest wonders near my place of birth has been the hill station, Mount Abu. It has one of the finest Jain temples, Delwara, with gorgeous Indian architecture from the eleventh century. The exquisite details are all meticulously hand-curved on marble, without parallels anywhere else in India. Those shapes and formations on the ceiling and columns with intricate details influenced me even when I was small child. Much later in my life, some of those incredible shapes made deeper and more tangible impacts on my research career.

Mukhopadhyay: Please tell me about your parents.

Mardia: I come from a business family. My father's and mother's names are, respectively, Vardichand and Sanghari. My father inherited the business of moneylending from my grandfather and he had a pawnbroker's shop in Bombay. My grandfather started with practically nothing but through his business acumen acquired a large fortune.

But, my father had to live through some tragedies. He lost his father, two brothers and their families in the span of one year in an epidemic. Due to the spread of some severe unknown disease in that particular area, many in his family perished. My father, about sixteen, was practically the lone survivor in his family.

Figure 1.1 Kanti Mardia on his uncle's lap, Bombay, 1940.

Mukhopadhyay: How did this episode affect your father and the family?

Mardia: It had a devastating effect. My father started taking life very philosophically and decided to take everything easy. His whole perspective of life changed. He passed on the family businesses to my uncles. One uncle was a compulsive gambler who piled up huge debts. Eventually, many of the family businesses and other assets (e.g., several buildings and movie theaters in and around Bombay) were lost as loan payments on those debts. By the time I turned ten, our family had already slipped down from a very rich status and joined the upper middle class.

Mukhopadhyay: What about your mother's side of the family?

Mardia: My maternal grandfather was a lawyer and writer. He was an original thinker. He wrote a number of novels. Any writing skills I may have, I probably inherited from him.

Mukhopadhyay: How about your brothers and sisters?

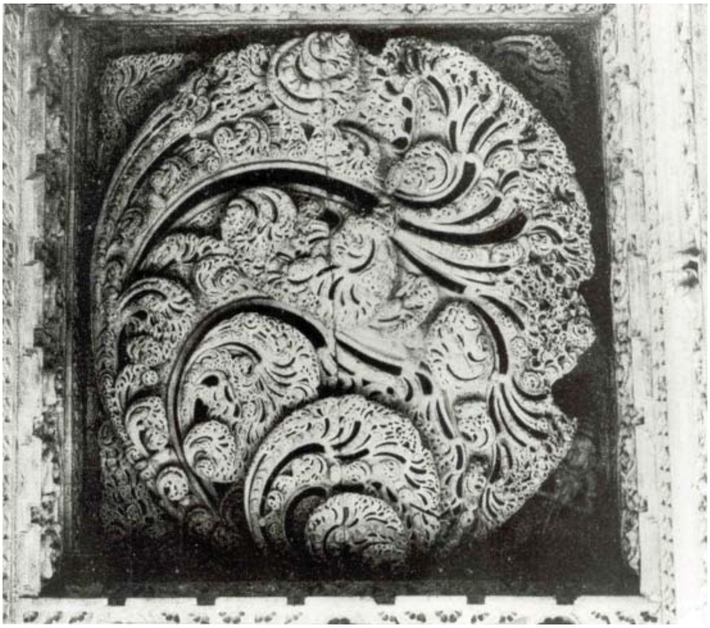


Figure 1.2 Ceiling from Jain Temple at Mount Abu (Rajasthan), Sirohi District. Original in white marble with tendrils circling in a fractal form, 1031 AD.

Mardia: I have four brothers and one sister. It is a large family. I am the one in the middle, a kind of the "median," a robust estimator. (Laughs.)

I became the first college graduate in the family. My brother Mangesh Kumarji looked after the family-run businesses. He earned real money to support the family while I had to study for my degrees! (Laughs.)

Mukhopadhyay: Was any of your siblings mathematically oriented?

Mardia: My younger brother Babu followed my footsteps and got a masters degree in pure mathematics. He is an Associate professor of Mathematics in Rajasthan University, Udaipur, India. During my childhood and school days, we lived in Sirohi as well as in Bombay, a major city center for all the businesses. We had to shuttle between these two places.

1.2 School days

Mukhopadhyay: Where and how did your schooling begin?

Mardia: In kindergarten, we learned numbers and even simple fractions. For example, at the age of four or five, we learned the concept of what is one-half of ten or one-quarter of eight! We had to memorize such multiplication tables and the teachers were very strict. We also had to learn to speak and write in Hindi, but this had to be mastered with the Rajasthani script and dialect, even though those styles were practically dead by then. It did feel like I was mastering a foreign language. This was on top of learning English.

Mukhopadhyay: Did you happen to have some inspiring teachers?

Mardia: In my time, there was only one high school in Sirohi, which I had to attend. Neither the teachers nor the curriculum had any flexibility and I did not like most of the subjects very much, except for mathematics. In the lower grades, we had a mathematics teacher who hailed from Ajmer, another part of Rajasthan, and he had an interesting habit. He used to assign challenging mathematical puzzles to the class and gave small prizes to whoever could solve the puzzles first. I was pretty good in solving such mathematical puzzles and won many prizes along the way. This math teacher had a big influence on me. I also enjoyed plane Euclidean geometry very much. I went through these constructions and proofs of theorems based on axioms. However, I have to confess that I preferred algebraic derivations and proofs with equations to the geometry-based arguments. (Laughs.)

Mukhopadhyay: (Laughs) Kanti, in quiet moments, sometimes you probably think what an irony of life that was!

Mardia: (Laughs) Nitis, you are right. Later in life, "geometry" became my mantra. What an irony indeed! I was not very interested nor considered particularly bright by others in nonmathematical subjects. I loved mathematics and sometimes I got into trouble because of this. Often I would come up with answers too quickly even for tricky problems. In higher grades, I became proficient in the factorization of quadratic equations, but some teachers did not appreciate that very much. Some teachers misjudged me, thinking that I was trying to show off or I was probably too clever. I was just being my enthusiastic self.

When I was about fourteen, I had to choose between the science stream or the arts stream. I did not care much about laboratory experiments and hence avoided pursuing the science stream. Instead I wanted to learn Sanskrit in the arts stream. So, I followed the arts stream.

Mukhopadhyay: Did you take the matriculation examinations from the same school?

Mardia: Yes, this was the only school in our area. I passed the matriculation examinations in 1951 and prepared for my transition to a college. But going to a college meant that I would have to migrate to another area and stay away from home, sweet home.

1.3 College life

Mukhopadhyay: Which college did you attend?

Mardia: The Jaswant College in Jodhpur (Rajasthan) was the closest to where we lived. I enrolled there for the two-year interscience degree. For someone like me who never took any science courses in school, there were not many choices of such interscience programs available in other universities or colleges. Jaswant College was about one hundred fifty miles from home. For the first time I stayed away from home. The hostel life was quite interesting.

I studied physics, chemistry and mathematics. I was terrible in the lab experiments (which we called the "practicals"). I dreaded the chemistry experiments with all those tubes and chemicals! I hardly had any clues! However, I used to enjoy the theories of physics, organic and inorganic chemistry, and the equations. But, when it came to lab experiments, I froze instantly. (Laughs.)

Mukhopadhyay: I can relate to this. I was quite weak in those chemistry practicals too. (Laughs) I assume that you fell in love with mathematics more.

Mardia: I really enjoyed learning the formative mathematics, for example, calculus, trigonometry, algebra, combinatorics. This was the first time I encountered the beauty of calculus. In the final examinations, I did well and came almost at the top of the graduating class. I got the I.Sc. degree in 1953 from Jaswant College which was affiliated with Rajasthan University.

Mukhopadhyay: What was special about those mathematics courses?

Mardia: The concepts of limits and derivatives were fascinating. I loved direct approaches through first principles rather than mechanically obtaining results. The principles and results from trigonometry were attractive. There was a book by S. Loney on this subject and I remember painstakingly solving every exercise from that book by myself. I really started enjoying the theoretical foundations and took studies very seriously. I surprised myself! (Laughs.)

1.4 Ismail Yusuf College — University of Bombay

Mukhopadhyay: I am sure that this helped in building your confidence. At this point, you were probably saying, “Look out University of Bombay, here I come!”.

Figure 1.3 Kanti Mardia in Jaswant College, Jodhpur, 1952.

Mardia: Nitis, you are correct. Earlier I was not qualified to attend the prestigious University of Bombay for the interscience degree. But now the door was open for me. By this time, most of our family had settled in and around Bombay, State of Maharashtra (earlier called the Bombay State). I was looking forward to attending the University of Bombay for the B.Sc. degree and at the same time I would stay close to the family. It was a great opportunity.

Mukhopadhyay: In the University of Bombay, what was your major?

Mardia: In 1953, I entered Ismail Yusuf College, a relatively small but prestigious college in a beautiful suburb, affiliated with the University of Bombay. I took mathematics as my major and physics as the subsidiary subject. I finished the physics requirement in the first year itself and thus in the final year I could concentrate only on mathematics. In physics, again there were those dreaded practical lab experiments, but I took care of them in my first year. What a relief it was for me! (Laughs.)

I found that my fellow students and others did not converse in Hindi. They all spoke in English. It felt like I was visiting another country altogether! I was used to writing in English, but I did not regularly speak in English. I slowly adapted, but first I had to get over a severe cultural shock.

It was an opportune time, though. The college had just started its math degree program. Professor Phadke, the Head of the Department of Mathematics, was an excellent teacher. Every student was required to have two elective papers, either in astronomy or statistics. I wanted to pursue astronomy, but the mathematics department had recently hired a young faculty member, Mr. Mehta, who graduated from the university with statistics. So I did not really have a choice. I ended up with both the elective papers in statistics — one paper on probability and another on inference and data analysis. It was a blessing in disguise!

Mukhopadhyay: In your course work, were you taught mostly from standard books or notes?

Mardia: Most advanced courses were taught by Professor Phadke. He was very smart and an excellent teacher. He came to each class fully prepared. He wrote clearly on the chalk board, taught interactively, and explained everything without looking at any notes or books. He was very impressive.

I was his favorite pupil in the class. After asking questions, he used to look at me and he fully expected me to come up with an answer. Sometimes I might not have been able to give answers as quickly as he expected and then I could sense that he was getting a little frustrated. One day he asked, "Where do two vertical planes intersect?" This question confused me. I could not feel the geometry at all. I was giving him two equations in three variables and then algebraically trying to find the common points. Naturally, this was taking some time. But, at that point the professor became very impatient because I was not seeing the answer that was obvious to him. He started explaining, "Look at this wall and that wall in the classroom. Where do they intersect?" As soon as he started drawing my attention to these vertical walls, it dawned on me that the answer was truly obvious. He wanted to hear some simple answer and I was throwing at him a couple of equations instead! (Laughs.)

I realize now why my teacher was getting restless. But I must tell you that Professor Phadke never meant any harm. He was refreshing and always challenging with very high expectations. I remain grateful to him forever.

Mukhopadhyay: Would you add anything about your other statistics professor, Mr. Mehta?

Mardia: Mr. Mehta was starting his own teaching career. He was young and very intelligent, and he focussed on doing everything right. He gave us excellent lecture notes. He did not have much experience and so he probably shied away from challenging us. He was very thorough and was available for extra help and guidance. We became very friendly. When I visit India, I make an attempt to go and see him.

1.5 University of Bombay

Mukhopadhyay: What influenced you to switch from mathematics to statistics?

Mardia: The top students normally opted for the engineering or the medical school. My parents wanted me to pursue that line too. I actually got admission to the engineering school of the prestigious Victoria Jubilee Technical Institute, Bombay. Someone hinted that if I went for a masters degree in statistics, I could become a fellow of the Ismail Yosuf College, a position which carried much honor and it also paid a stipend. In 1955, the subject of statistics was growing and so one could be very innovative. I heard from others that this discipline would offer a challenging future for bright people. The opportunities were plentiful as I understood.

I asked Mr. Mehta for advice and he said that "statistics" was the way to go. Each bit and piece of information convinced me that this route was more appealing than becoming an engineer or a doctor. My family had to be convinced that pursuing a two-year masters degree program in statistics would be more useful in the long run, and eventually they agreed.

The Department of Statistics at the University of Bombay was very highly regarded and it was quite special because this department was allowed to award its own masters degrees. Unlike in pure mathematics, there was tough competition to get admitted in statistics at the University of Bombay. I did not understand all the ramifications of what I was getting into. But because it was so hard to enter a program like that, I took it as a challenge and applied for admission during the first part of 1955.

Mukhopadhyay: What did you experience when you entered the masters program in statistics?

Mardia: Professor M. C. Chakrabarti, a great expert in the combinatorics and constructions of designs, was the Head of the Department of Statistics. I recall that he taught us probability. He was very methodical and an excellent teacher. Multivariate analysis and statistical inference were taught, respectively, by Professors A. M. Kshirsagar and Kamal Chanda (both of whom have been living in the United States for many years now). Professor K. S. Rao, an economist, was also on the faculty. All the professors were excellent. I am still in touch with Professor Kshirsagar.

There were very few textbooks and frequently we had to learn the materials directly from the journals. We often referred to *Biometrika*. In addition to the class notes and journal articles, I remember studying page-by-page from H. Cramér's *Mathematical Methods of Statistics* (1946) and C. R. Rao's *Advanced Statistical Methods in Biometric Research* (1952). Because of Professor Chakrabarti's eminence, we were taught a variety of materials on design of experiments for which we essentially relied upon W. G. Cochran and G. M. Cox's *Experimental Designs* (1950) and O. Kempthorne's *The Design and Analysis of Experiments* (1952). I was young and I came here with an impression that I was very good. But, once I landed in this department, it took me no time at all to realize that there were other intelligent people too! (Laughter.)

On a serious note, I immediately felt the challenging aspect of the teaching and research led to high expectations of the best and brightest students in the department. One had to be real sharp to survive such a level of tremendous pressure.

Mukhopadhyay: Who were some of your fellow students at the University?

Mardia: Babubhai Shah was my classmate. He was very bright. He has been at the Research Triangle Institute in Research Triangle Park at Raleigh, North Carolina. Jon N. K. Rao at Carleton University was one year senior to me. Jon was very sharp and popular. When we got stuck in a problem, sometimes we would ask him for advice. I remember that one time I was working on the distribution of the range in a random sample from some distribution and Jon Rao instantly came up with important suggestions on possible plans of attack. C. G. Khatri was two years ahead of me and unfortunately he is no longer alive. Kirti Shah at the University of Waterloo was one year junior to me. G. S. Maddala was my contemporary too. He was very clever and very good with the statistics practicals. I have lost touch with most of these friends. I have been able to keep in touch only with Babubhai and Jon Rao through all these years.

Mukhopadhyay: In the M.Sc. curriculum, which areas in statistics attracted you the most?

Mardia: Multivariate analysis and matrix algebra were definitely my two favorite subjects. The derivations were mostly algebraic, rather than geometric. I will say that statistical inference was the next in line. Wishart's (1928) original derivation of the Wishart Distribution fascinated me. At this stage, I was not exposed to the geometric approaches in statistics. I relied heavily upon algebraic and analytical derivations rather than the more intuitive geometric validations. Even now I do not have full faith in purely geometric "proofs."

Mukhopadhyay: Do you recall any aspect of the masters program that you did not enjoy much?

Mardia: I did not enjoy statistical calculations with the Facit machines. One had to turn the handle in one direction for addition/multiplication but in a opposite direction for subtraction/division. We depended on this machine for evaluating the square root of a number or for inverting a 4×4 matrix, and for that matter in all statistical calculations. During an exam, the whole room would be so noisy that it sounded like a factory. My usual problem was that if I repeated the steps to check any calculations, I very rarely got the same answer again! That was very frustrating. There was no way to be sure that the Facit machine's handle was turned in the right direction and the right number of times, particularly during an exam! I still remember that. (Laughs.)

Mukhopadhyay: Kanti, please excuse me. I cannot resist the urge to say this. It seems that you could not shake off the "ghosts" of the "practicals in physics and chemistry" that easily. You thought that you did, but the "ghosts" reappeared to haunt you with the disguise of Facit machines. (Laughs.)

Mardia: (Laughs.) You are right. I just could not get away from the so called "practicals," even in statistics! I always struggled with those Facit machines. You can only guess the relief and mental peace I derive from the personal computers I have.

Mukhopadhyay: Would you say that in the mid- to late 1950s, the statistical research program at the University of Bombay was in the forefront?

Mardia: Yes, the statistical research program at the University of Bombay was in the forefront. Professor M. C. Chakrabarti was internationally known and he was the star of the group. In 1956, there was a meeting of the International Statistical Institute in Calcutta and many notable personalities participated. On their way to or from Calcutta, some of the delegates came to Bombay to visit the Department of Statistics. I remember that Professor S. N. Roy came and gave a lecture on multivariate analysis. Professor Roy was wearing a typical Bengali attire, dhoti and punjabi. Everybody had so much regard for him that during his talk nobody said anything. Everyone listened intensely to whatever Professor Roy had to say. Professor Jerzy Neyman also came and gave a lecture on maximum likelihood and other estimators. I liked Professor Neyman's style of presentation very much. He raised issues regarding consistency, efficiency and so on by asking questions and then pointing out deep logical flaws in some of the obvious "answers." Such interactive exchanges with the audience continued without any notes while Professor Neyman paced up and down. His forceful seminar was so impressive.

Mukhopadhyay: Were there something like "student seminars" too?

Mardia: I remember that "linear programming" was not included in the masters curriculum. I started reading about linear programming and constrained optimization by myself. Later I gave a talk on this topic in the "student seminar" series. Senior masters students often took part in the "student seminars." These gave students important exposure and some good practice in talking in front of a audience and answering questions.

During this formative period I learned some important lessons: everything we read in print was not necessarily correct and I also understood that some results printed in books or research papers could be extended and sharpened. These realizations gave me the confidence and hope for future creative work.

Mukhopadhyay: The University of Poona is not far away from the University of Bombay. Did you see any interactions among the statisticians at these two sister institutions in the mid- to late 1950s?

Mardia: I do not recall any major interactions. I thought that the University of Bombay had the most reputable group of statisticians and they were the leaders in that geographical area. My memory has faded about the specifics of Poona's statistics program. The University of Bombay used to invite some external examiners from Poona, I am sure.

1.6 A taste of the real world

Mukhopadhyay: After receiving the M.Sc. degree in statistics in 1957, what was in store for you?

Mardia: Overall, the two years at the University of Bombay were great. I did not, however, do too well in the examinations. Again I partly botched the "practicals." When I graduated from the university, the State Bank of India was hiring people after screening through their highly competitive examinations. Many bright individuals sat in those exams with the hope that they would be selected. Some of my classmates ranked high enough in the examination and succeeded in getting jobs in the bank. I applied for a position too, but I was not selected! Now I may add that *fortunately* I was not selected! (Laughs.)

At that time, I did not aim for an academic career. Incidentally, I became very close to Professor Chakrabarti. I went to his house a number of times and he used to offer delicious Bengali munchies and snacks. Apart from the statistical discussions and help I got from visiting him at home, I admit that those delicious snacks were major attractions too. Babubhai Shah started working with Professor Chakrabarti on a Ph.D. thesis topic. My parents were hoping that I would take up a real job, earn a living and settle down in life. I was hesitant, but Professor Chakrabarti was advising me to pursue a Ph.D. degree in statistics.

Mukhopadhyay: You came in contact with Professor P. Masani. How did that happen and where did this connection lead you?

Mardia: I was not getting any job offers and I was already wondering about joining my family-run business. Professor Chakrabarti asked me to go and see Professor Masani, whose office was almost next door. He was Head of the Department of Mathematics. When I went to see him, he became excited and wrote me a letter offering a teaching position. He wanted an immediate reply. I was not too sure about an academic career at that point. But Professor Chakrabarti told me that if I ever wanted to pursue a Ph.D. degree or seek opportunities overseas, then I would be better off in the future if I accepted this offer from Masani. I decided to accept this one-year offer and started to teach.

Mukhopadhyay: This was a big break for you then. Any other recollections about those days?

Mardia: Professor Masani was very well known for his diligence and hard work. He would work day in and day out without letting up. He told me to get a solid foundation in mathematics including measure theory. I started learning the material from him. During my childhood, I had solved many challenging puzzles. When I grew up, I became more interested in finding what is in a “theorem” rather than proving the “theorem” itself. Professor Masani taught me proofs of very many deep theorems in measure theory, but I wondered about their inner meanings and beauty. The Institute of Science had connections with the prestigious Tata Institute of Fundamental Research (TIFR). Professor Pitt came from Nottingham University to visit TIFR and gave some lectures on measure theory which later shaped his book, *Integration, Measure and Probability* (1963). I attended those lectures very seriously but there was no fire. I did not get too excited and that puzzled Professor Masani.

In addition to my regular duties of teaching both mathematics and statistics, I also sometimes substituted in Professor Masani’s classes. Professor V. Mandrekar of East Lansing was doing his B.Sc. degree in mathematics in this Institute. In his first year, I had him as a student in my class.

Incidentally, you will recall that Babubhai Shah was in the other building in the university. He was already doing research in the design of experiments and I would regularly exchange ideas with him. For some time, I was interested in Pareto distributions and distributions of a range and other related problems. I also got some partial results. But I was not sure where my career was going.

1.7 Changes in the air

Mukhopadhyay: But you could sense that major changes were in the air, right?

Mardia: Yes, you are right. My family started getting impatient and wanted me to get married and get settled, and so on. I had been engaged since 1955 and that meant two years went by but I did not get married! Finally, I got married to Pavan in 1958 at the age of 23, which was considered “old” according to our custom. My younger brother also got engaged to be married on the same day so as to minimize his waiting time. (Laughs.)

I did not get any time to enjoy life very much. Immediately after I got married, I went back to the Institute and immersed myself in the studies of mathematics again. Professor Masani was planning to leave the Institute, probably in 1959, and go to the University of Pittsburgh. I seriously started to think about making a career move for myself. I applied for a position elsewhere. I vaguely recall that I got an opportunity to go to the University of Iowa, but for family reasons that did not materialize.

Mukhopadhyay: I guess that this was your period for job as well as soul searching.

Mardia: You are correct. I was looking for an opening to the right career path. Then I heard that Ruia College, another prestigious college affiliated with the University of

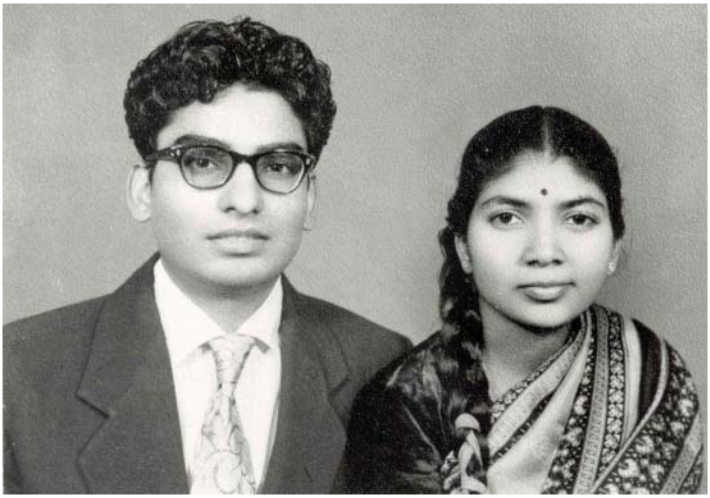


Figure 1.4 Kanti and Pavan Mardia's marriage photo, Bombay, 1958.

Bombay, was looking for someone to teach statistics courses. I applied for this position and got the job.

Our first child, Bela, a daughter, was born in 1959. At that point it became very clear to me that I would go overseas if and only if it would be financially feasible for my family to accompany me for the trip. I taught in Ruia College during 1959–1961. Unfortunately I do not recall the specifics from that period, but I do remember that the head of the department and other colleagues were kind and helpful to me. Also, I decided to improve my background in pure mathematics by earning externally an M.Sc. degree from Poona University in 1961, where I topped the list. I studied everything by myself for three months or so for the examinations.

With one baby at home and another one on its way the hustle and bustle of the city life of Bombay started to take its toll on both my wife, Pavan, and myself. We decided to move away from Bombay for some quiet and peace. Without a Ph.D. degree it seemed nearly impossible for a visit overseas. By this time, I had written a paper on multivariate Pareto distributions (Mardia 1962). I was gaining confidence and then the idea of seriously pursuing the Ph.D. degree crossed my mind.

1.8 University of Rajasthan

Mukhopadhyay: Did you make a career move then?

Mardia: In 1961, Rajasthan University in Jodhpur was starting a separate statistics department and they were looking for qualified teachers in statistics. Its close proximity to Jaipur, where I had spent the first part of my college life, made this opportunity very appealing. I moved to Rajasthan University to start their masters degree curriculum in statistics. There was another appointment (Dr. B. L. Sharma) junior to mine and we both taught at the masters level. I was more responsible for formulating the curriculum. The acting head of the department was Professor G. C. Patni, from the mathematics department. He suggested that I should pursue a Ph.D. degree, particularly because I already had some publications. In 1961, our son, Hemant, was born.

I registered under Professor G. C. Patni as a Ph.D. student and the research work that I was doing myself was progressing well. More students were enrolling in the courses I was teaching. Professor B. D. Tikkiwal, who was well known in sampling theory, joined the department a year later. He wanted me to work under his supervision. But I was not about to work in sampling, and then tension started to build. On the other hand, Professor Patni was always more than gracious and kind to me. I had both good and bad fortune. When I first arrived, I seriously thought that I was going to retire there. But quickly my views changed drastically. I again went into the transition mode and started looking around for a position abroad.

Mukhopadhyay: You were a junior faculty member and your life was miserable. How did you come out of this tight corner?

Mardia: The Commonwealth Scholarships became available in 1964. Professor Patni encouraged me to go abroad. The vice-chancellor of Rajasthan University was supportive of me. Because of their support, I applied and received one of the Commonwealth Scholarships. When I applied for leave without pay, I faced tremendous hurdles at the departmental level. Unfortunately, I could not persuade Professor Tikkiwal to help me this time.

1.9 Commonwealth scholarship to England

Mukhopadhyay: There was a period when your mind was set for overseas travel, but you had not yet left India. What was going on around that time?

Mardia: Before I left India on a Commonwealth Scholarship with my family, I submitted my first Ph.D. thesis to the Rajasthan University. At the time of my departure from India, that Ph.D. thesis was being examined by eminent external referees. I came to know much later that Professor Henry Daniel[s] from Birmingham University was one of the external examiners. My first Ph.D. degree came in 1965.

Mukhopadhyay: Where overseas were you heading as a Commonwealth Scholar?

Mardia: I left India with my family on September 13, 1964, on way to the University of Newcastle for a Ph.D. degree under the supervision of Professor Robin Plackett. He was well known for contributions in linear models and design of experiments. He was very knowledgeable in all aspects of statistics.

I was in a large group of Commonwealth Scholars from India in different subjects. The group was given a high profile reception upon arrival in London. The Mayor of London came to welcome the scholars. We went through a series of receptions and orientations lasting nearly ten days. Hemant was three and Bela was five. We stayed in a good hotel but there was no real facility for vegetarian meals. We were tired and waiting for the day to go and settle in Newcastle, the final destination.

Mukhopadhyay: I gather that you reached Newcastle after spending about two weeks in London. Did you adjust to the new surroundings and culture quickly?

Mardia: We got the culture shock of our life! We stayed temporarily with a host family arranged by the British Council. In this host family, the husband was Indian and the wife was English. Our children were hungry by the time we arrived at their residence. Fruits were on the table but these were refused to the children. Apparently, there were appropriate times to eat fruits! It was the wrong time to get hungry. I remember the incident vividly. This period was very trying.

Soon a representative from the British Council took us to a place where we could live more permanently as a family. Nearby, there was another Indian family, Ghura, who showed us around. They were very helpful. We immediately moved in and became very close to the landlord and his family. What a relief and joy it was to eventually find a place where we could buy Indian spices and groceries! I still remember the first homecooked meal in a foreign land after missing it for over three weeks. We lived in this one place for as long as we stayed in Newcastle. Subsequently, due to the children's schooling we came in contact with a much larger community.

1.10 University of Newcastle

Mukhopadhyay: In the University of Newcastle, which department did you join as a student?

Mardia: I went to the Department of Mathematics, which had a section on statistics. Professor Plackett, my assigned advisor, was Head of the Department of Mathematics. He was probably Editor of the Journal of the Royal Statistical Society, Series B, right around this period. He was a very busy man but he always had time for me. I became a full-time student all over again.

Mukhopadhyay: Did you think ahead about possible topics for a Ph.D. thesis?

Mardia: Professor Plackett and I were exchanging ideas. He had just finished a paper (Plackett 1965) where he formulated a bivariate family of contingency type distributions and he gave me a copy to study. I quickly realized that the same family could have been generated by quadratics having unique roots which led to interesting conclusions. This paper of mine appeared in *Biometrika* (Mardia 1967d).

Another problem which interested me all along was to find the joint distribution of two sample ranges obtained from bivariate random samples. I found simple expressions for the means, variances, and even the correlation coefficient between the sample ranges. The formula for the correlation coefficient was derived earlier by H. O. Hartley in *Biometrika* (1950) but my answer did not match with his and so I was puzzled. My paper (Mardia 1967a) was published in *Biometrika* where I wrote that Hartley's expression of the correlation coefficient was wrong! (Laughs.) Later, H. O. Hartley published (1968) a note with W. B. Smith, one of his students, showing that his formula was not wrong. It turned out that my approach was just simpler. (Laughs.)

Mukhopadhyay: So I suppose that no serious harm was done.

Mardia: (Laughs.) Right, no serious harm was done. Another work of mine that has survived all these years had to do with a nonparametric test for locations in a bivariate distribution (Mardia 1967c). This work was also done in the University of Newcastle.

Mukhopadhyay: Did you take any courses? How was the Ph.D. program structured?

Mardia: I did not have to go through any course work. I began exploring various research problems right away. Students were expected to attend regular colloquia. I remember that one time George Barnard came and gave a lecture. There was a symposium once where I presented a paper and I think that O. Barndorff-Nielsen was present.

Mukhopadhyay: You went to Newcastle with a wealth of knowledge about statistics. How did you proceed to learn new techniques and areas?

Mardia: In the beginning, my thesis topic was quite open. I had frequent discussions with Professor Plackett. He guided and exposed me to a broader horizon. It was the time when I started learning more things directly from the published papers. I was attracted by H. Chernoff's and E. L. Lehmann's nonparametric papers.

I kept researching by myself and I was totally independent. They had the KDF9 computer which ran on the language called ALGOL. I began having some difficulty working

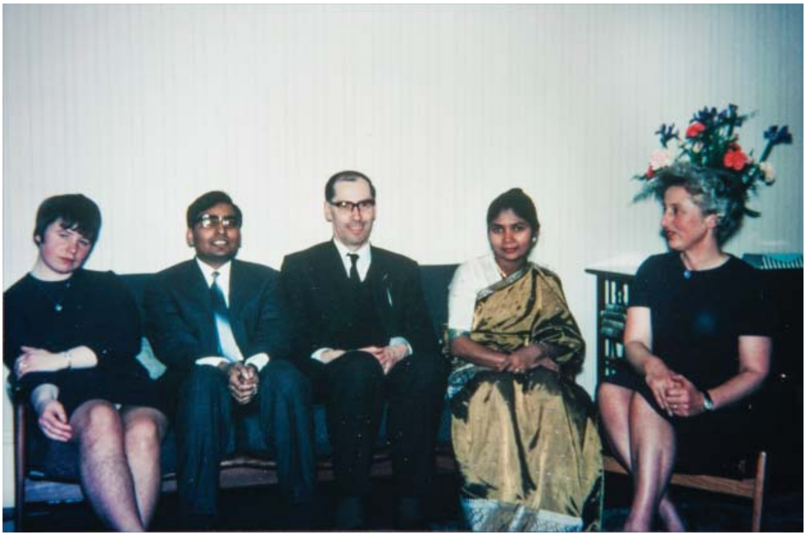


Figure 1.5 Left to right, Mrs. Brook, Kanti Mardia, Robin Plackett, Pavan Mardia, and Mrs. Plackett, at Newcastle, 1966.

with this machine. Professor Plackett was persistent that I must learn this language and eventually I became quite efficient in programming. I used computing tools extensively for my work in nonparametrics.

I also attended some of the Royal Statistical Society meetings. The invited papers with discussions always fascinated me. I heard some lectures of Vic Barnett and Toby Lewis on extremes. I commented (Mardia 1967b) on their paper, but I had to do so within five minutes of allotted time, something very new to me. In the middle of my comments, the bell started ringing. It was a very shaky but unique experience! (Laughs.)

Mukhopadhyay: Eventually what turned out to be your thesis topic in Newcastle?

Mardia: I already had two papers in *Biometrika* (Mardia 1967a; 1967d) and another two in *J. Roy. Statist. Soc. Ser. B* (Mardia 1967c; 1968). I finally wrote my thesis on “Some contributions to bivariate distributions and nonparametric methods.” This work was finished in approximately one and one-half years, but I did not know what to do after getting the Ph.D. degree and so I stayed on for a while. I passed my final thesis defense in January or February, 1967. Meanwhile, the two children were growing and my wife, Pavan, was pregnant with our third child. We had a baby girl, Neeta, in March, 1967.

For my Ph.D. thesis examination, the external examiner was Alan Stuart. This was the first time I met the “Stuart” of the famous “Kendall and Stuart.” He asked me pertinent questions and then kindly suggested how I might move ahead in different directions for further research. In the end, he remarked, “Two Ph.D. theses could have been made out of this one thesis.” I felt honored by the fact that this praise came from someone like Alan Stuart.

Mukhopadhyay: After finishing the Ph.D. degree, did you contemplate going back to India?

Mardia: I was thinking about this. But then Professor Tikkiwal from India hinted that when I returned to India after fulfilling the terms and conditions of the Commonwealth

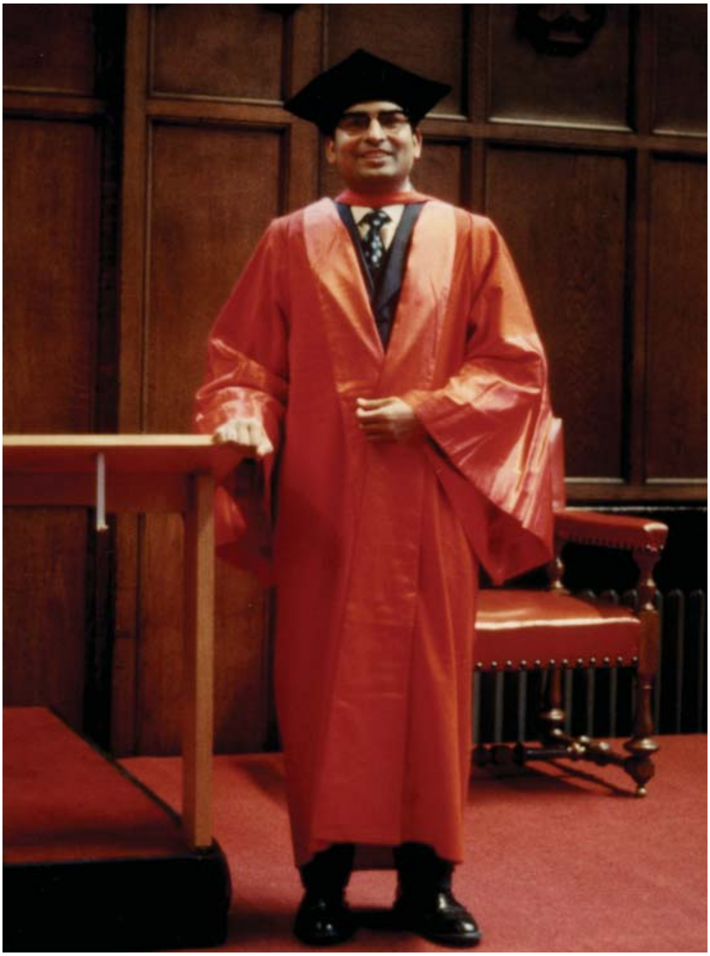


Figure 1.6 Kanti Mardia received the D.Sc. Degree, at Newcastle, 1973.

Scholarship, I would be transferred to teach in an undergraduate college. I felt unbelievable pressure building upon me from so far away!

Again, I heard the call for drastic changes in our lives. Some major decisions were hanging in the balance and I had to make a "statement." A lecturer's position became available in Newcastle and Robin advised me to apply. I went through the process, but the official waiver of my obligations to India arrived much too late, and hence I could not be offered a position. Robin asked me to withdraw my application and I followed his advice.

1.11 University of Hull

Mukhopadhyay: You then applied to the University of Leeds and what happened next?

Mardia: Once I got all the clearances from the Government of India, I applied for a position in the University of Leeds, probably in January or February, 1967. But, I was not selected. (Laughs.)

Meanwhile, I got an interview with the University of Hull for a lecturer's position. Hull is on the east coast of Britain, about sixty miles from Leeds. I liked everything in Hull. In April, 1967, I joined the statistics section in the Department of Applied Mathematics. They had two lecturers, Jim Thompson and Edward Evans. Jim worked with J. L. Hodges, Jr. and came from Berkeley. Edward worked on entropy but later switched to statistics. Subsequently, Michael Bingham, a student of K. R. Parthasarathy from Sheffield, was hired. This was a very good group.

Mukhopadhyay: I hope that your move to Hull was smooth.

Mardia: We bought a house straight away and arranged schools for the two older children. Our infant daughter Neeta came down with a bad strain of whooping cough and she was quarantined. The initial period was rough. After I had been a few days in the department, Professor Slater, the in-charge, asked me to describe the location of my house. I described the exact location and then Professor Slater said, “Kanti, would you believe! Your house is exactly opposite my house.” I thought to myself, “Oh God!” (Laughs.) One can surely guess that Professor Slater did not drive and frequently I gave him rides! (Laughs.)

Mukhopadhyay: What courses were you assigned to teach?

Mardia: I taught multivariate analysis to third-year students and had a very large class of second-year students. So I prepared my own lecture notes and adjusted the teaching style accordingly. I remember being asked to teach some traditionally unpopular courses, but those were extremely successful when I taught them. I also taught statistical inference to third-year students and some of my initial Ph.D. students came from this course. I was strengthening as well as teaching the department's course offerings. At the same time, my own research program started to flourish.

Mukhopadhyay: Did you then handle both the undergraduate and graduate students in Hull?

Mardia: The two systems in the United States and England are quite different. In England, one does not customarily go through a rigid course work in a Ph.D. program. One may opt to enter a Ph.D. program right after finishing an undergraduate degree. A third-year undergraduate in statistics learns through courses and substantial honors project, many modern aspects of statistical theory and applications. A student with such preparation and maturity is normally guided by a supervisor to explore research topics that may later develop into a Ph.D. thesis. This process may need about three to four years to culminate into a Ph.D. degree.

Mukhopadhyay: Kanti, I realize that you went to Hull as a lecturer with substantial experience. Were you happy?

Mardia: Not exactly, but I had no choice. I felt bothered mentally. I started looking for a more suitable position elsewhere in 1969. A senior position became available in Hull and Toby Lewis joined as Professor of Statistics, with the understanding that he could immediately hire a senior lecturer. I applied for the position. Obviously there was some competition but, in the end, I got the senior lectureship.

Mukhopadhyay: As you looked for a right position, did you ever consider moving away from England?

Mardia: The racial overtones and related flareups now and then in England bothered me greatly. I also wondered about the prospect of my eventually becoming a professor in England and worried that the chance was nearly zero. I could think of only K. R. Parthasarathy who became a professor in Sheffield.

I started looking for an opportunity to go abroad. In 1969, Madan Puri made arrangements to get me an offer from Indiana University, Bloomington, to become a nontenured associate professor. But, having heard horror stories about nontenured positions, I started negotiating with them and later decided that I was not about to go to Bloomington with my family with a nontenured job. That offer fell through.

Mukhopadhyay: So you stayed in Hull, I presume. What came next?

Mardia: The position of a reader is reserved only for good scholars. Monetarily this position is not very different but it has a lot of associated prestige. Each university in England has a unique system outlining the process of appointing readers. I was interviewed for the readership position in Hull with David R. Cox as the external and I became a Reader in 1971. I stayed in Hull through August 1973.

Mukhopadhyay: What were some of the research topics of your Ph.D. students in Hull?

Mardia: Barry Spurr worked on tests for multimodal axial circular distributions (Mardia and Spurr 1973). This developed nonparametric methods that later became a part of directional data analysis. Another student, T. W. Sutton, had worked on blocking problems in meteorology and regression analysis on a cylinder with temperature as a variable (Mardia and Sutton 1975). This work needed methodologies for some distributions with cylindrical variables and so this student developed both parametric and nonparametric methods for cylindrical distributions (Mardia and Sutton 1978). In the University of Hull, I essentially focused on guiding these two Ph.D. theses.

1.12 Book writing at the University of Hull

Mukhopadhyay: Kanti, you are well known for your books and edited volumes in a variety of areas. How and where did all these begin?

Mardia: The first thing I ever published that I could call my own was a short story written in Hindi for the college magazine in Bombay. The serious book writing started in Hull.

Recall that Alan Stuart was the external examiner for my Ph.D. thesis in Newcastle. He saw the great potential in my thesis area and mentioned that there was no book dedicated solely to that subject. Alan suggested that I should write a book on bivariate families of distributions. He said that his former student, Keith Ord, was writing the univariate part (Ord 1972). So, I immediately started writing the book, *Families of Bivariate Distributions*.

Alan was Editor of the Griffin's Statistical Monograph series and he urged me to finish the manuscript quickly. From time to time he would call and ask about my progress. The project was moving along very slowly. After a while, he said, "Kanti, look, there is no perfect book. I will tell you an anecdote which you should always remember. Harold Hotelling once had a contract with a publisher to write a book on multivariate analysis. He started writing some chapters and some years went by. At the end of each year, when the publisher inquired about the progress, Hotelling reported which chapters he was writing or revising and so on. During this time, C. R. Rao's biometric research book (1952) and T. W. Anderson's multivariate analysis book (1958) came out, and Hotelling felt that there was no more urgency for another book on multivariate analysis. Kanti, don't fall in such a trap." Alan said, "The moral is this: do not wait for someone else to write 'your' book in your subject!" (Laughs.)

I took Alan's advice very seriously. I moved on with this project, collected all the necessary materials quickly, and I completed the book in about one and a half years. My first book, *Families of Bivariate Distributions*, appeared in 1970 (Mardia 1970a). At the time when I wrote this book there was nothing else in the area. Then came the book of N. L. Johnson and S. Kotz (1972). Of course, the Johnson and Kotz series of books were superior.

1.13 Directional data analysis

Mukhopadhyay: How did you come upon the area of directional data analysis?

Mardia: In Newcastle, I began developing nonparametric methods by way of Hotelling's T^2 test. But, I was never too keen on working with ranks and asymptotics. In the latter part of 1964, I started thinking about some simple tests. I wanted to have a slick way of doing bivariate nonparametrics and not lose much power. I centered the two distributions, projected them on circles and worked with the uniform scores. Then I examined how these scores in the two populations were distributed. When I did this sort of thing fully in my thesis, I did not know anything about Geoff Watson's work on directional data. I did not even know what "directional data" was. Then Robin Plackett pointed out to me that there was a short note (Wheeler and Watson 1964) proposing a test that came to be known as the "Wheeler-Watson test." That paper came to my attention after I had submitted my thesis in Newcastle and my paper (Mardia 1967c) was published. It turned out that I had independently derived the Wheeler-Watson test.

Mukhopadhyay: Would you please explain briefly what this area is about?

Mardia: One may consider, for example, migrating birds and their homing directions. In this context, one may like to investigate whether there is a preferred direction or measure the variation from the homing direction, if any. Most navigational problems and many problems in astronomy involve measurements with directions. There were quite a few data analytic problems involving directions. Usual statistical entities such as the sample average and sample standard deviation are not so meaningful when observations are directions. One must take into account the geometrical structure and topology in order to arrive at appropriate analysis of such data.

Mukhopadhyay: Who were some of the major contributors in this field?

Mardia: Of course, R. A. Fisher did some early and fundamental work on the dispersion on a sphere (Fisher 1953). Geoff Watson was probably the next most important contributor to this field. His students (e.g., Michael Stephens and Rudy Beran) wrote theses in this area. Also, J. S. Rao wrote his thesis in 1969 on directional data at the Indian Statistical Institute, Calcutta, under C. R. Rao.

Mukhopadhyay: Did your directional data book originate in Hull?

Mardia: Yes, it did. As I was writing the bivariate distributions book, I felt that I got in me the bug of writing books. (Laughs.)

Substantial amount of material was available, but this material was all scattered. It was time to make a synthesis of the papers and dissertations and present this in a more accessible form. My research students and I were collecting these materials, and I thought that I already had enough for a book. Thus, the book on directional data was born.

Mukhopadhyay: How did you proceed?

Mardia: I wrote to Eugene Lukacs, an Editor for Academic Press for the series on probability and statistics, explaining my intent. Then, following Alan Stuart's valuable advice given to me before I wrote my first book, I immediately moved forward with the project with full steam. Toby Lewis was very supportive and he asked me how much time I needed to finish this book. He first approved a one-term sabbatical, followed by another, which were both immensely helpful for concentrating on book-writing. Some of the works on spheres

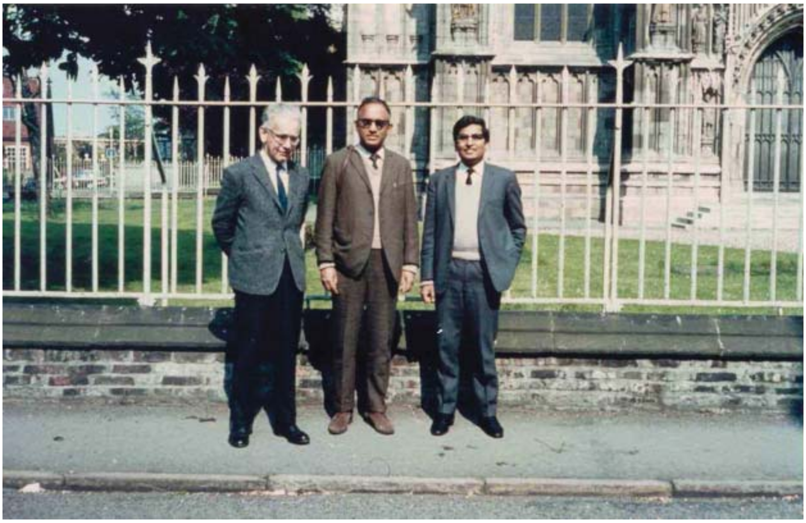


Figure 1.7 From left to right: E. Lukacs, D. Basu, K. Mardia, at Beverley Minster (near Hull), 1971.

were either incomplete or not very satisfactory, and so I started developing the needed material as I went along. By that time, my second sabbatical was gone and Toby suggested that I finish the book with whatever available material there was. Otherwise, the work could have dragged on much longer.

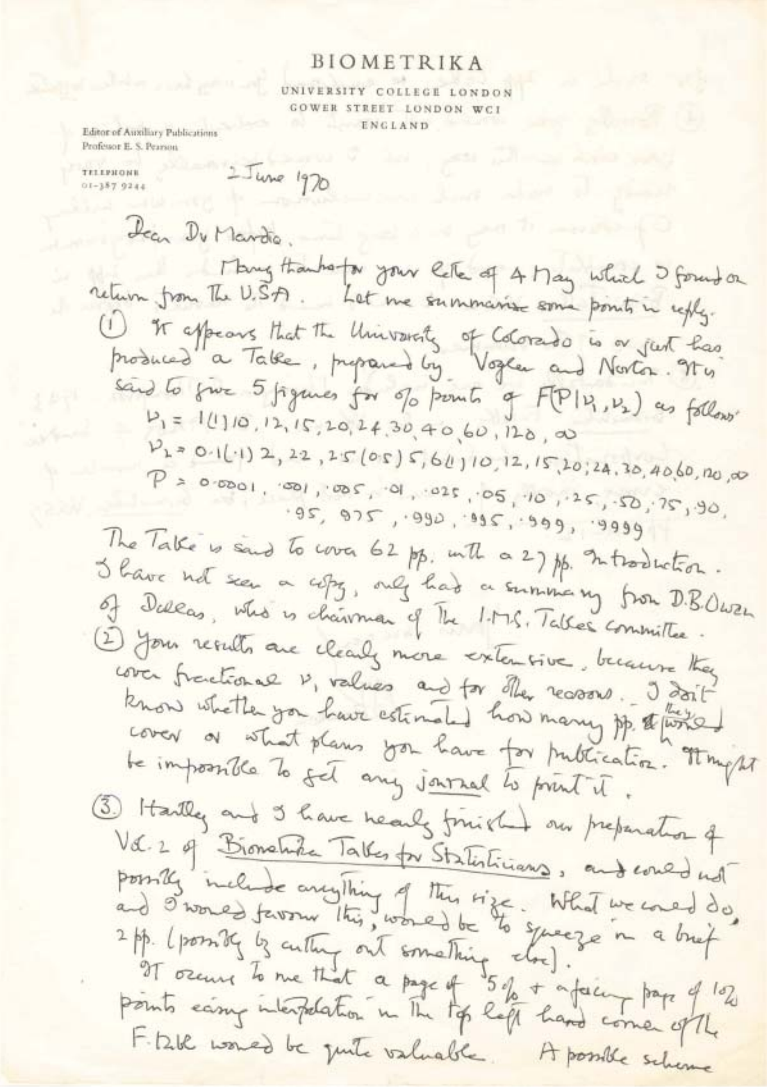
I finished writing the book in 1971. The *Statistics of Directional Data* was published in Mardia (1972) and it was an immediate hit. Geoff Watson (1973) wrote a very nice review of that book.

Mukhopadhyay: This book included a number of valuable tables. You produced several tables yourself. But you had to expend quite some energy to get permission to reproduce some of the other tables. Do you want to mention that story?

Mardia: This book needed many tables and I requested permission from Michael Stephens to reproduce some of the tables from his published works. He was hesitant because he was also writing a book in the same area. As many of those tables were from the journal *Biometrika*, I then approached its Editor, E. S. Pearson, for permission to reproduce the tables. Pearson said that Michael Stephens could be justified in being hesitant and he hinted that there could be a conflict of interest here because some of these tables were going to be included in the forthcoming E. S. Pearson and H. O. Hartley (1972) volume. He was not too sure that he should give me a "go ahead." I was kept in suspense while I waited with an almost finished book!

I had lot of correspondence with Pearson regarding my directional data book including many exchanges among Michael Stephens, Pearson and me regarding the copyright issues in reproducing some of the tables published earlier in *Biometrika*. Toby Lewis suggested that I go and see Pearson personally. I may add that I met E. S. Pearson only once, probably in 1970 or 1971, in his office in the University College, London.

When I saw Pearson, I sensed that he was not very comfortable with the whole episode and he was not happy about how the events turned out and became so complicated. He was a very kind person. By that time, I had become quite proficient with computers and I was preparing tables of the F-distributions with fractional degrees of freedom. So I dropped the hint that Pearson could include some of my F-tables with fractional degrees of freedom in the upcoming Pearson-Hartley volume. He then suggested that I should recalculate Stephens's tables as much as possible, but he would permit me to reproduce the difficult parts of his tables. My F-table was inserted on pages 171–174 of the Pearson-Hartley (1972) volume.



BIOMETRIKA

UNIVERSITY COLLEGE LONDON
GOWER STREET LONDON WC1
ENGLAND

Editor of Auxiliary Publications
Professor E. S. Pearson

TELEPHONE
01-387 9244

2 June 1970

Dear Dr Mardia,

Many thanks for your letter of 4 May which I found on return from the U.S.A. Let me summarise some points in reply.

(1) It appears that the University of Colorado is or just has produced a Table, prepared by Vogler and Norton. It is said to give 5 figures for % points of $F(P|\nu_1, \nu_2)$ as follows

$\nu_1 = 1(1)10, 12, 15, 20, 24, 30, 40, 60, 120, \infty$

$\nu_2 = 0{\cdot}1({\cdot}1)2, 2{\cdot}2, 2{\cdot}5(0{\cdot}5)5, 6(1)10, 12, 15, 20, 24, 30, 40, 60, 120, \infty$

$P = 0{\cdot}0001, {\cdot}001, {\cdot}005, {\cdot}01, {\cdot}025, {\cdot}05, {\cdot}10, {\cdot}25, {\cdot}50, {\cdot}75, {\cdot}90, {\cdot}95, 975, {\cdot}990, {\cdot}995, {\cdot}999, {\cdot}9999$

The Table is said to cover 62 pp. with a 27 pp. Introduction. I have not seen a copy, only had a summary from D.B. Owen of Dallas, who is chairman of the I.M.S. Tables Committee.

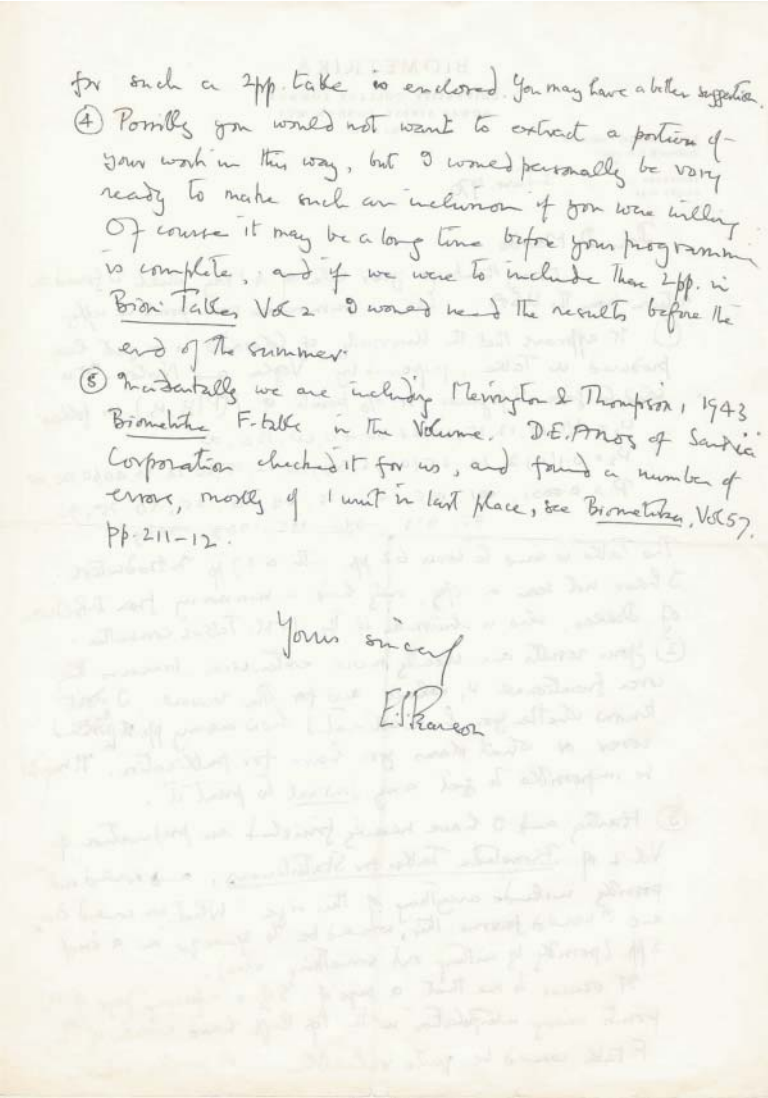
(2) Your results are clearly more extensive, because they cover fractional ν_1 values and for other reasons. I don't know whether you have estimated how many pp. ~~it~~ they would cover or what plans you have for publication. It might be impossible to get any journal to print it.

(3) Hartley and I have nearly finished our preparation of Vol. 2 of Biometrika Tables for Statisticians, and could not possibly include anything of this size. What we could do, and I would favour this, would be to squeeze in a brief 2 pp. (possibly by cutting out something else).

It occurs to me that a page of 5% + a facing page of 10% points easing interpolation in the top left hand corner of the F-table would be quite valuable. A possible scheme

(a)

Figure 1.8 (a) A sample of E. S. Pearson's letter on K. Mardia's F-tables with fractional degrees of freedom. (b) The last part of E. S. Pearson's letter.



for such a 2pp. table is enclosed. You may have a better suggestion.
④ Possibly you would not want to extract a portion of
your work in this way, but I would personally be very
ready to make such an inclusion if you were willing.
Of course it may be a long time before your programme
is complete, and if we were to include these 2pp. in
Biom. Tables Vol 2 I would need the results before the
end of the summer.
⑤ Incidentally we are including Merrington & Thompson, 1943
Biometrika F-table in the Volume. D.E. Amos of Sandia
Corporation checked it for us, and found a number of
errors, mostly of 1 unit in last place, see Biometrika, Vol 57.
pp. 211–12.

Yours sincerely

E S Pearson

(b)

Figure 1.8 (*Continued*)

Mukhopadhyay: It sounds like a very high level negotiation!

Mardia: It was understandable, but frustrating nonetheless. Again, Toby's advice came in so handy.

The interesting thing is that Nick I. Fisher, Toby Lewis and B. J. J. Embleton (1987) later wrote a book that dealt with the spherical data. I am very glad that they came up with their book, which included many details of the associated exploratory analysis. This book beautifully supplements what was lacking in my 1972 book.

Mukhopadhyay: What else was going on during this period?

Mardia: Meanwhile, the children were growing up. My wife, Pavan, already had a masters degree before we came to England, but she was not getting any suitable jobs. Pavan wanted to teach mathematics in a school. In 1969, she went through the three-year full-time certification program for education. All of us in the family had to endure a long and busy period. It was delightful when Pavan became a permanent school teacher in 1973, in Leeds. To her credit, she maintained the stability in the family through the whole ordeal.

A number of interesting people lived in the same neighborhood where we lived. One of them was Alan Plater, a very well-known playwright. The BBC often broadcast his plays. His son and my son, Hemant, were classmates in school. Phillip Larkin, a great poet, was the Chief Librarian at the University of Hull. Also, Sheldon, a novelist, was Larkin's second-in-command. These contacts with literary people were lots of fun for both my mind and soul.

1.14 Chair Professorship of Applied Statistics, University of Leeds

Mukhopadhyay: You were settled in at Hull. Why did you then decide to move?

Mardia: I liked Hull very much and I enjoyed doing what I did there. But, even so, for some time I was itching to become a full professor and worrying about my chances to hold such a position. One time Toby (Lewis) jokingly said, "Kanti, you don't move to the Sahara Desert simply because someone from there offers you a professorship. Take it easy and don't get so worked up. In time a position will come along anyway." Toby was correct and I did not even have to move to "Sahara." (Laughs.)

In the United Kingdom, some universities have a system which awards "personal" Chair positions and only exceptionally qualified individuals can be promoted to a Chair. The personal Chair Professors normally are not administrators, although there are some exceptions. The University of Hull did not have this system of personal Chairs. It was clear that I would have to move in order to become a full professor. In 1973, some of these positions were openly advertised and I applied. There were positions both in Salford, which is close to Manchester, and Leeds. I was offered a Chair Professorship at both Salford and Leeds. When I started my career in the United Kingdom, the University of Leeds did not offer me a junior position, and so I did not think twice! I decided to join the University of Leeds. (Laughs.)

But seriously speaking, there were important reasons to move to Leeds. Bernard Welch who had worked, among other things, on the Behrens-Fisher distribution of the two-sample statistic and robustness, was the Professor and Head at Leeds. This was a good department and I thought that I would never have to be the Head because there would always be two Chairs in the department. The Vice-Chancellor Lord Boyle, who interviewed me, had great sympathy and regard for Indian scholars and other minorities. He was a former Cabinet Minister of Education and held very broad ideals. In Hull, Toby Lewis was very supportive and he was one of my referees. My colleagues in Hull understood fully that this was a career move for me, and they all helped and supported me throughout the ordeal, for which I remain grateful.

The offer from Leeds came in May, 1973, and I joined in September of that same year. The position came with a personal secretary and a statistical assistant. A Ph.D. student, Dick

Figure 1.9 Kanti Mardia presenting a Discussion Paper at the Royal Statistical Society, London, 1975.

Gadsden, had followed me from Hull to Leeds. He worked with me on sequential methods for directional data. He is now a senior lecturer in Sheffield, in the same department as Gopal Kanji.

Mukhopadhyay: What were you doing when you first arrived in Leeds?

Mardia: I started with a statistical assistant who helped me with the computer programming. Peter Zemroch was my student and then he became my research assistant in Hull. He had also moved with me from Hull to Leeds. I had a grant from the Science Research Council to construct the tables for F- and related distributions. I already had a contract with Academic Press to write the multivariate analysis book. I was very busy with research problems in directional data, as I was preparing a paper to be read at the Royal Statistical Society meeting in 1975.

Mukhopadhyay: Why were you so involved with the F-tables?

Mardia: I was fitting univariate distributions using the first four moments. This exercise needed F-tables with fractional degrees of freedom.

In Hull, I got the idea of writing on multivariate skewness and kurtosis for testing multinormality. This was conceived via multivariate linear model and permutation tests. The second moment in a permutation test depends on the multivariate kurtosis. I gave most of the details in my *Biometrika* paper (Mardia 1970b). Then I got down to the F- or beta distributions and I needed extensive sets of F-tables for checking the goodness of the fitted distributions. Peter Zemroch developed the computer programs in ALGOL60 language and eventually Peter and I published a monograph, *F-Tables and Related Algorithms* (Mardia and Zemroch 1978), which has since been translated into Russian. Peter continued working on algorithms for directional data for about three years and we published some joint papers.

Mukhopadhyay: What else was going on during those initial years in Leeds?

Mardia: After we moved to Leeds, Pavan got a job in a school right away. Our son received scholarships to attend a prestigious private school. Our daughters were progressing beautifully. In the family front, everything felt just right for a change.

In the department, I was given the responsibility for the masters program. To energize the curriculum, I introduced new courses. Apart from adding a course on directional data, I pushed for more vocational courses. Geostatistics and statistical computing courses were added around 1975. A set of new computers arrived in 1977 to modernize the computing environment.

I started the statistical consulting component to foster collaborative research with scientists from other disciplines. The routine consulting requests were passed on to the postgraduate students and they learned what real statistics was all about. Substantial consulting projects were shared by colleagues for broadening the scope of research in other fields and also for preparing grant applications.

Bernard Welch and I overlapped for about three years while he was preparing to retire in 1976. He lost interest in the day-to-day administration of the department. He was, however, still teaching. Outside of statistics, one of his main interests was the game of cricket. He often said, "I recommend retirement to do other things full-time."

I was brought to Leeds with the charge to energize the teaching, research and consulting programs. I started doing just that with vigor and vision for the future, I hope. I got the support I needed from my colleagues and the higher administration alike.

Mukhopadhyay: What was the administrative structure and who were some of your colleagues in Leeds?

Mardia: The statistics department was, and still continues to be, one of the three autonomous departments within the School of Mathematics. The school had a chairman and these three departments had respective department heads.

Apart from Bernard Welch, we had Harry Trickett who was a senior lecturer. He did some research in statistics, but his strength was in administration and teaching. Harold Peers had worked on invariance. I had also other colleagues. We had people working on, for example, distribution theory and time series analysis.

Mukhopadhyay: Between a department head and the Chairman of the School, who is more powerful? Where did you fit in this bigger picture?

Mardia: The department heads are traditionally more powerful. The role of the Chairman of the School is to coordinate its total program and services. If conflicts or duplication of programs or services arise among the departments, the school chairman then intervenes to mediate and guide all parties to a common ground for the benefit of the school. A department head is responsible for running the department, whereas the school chairman acts as a liaison.

When I arrived in Leeds, I found a wonderful administrative structure. I did not have to worry at all about the undergraduate administration. The school had a Director of Undergraduate Teaching who looked after all courses and related matters in the three departments. Each department was, however, responsible for formulating its own curriculum requirements, develop teaching modules, update future planning, and so on. I introduced tutorials with smaller groups of students and added modern course materials; for example, we created an exploratory data analysis course as a requirement for the third-year students.

Mukhopadhyay: In the mid-seventies, the university had to endure serious financial hardship and I am sure that your department had to streamline its priorities. How did you "reposition" yourself?

Mardia: There was a period in 1976 when circumstances changed and finances became hard to come by. I had to become the department head, quite reluctantly, to lead the group of ten statistics faculty members. Subsequently, I was allowed to hire new faculty members.

In the meantime, I received a large Symposium Grant. With this grant, I could invite short-term visitors from abroad to Leeds for collaborating on projects related to directional data. During that period, C. G. Khatri came to Leeds when my joint works with him started. Subsequently, O. Barndorff-Nielsen, Rudy Beran, Kit Bingham, Tom Downs, John Kent and J. S. Rao came to visit Leeds. John Kent was a graduate student of David Kendall and later he joined our department as a Lecturer. Ian Dryden also joined the department subsequently. It turned into a wonderful period to move ahead in the areas of directional data and non-Euclidean geometry in statistics.

Mukhopadhyay: I understand that in Hull, your multivariate analysis book was also conceived. When you moved to Leeds, work on that book continued too. What do you recall?

Mardia: The work on my multivariate book was continuing. When I was in Hull, Toby Lewis pointed out that John Bibby from St. Andrews was writing a book on the same topic, and I took John Bibby as a coauthor. But slowly I came to understand that his style was very different. I first rewrote and verified everything he used to send. Then I took John Kent as another coauthor to make real progress. The book *Multivariate Analysis*, jointly authored with J. T. Kent and J. M. Bibby, appeared in 1979, much later than it should have (Mardia et al. 1979).

Mukhopadhyay: At some point, you went to Canada for a semester to try out a tenured position. Obviously you did not stay there. Would you care to comment?

Mardia: It was 1977. Racial tensions in the United Kingdom were on the rise. Many politicians and other people were giving negative speeches. My wife, Pavan, said “Let us get out of this country before it becomes too late for us.” We were seriously debating whether we should permanently move away from the United Kingdom and around that time I received an offer from the University of Windsor, Canada, for a tenured position. I thought that I should try out this change of venue for a semester. In January, 1978, I arrived in Windsor, Canada, by myself with a leave of absence from Leeds.

This was the coldest winter I had ever faced. Because it was the middle of the school year, my family could not join me. In Windsor, I was given a substantial teaching load. I was asked to teach a very elementary course with two hundred students! I had never taught any class nearly as large as this one. The departmental environment was very good and I liked my new colleagues. John Atkinson, who was the department head, and Dick Tracy were both very helpful. My family joined me in Windsor when they had the Easter break in the United Kingdom but the overall systems and cultures in the two countries were very different. My family did not take to it and deep down I also did not. It might have been different if my family had had more time to spend in Windsor or perhaps if I had not had to teach this heterogeneous and huge class as soon as I arrived. In May, 1978, I returned to Leeds. (Laughs.)

1.15 Leeds annual workshops and conferences

Mukhopadhyay: You created the tradition of annual workshops in Leeds. You should feel genuine pride and satisfaction when you look back. Any highlights to share?

Mardia: If one wants to expose interested colleagues to a new subject, it works well to invite an expert in that area and learn the subject from the lectures. For example, when I was writing the multivariate analysis book, I realized that the multidimensional scaling and

Procrustes methods will have significant impacts on the area. John Gower (1975) had just published some work on Procrustes methods. I knew him from Hull, where I had invited him to a symposium on multivariate analysis which I had organized in 1973. I invited John to Leeds for detailed lectures on multidimensional scaling and Procrustes methods. A couple of other faculty members and I went through that workshop very diligently. Such workshops are now integral parts of the statistics department in Leeds. In subsequent workshops, I invited other scholars, including Julian Besag and Brian Ripley, because we felt that a lot of activity was imminent in spatial statistics and in image analysis. These gatherings are now internationally recognized as the Leeds Annual Statistical Research (LASR) workshops. We have an open-door policy. Anyone interested should feel free to participate at any time.

Also in 1985, Subha-Rao visited us and gave lectures on aspects of time series analysis that had direct bearing on spatial statistics. As early as 1979, we organized a conference in geostatistics. This was a workshop, but it was also open to several invited speakers along the lines of a conference. We had an invited speaker from France, A. Marechal (Centre de Geostatistique, Fountainbleau), from the G. Matheron group. This conference was quite a success.

In 1984, I organized a workshop on image analysis. This was the first time such a workshop had been held in a statistics department anywhere. Researchers from all over the world had showed a lot of interest in this workshop. One of the workshops on shape analysis was attended by David Kendall; Fred Bookstein also participated many times in these important workshops on shapes.

Our forthcoming workshop (the eighteenth one) will address spatio-temporal modeling with emphasis on applications. The applications will include tracking in machine vision, functional MRI in medical imaging and ecology. Some well-known statisticians (e.g., David R. Cox, Peter Diggle and Noel Cressie) have been invited. But there will also be experts in functional MRI, epidemiology, tracking and ecology. From Oxford, Andrew Blake (now with Microsoft, Cambridge) will participate. He is an expert on tracking. These

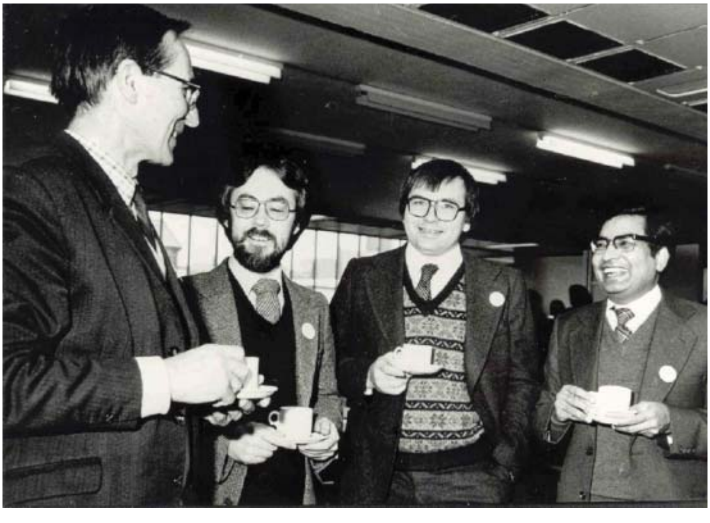


Figure 1.10 From left to right: D. G. Kendall, P. A. Dowd, A. Marechal, K. V. Mardia. Geostatistics Meeting, Leeds, 1979.

annual workshops try to build a bridge of communication and collaboration among experts in statistics and other substantive scientific fields where fresh ideas and methodologies are urgently needed.

Mukhopadhyay: Do you normally edit and publish the proceedings from these workshops and conferences? The proceedings can reach a much wider audience.

Mardia: In the beginning, we did not publish the proceedings. But, subsequently we started publishing these Proceedings starting in 1995 to reach a wider audience than the limited number of participants. These have been well received by the scientific community in general.

Mukhopadhyay: In your view, what have been your two best accomplishments in Leeds?

Mardia: I think that our department is recognized internationally. Our research program has been in the forefront of fundamental breakthroughs in information technology (IT). The department is undeniably on the map. I would like to think that I have helped in creating and strengthening this visibility. This has been the most important accomplishment. It has been such a gratifying journey for me.

Another major accomplishment, if I may say so, has been the modernization of our course offerings, including computer-aided teaching utilizing the latest available statistical software packages. A long time ago, in teaching our courses, we implemented our own software, even before the well-known software, MINITAB, came on the market.

Mukhopadhyay: Now please describe the worst episode during your tenure in Leeds in the sense that you would happily change your course of action right now if you could turn the clock backward?

Mardia: Very early on, the Science Research Council (SRC, now EPSRC) threatened our masters program. The SRC came up with a new policy to cut the number of courses in order to streamline programs across the university. I fought against this decision and made an appeal to the higher administration. All our M.Sc. courses were then reinstated.

Figure 1.11 From left to right: F. L. Bookstein, W. S. Kendall, K. V. Mardia, J. T. Kent, C. R. Goodall. The 15th LASR Workshop, Leeds, 1995.

Figure 1.12 Workshop in action at home in Leeds. Playing bowls. From left to right: D. Terzopoulos, J. Koenderink, E. Berry, K. Mardia, L. Marcus.

But after three years of some calm and quiet, the SRC went back to the drawing board and decided to drop the M.Sc. program in statistics. Our Ph.D. program was hurt because our M.Sc. program fed students into our Ph.D. program and now that channel was eliminated! Difficult financial situations were knocking at the door. We lost some regular faculty members and some were replaced by temporary positions in order to cut costs. We continue having an active but smaller Ph.D. program. If I could turn the clock backward, I would definitely fight more to save the M.Sc. program.

Mukhopadhyay: Statistical methodologies have certainly changed focus over the years. In your view, where are we heading?

Mardia: During the periods of R. A. Fisher and P. C. Mahalanobis, statistics brought revolutions with path-breaking applications in the areas, for example, of agriculture, biology and sampling, with great impact on population census and economic planning. In the past ten or fifteen years, new statistical ideas and methodologies have energized IT which is a general name to describe subject areas such as computer vision, image analysis and machine learning. In my department, we have a large group of internationally recognized experts in these and related fields. Fundamental challenges in data handling in IT have enriched the field of statistics tremendously. My feeling is that this change in emphasis and directions will continue in the foreseeable future. In Leeds, we have been preparing for such changes for quite some time.

1.16 High profile research areas

Mukhopadhyay: Kanti, in your opinion, what are your primary areas of research expertise?

Mardia: Broadly speaking, the major thrust areas include multivariate analysis, directional data, shape analysis, spatial statistics and spatial-temporal modeling. Another big area, which goes hand in hand with these can be categorized as applications involving imaging, machine vision and so on.

1.16.1 Multivariate analysis

Mukhopadhyay: Please highlight some of your important contributions in multivariate analysis.

Mardia: Classical multivariate analysis heavily depended upon the multivariate normality assumption of the parent population. I developed methods for checking multivariate normality (Mardia 1970b) by introducing multivariate analogs of skewness and kurtosis and gave measures to quantify departures from normality. The impact of this paper has lasted more than that of some other papers of mine. When others come up with newer measures of multivariate skewness and kurtosis, they compare performances with earlier measures given in Mardia (1970b).

Mukhopadhyay: What do you suggest users do when multivariate normality is suspect?

Mardia: Unfortunately, I have not addressed that aspect. I would expect one to use multivariate Box-Cox transformations as a possible route. But it is not always an easy task to accomplish. One may alternatively use permutation tests for the mean or the location parameter. I wrote a related paper (Mardia 1971) describing the effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. In that paper I gave permutation tests which may provide plausible solutions.

Figure 1.13 From left to right: C. R. Rao, B. Mandelbrot, K. V. Mardia, at Penn State, 1994.

Mukhopadhyay: What other kinds of multivariate problems attracted your attention?

Mardia: I have enjoyed creating new and interesting multivariate distributions and deriving some of their important properties. For example, I elaborated Plackett's family of bivariate distributions.

I worked on multidimensional scaling. The subject of multidimensional scaling helps one to come up with similarity measures. When I examine configurations, I can come up with numerical measures which will in turn tell us how similar or dissimilar these configurations are.

My 1977 work on how the singularity of the variance-covariance matrix Σ affects inference techniques involving the Mahalanobis distance has also been well cited in the literature. In Mardia (1977), I had defined what is known as the Mahalanobis Angle.

Perhaps the most important contribution was my Mardia et al. (1979) multivariate analysis book, jointly written with John Kent and John Bibby which we talked about earlier. This book has met the test of time. My feeling is that a good book should last at least ten years. Some of my books, for example, the Families of Bivariate Distributions, have not been of this caliber. (Laughs.)

Mukhopadhyay: Since the early days of Fisher, Mahalanobis, Hotelling, Hsu, Roy, Bose, Rao and others, multivariate analysis has come a long way. Where will this field take us next?

Mardia: My best guess is that the field will become more exploratory and data oriented. There will be more emphasis on statistical modeling, for example through elliptic distributions, and in nonparametric or semiparametric models. There was a time when distributions were discarded as models if there were no analytical expressions for the maximum likelihood estimators of the parameters. That scenario has changed for the better. Model checking has become more a visual art than anything else. With easy accessibility to computers, statisticians are now driven more by the complexity of the problems rather than opting for a narrow set of "nice distributions" for analytical reasons alone.

1.16.2 Directional data

Mukhopadhyay: Now we move to directional data. Please highlight some of your important contributions.

Mardia: There were quite a few important problems involving directions on which I had the opportunity to work. I studied, for example, flying patterns of migratory birds (Mardia 1975), various problems in geology, analysis of megalithic-yard data in archeology (Mardia and Holmes 1980), and in astronomy the behavior of long-period comets (Jupp and Mardia 1979).

John Kent and I had worked with Jim Briden, an earth scientist, on the formation of the earth, its various layers, continents and their movement patterns over time. The data reduces to the directions of the prevailing magnetic field of the earth, but sudden movements of some layers may change the course. Using the natural remnant magnetization in rocks, our work shows how to find out where a continent was located when a particular rock was formed and involves identification of linear segments given a set of ordered points. Our paper (Mardia et al. 1983) is highly regarded.

Once the magnetic components have been extracted, the objective is to follow the movement of the continents over geologic time — that is, the apparent wander paths. This problem was investigated with Dick Gadsden (Mardia and Gadsden 1977). Also, a related problem is the movement of the area of vulcanism or hot spots. As the plates move, a chain of hot spots is assumed to be formed on the earth's surface. Both can be viewed as following points along the arc of a small circle on the earth's surface and thereby determining (fitting) that circle. We looked at the actual data for validation of the theoretical model. Further distributional work was developed with C. Bingham (Bingham and Mardia 1978).

One time I worked with a physicist, Professor Alan Watson of Leeds, on high energy particles. It was believed that these particles could have arrived on earth from one of two possible galaxies. The question was whether these particles came from one single source.

The points where these particles hit the earth may be thought of as a cap on a sphere. Jointly with Rob Edwards, I came up with an appropriate distribution and analyzed the observational data (Mardia and Edwards 1982). I understand that the physicist's postulates have since been modified.

Another project was on central place theory. Suppose that a town grows in a regular fashion. Then, using the principles of Delaunay's triangle (Dryden and Mardia 1998) for the set of sites of the towns, one should claim that these triangles should be equilateral. With Madan Puri and one of my students (Robert Edwards), we developed a statistical test (Mardia et al. 1978) to check whether the triangles are equilateral. This work is again often cited and, in a way, inspired some shape work by others later on. We found the distribution of shapes of the equilateral triangles assuming that they were independent. But they were not really independent! More works followed later, including those by other researchers. If one looks at wind directions at two time points, they will be naturally correlated. The analysis of such data had led to another collaboration with Madan Puri (Mardia and Puri 1978).

Mukhopadhyay: Your Mardia (1972) book, *Statistics of Directional Data*, was certainly major work.

Mardia: The directional data book has been a success. The field really took off after this book was out. My Mardia (1975) discussion paper, read at the Royal Statistical Society, also created much enthusiasm among researchers in this field.

Mukhopadhyay: Preparations for its second edition have gone on for years. Will it be out soon?

Mardia: P. R. Halmos once wrote that one should never go for a second edition of a book. But, Halmos himself published second editions of some of his works! (Laughs.)

I was hesitant to prepare a second edition of my book. Peter Jupp, who had a background in pure mathematics and differential geometry, later worked with me as a postdoc around 1976–78. He is now a Reader in St. Andrews. Peter and I have completely rewritten and updated the material. What one will find is a new book, *Directional Statistics*, which is expected to be out soon (Mardia and Jupp 1999). The rewritten book took us close to eight years to finish!

1.16.3 Shape analysis

Mukhopadhyay: Kanti, how did you get into the area of shape analysis?

Mardia: I have been fascinated by shapes, being brought up in the midst of the famous Jain temples with intricate marble carving (see Figure 1.1 of Mount Abu). I always wondered "How were these shapes generated? Are the replications accurate?"

Another exposure has been in childhood through palmists who would make claims based on various features of palms, for example, palm shapes. Apparently, there are seven basic types of palm shapes. This intrigued me — why are there seven?

Mukhopadhyay: Actually, you have been intrigued by palmistry for a very long time. Would you care to explain?

Mardia: When I was a small child, I was brought up with the expectation that I would pursue the family-run business, but a palm reader looked at my palm and forecasted that I would end up going abroad for higher studies. Notions like "higher studies" or "going abroad" were not even on the horizon. I have no idea how this palm reader could forecast my fate! (Laughter.)

Mukhopadhyay: Regardless of how the forecast was made, it turned out to be accurate. Where is that palm reader now? (Laughter.)

Mardia: I have also asked myself, "Where is that palm reader now?" Perhaps he should be invited to come and give some lectures in one of my workshops! That will constitute practical statistics on shapes! (Laughter.)

In fact, in 1980 I collected some preclassified palm shapes in the literature and constructed various landmarks to characterize the palm outline. I obtained what are now called Bookstein shape variables. It turned out that there is quite a large overlap between the shapes. I think that Ian Dryden also used this data in the initial period of his Ph.D. work.

Indeed, whenever there are claims related to "palmistry," I try to get involved! There was an article in the *J. Roy. Soc. Medicine* (1990) by a medical doctor (Dr. P. G. Newrick) and his collaborators in the United Kingdom claiming that longevity depends on the length of what is called a life line. I got the data and analyzed it, but found that even the life line was not well defined (Mardia 1991).

Scientific studies of ridge-patterns of the hand are important to detect genetic disorders and malformation. The field of scientific studies of such patterns is called dermatoglyphics. Now, there are various known features which are used to describe ridge-patterns. In the 1960s, L. S. Penrose — a Galton Professor — proposed a number of feature variables which are in use. I wrote a joint paper (Mardia and Constable 1992) characterizing a unique special feature. We also provided software. Our proposed theory along with its computer algorithm have enhanced automatic recognition of fingerprints in forensic investigations (Mardia et al. 1997a). The software is slow but it does provide a unified statistical approach!

Mukhopadhyay: How and when did the actual transition to shape analysis take place?

Mardia: Fred Bookstein's paper "Size and shape spaces for landmark data in two dimensions," appeared in *Statistical Science* (1986) and I thought that this was the kind of paper that I had been waiting for quite some time. Fred's paper showed me the light.

The following year, I believe, Ian Dryden came from Nottingham to Leeds to work on his Ph.D. with me. We started working on a joint project with an anatomist who had a problem which had originated from experimental breeding with mice. The anatomist started with big (heavy weight), average, and small (low weight) mice, and then let the breeding process go through some generations within the weight groups. One question was whether the shapes of mice, within a weight group, remained the same across generations. The anatomists were comparing shapes of the vertebrae of mice in each group. This is how I got into this area which gave me the impetus to start a brand new career, so to speak.

I should add that the subjects of shapes and directions are closely related. Ideas of constructing distributions are similar. In shapes, what was lacking was that there was no analogous "normal" distribution to work with. The question we faced was whether there could be an exponential family of distributions for shapes. In a series of papers, Ian Dryden and I considered the marginalization approach by integrating out the nuisance parameters (Dryden and Mardia 1991; Mardia and Dryden 1989a,b).

Mukhopadhyay: These investigations eventually led to the distributions which are known in the literature as Mardia-Dryden distributions.

Mardia: I presented the distribution in my discussion (Mardia 1989) of David Kendall's (1989) *Statistical Science* paper. The distribution I proposed was clearly too simple for a problem that had seemed so intractable for some time! David liked the distribution but he did not believe the answer at first. Later, David validated the distribution using a stochastic formulation.

I worked as a catalyst. I got several people interested in a new and interesting subject. In turn, I was able to accomplish new results too, both theoretical and applied in nature.

Mukhopadhyay: You have not yet mentioned your recent book, *Statistical Shape Analysis* (1998), coauthored with Ian Dryden.

Mardia: Ian Dryden has been my colleague in Leeds since 1989. The book was finally launched as I was giving the special invited lecture on the same subject at the 1998 Joint Statistical Meetings in Dallas, Texas. I advised Wiley to print a large number of copies of the book. They did not take a statistician's forecast too seriously! (Laughs.)

It turned out that they ran out of copies within four months of publication. The reprinted version came out in April, 1999.

In shape analysis, there are many unsolved theoretical aspects. A shape or an image looks different when viewed from different angles or subspaces. If one rotates the axes or stretches or squeezes the axes, the basic characteristics of a regular shape should be preserved. For example, I considered projective shape space and the associated distributions. A theory paper, jointly written with C. Goodall and A. N. Walder, has appeared as (Mardia et al. 1996). Walder was formerly my student and then he became a postdoctoral fellow. One of the fundamental challenges in the field of computer vision is to enable computers to "see," that is, to emulate human vision, and these projective invariants allow object recognition. This latter piece is a joint work with Colin Goodall which included machine vision applications (Goodall and Mardia 1999).

1.16.4 Spatial statistics

Mukhopadhyay: Another of your major interest is spatial statistics. How did this interest arise for you?

Mardia: Early on, I became interested in kriging within geostatistics and spatial statistics. I taught an M.Sc. course on geostatistics as early as 1978. I was charmed by the methodologies. Given some of the coordinates in a space, I was thrilled to learn how practical models were built with the help of variograms and covariograms. One could fit a surface, if nothing else was feasible.

I had a grant on geostatistics on which R. Marshall and I worked (Mardia and Marshall 1984) to develop a spatial linear model under normality where the errors were correlated. The parameters were estimated by the maximum likelihood method. But there were some crucial difficulties. We had one realization from a stochastic process. We were not too sure whether we should proceed with asymptotics by increasing the size of the grid or by making the grid more dense. In other words, we were unsure whether we should "fill in" or "fill out!" Eventually, I thought that "fill out" should be the way to go because then the information would steadily increase. We could obtain asymptotic results with complicated looking criteria under which the distributions of the parameter estimates were multivariate normal. One of my students, Alan Watkins, worked on multimodality, bias and other criteria in related spatial problems (Mardia and Watkins 1989).

I am pleased to say that my (Mardia and Marshall 1984) *Biometrika* paper with Marshall is highly regarded in geostatistics. Here, we modified some of the classical ideas to come up with appropriate linear models in the new area of spatial statistics. Because of the general acceptance and popularity of linear models in statistics, I believe, our approach to spatial statistics with linear models has caught on rapidly.

Mukhopadhyay: Subsequently, you became more involved with research related to kriging.

Mardia: Yes, you are correct. John Kent's interest also turned to kriging. Fred Bookstein was working on comparing images. This involved comparing landmarks of two or more averages in the space. Suppose that we consider one plane in the (x, y) coordinate system and another one in the (u, v) coordinate system. The question may be whether these two planes are similar. If they are similar, then one should be obtainable by the identity mapping from the (x, y) to the (u, v) systems. But, if the two planes are not similar, then one may try to find the corresponding mapping of the plane in the (x, y) system to the one in the (u, v) system, and examine how deformed or stressed this mapping is if compared with the identity map. Fred did some important work (Bookstein 1989) using thin-plate splines. One should bear in mind that this kind of mapping should not depend upon rotation or other similarity transformations of the shapes under consideration. Fred used thin-plate splines with linear terms which "kill" affine transformations. So one sees local shape changes.

This approach of Fred Bookstein helped to identify one kind of deformation. But in many applications, deformation may arise from a larger class which consists of kriging. Here, self-similar processes provide the necessary background. Then one may not only ask questions about the landmarks, but the tangent directions may also be included in our considerations. Fred had handled (Bookstein 1996) this more general situation. This aspect of spatial analysis has a bright future.

Mukhopadhyay: I am quite certain that you have other ongoing book projects as we speak.

Mardia: I actually started writing another book on spatial analysis. When I introduced and taught the M.Sc. course on geostatistics, perhaps as early as 1978, I prepared my own lecture notes. I was using those lecture notes instead of any book. In the meantime, Brian Ripley's *Statistical Inference for Spatial Processes* came out in (1981). Now I was in the same boat that Hotelling had been in his aspiration to write a multivariate analysis book! Even though I had a contract with a publisher to write this spatial statistics book, I could not see what purpose such a book would serve, particularly because Ripley had just written an excellent book on the same subject but also covering spatial point patterns. However, I lately have been actively writing the spatial analysis book with John Kent as my coauthor.

Mukhopadhyay: Sometimes, you have used Bayesian analyses. Are you a Bayesian now?

Mardia: Personally, I am a very pragmatic Bayesian. If there is prior information available, I tend to use it, especially in situations where there is no readily available better technique. I started relying upon Bayesian techniques when I began working on image and shape analysis. I do not see any practical value of using Bayesian techniques indiscriminately.

1.16.5 Applied research

Mukhopadhyay: Methodologies you have been vigorously pursuing in spatial statistics, directional data and shape analysis are clearly in the cutting edge of statistical computing. Any thoughts?

Mardia: The images in general are very large and therefore techniques are developed which can use local contextual information. This is more so in low level image analysis where the aim is segmentation. Hence the use of Markov random fields as priors has come into use. On the other hand, for high level image analysis, such as in object recognition, structural information of objects in priors (e.g., deformable templates) reduces computational complexity to some extent.

In 1979, I attended a conference on geology in Paris where Paul Switzer gave a talk using some Landsat and showed how he had classified types of rocks. It was delightful. I requested the data and he kindly gave it to me. The pixels had very low resolutions, perhaps 5 × 5 km, I vaguely recall. The rock types were overlapped. At that time most statisticians had not even heard about pixels. (Laughs.)

I got some work done and submitted the paper to the *J. Roy. Statist. Soc. Ser. C*, but the referee did not like my work. I got an impression that the referee thought this approach of mine would go nowhere. But I felt otherwise. Soon, Paul Switzer presented a related paper (Switzer 1983) at the International Statistical Institute meeting in 1983 where Julian Besag was one of the discussants. Besag (1986) later pursued the iterated conditional mode approach. Geman and Geman (1984) on the other hand took the statistical computations to another level by exploiting the ideas of stochastic relaxation and Gibbs distributions. Now these are labeled Markov chain Monte Carlo (MCMC) methods.

I started working more vigorously on low level image analysis. Then I published two papers in *IEEE Trans. on Pattern Analysis and Machine Intelligence* (Mardia and Hainsworth 1988; Mardia and Kent 1988) on image segmentation and spatial classification, respectively. The work with John Kent developed spatial classification using fuzzy membership models. Some of these methods are robust and fast.

Mukhopadhyay: Suppose that a criminal has been on the run for five years. The investigating agencies try to reconstruct a “recent” photo of this criminal based on his file photos which are between five and fifteen years old. The theory and practice behind any such reconstruction fall right in your alley, I am sure.

Mardia: You are absolutely right. Modeling image warping is an area I have worked on. There are many difficult problems here. For example, suppose we have four photos of someone’s face at different ages. What kind of image should be called an “average” of these four photos? How different are the four photos from the so-called average image? These are important, interesting and challenging problems. For researchers in machine vision, the problems of identification and tracking are crucial. Success in this area of research depends heavily on one’s expertise with the methodologies of spatial statistics, shape analysis and computing.

Mukhopadhyay: I understand that spatial and spatio-temporal modeling are important for environmental monitoring too.

Mardia: Indeed, spatial and spatio-temporal modeling are *essential* for environmental monitoring. For example, what should be the location of the next monitoring station? There is no quick-fix answer for this. Such a question can be addressed with the help of a complex interplay between spatial and spatio-temporal modeling. I had worked with Colin Goodall on some spatio-temporal analysis of multivariate environmental monitoring data (Mardia and Goodall 1993), and the results were presented at the 1993 Multivariate Conference held at Pennsylvania State University. I read related papers at the 1994 Biometric Society meeting in Hamilton, Canada (Goodall and Mardia 1994) and in the University of Granada, Spain, in 1996. A discussion paper on the Kriged-Kalman filter was read at the Spanish Statistical and Operations Research Society meeting in November, 1998. This paper marries the two prediction approaches, kriging for space and Kalman filter for time (Mardia et al. 1998). In July, 1999, a workshop was arranged: Spatial-Temporal Modeling and its Applications in Leeds. In the coming years, I expect a lot of activity in these exciting areas.

Mukhopadhyay: Automatic classifiers are used in harvesting and packaging. A robot does the work, but what can you say about the behind the scene modeling which creates the "brain?"

Mardia: In automatic harvesting of mushrooms, for example, how does one design a robot which will pick only the good mushrooms of a certain size? The problem may appear very simple on the surface, but the mathematics and the implementation of the model behind the algorithms are both far from trivial. In Mardia et al. (1997b), an appropriate Bayesian methodology was developed. Such techniques have a good future in general.

Mukhopadhyay: Would you mention one or two upcoming papers with important applications?

Mardia: Right now I am writing a paper with Fred Bookstein and another one with John Kent. Both papers have to do with bilateral symmetry. In some individuals, one half of the face does not look the same as the other half because one half of the face is distorted. This phenomenon is called hemifacial microsomia and can be corrected only by surgery. We are collaborating with a surgeon, Jim Moss, and a physicist, Alf Linney, at University College, London. The common practice is to take laser scans of the face both before and after the surgery. But how should one go about comparing the before-and-after pictures of the face? How should one compare the images of two brains, one normal and another schizophrenic? This is not a routine matter. Many scientists from different fields are working on these types of problems. Some of my recent work with Fred and John falls in this area.

Mukhopadhyay: How do you get ideas? How do you know which ideas to pursue?

Mardia: Most of my ideas are data driven. Somebody may give me a set of data or it may come out of our consultancy or collaboration. I enjoy looking at data inside out and try to understand the hidden message it has for me. The data gives clues in every turn, but I have

Figure 1.14 Launching of CoMIR. Seated from right to left Three Founding Directors, Kanti Mardia, Mike Smith, David Hogg. Elizabeth Berry (seated between Smith and Hogg) is an additional new Director since 1998.

to discover the punch line. I remember the fun I used to have when solving different puzzles as a small child. I assume that the data is challenging me to uncover its message, and then it becomes a lot of fun. But in such exercises, I often find that I need newer and sharper tools to proceed. This leads to deeper data analysis and more methodological research. My ideas have been predominantly driven by some kind of data and my attempts to make sense of this data. The bottom line is the challenging data analysis where my research ideas germinate.

1.17 Center of Medical Imaging Research (CoMIR)

Mukhopadhyay: You are a founding director and now Director of the Center of Medical Imaging Research (CoMIR) in Leeds. The creation of this prestigious center within the university has become a benchmark in your career. Where would you like to start?

Mardia: Within our department, collaborative activities and research in imaging, especially for medical diagnostics, kept growing tremendously through the 1980s. Obviously there was real need for this type of fundamental research in this area. In 1992, three departments in Leeds got together for a joint venture with myself as the founding director. Professor David Hogg, an expert in artificial intelligence from the Department of Computer Science, joined hands. Professor Mike Smith, Director of the Research School in Medicine and Head of the Department of Medical Physics, joined the team. Three of us got together. The University of Leeds pumped in a lot of money and we got some external grants too. The key idea was to bring the three groups of researchers together to solve practical and important problems in medical imaging with clinical imports from the university hospital and other nearby hospitals. The area of research problems may arise from the interface of medicine, physics, imaging, modeling, design, computer hardware and/or software and so on. The CoMIR has been extremely successful.

Mukhopadhyay: Would you please describe briefly an ongoing CoMIR project?

Mardia: One substantial project consists of longitudinal data collected by an orthopedic surgeon, Professor Bob Dickson, on spinal scoliosis for a cohort of one thousand children in the age group 9–14 years over a period of five years. Images of the spinal columns viewed from two important orthogonal directions have been recorded for these children. The challenge is to be able to forecast the onset of a debilitating disease, spinal scoliosis, as early as practically possible. One has to pinpoint the presence (or absence) of the disease with a very high accuracy. The criteria for recommending the presence (or absence) of the disease have to be formulated, implemented and tested medically. Assuming that everything goes as planned, in the end, the large group of health providers in the clinics have to be trained so that they can diagnose and treat the disease appropriately. Every aspect of this project's successful completion depends heavily on each team player's full participation.

Mukhopadhyay: The project sounds very challenging. Where are you now in this project?

Mardia: The criteria to quantify the curvature of the spinal column are being developed. Alistair Walder and I have developed some of these criteria. The medical trials are continuing to test both the feasibility and validity of the suggested statistical and physical models for the early detection of the onset of spinal scoliosis. Significant theoretical as well as methodological research in statistics would have major impact on children's health. This is the kind of project of national importance for which one needs a center to attract experts from many areas under one roof. The CoMIR has been doing some fundamental work in this area (Mardia et al. 1999).

Mukhopadhyay: Would you be willing to describe another significant project under CoMIR?

Mardia: The CoMIR has been working on another project of immense practical importance. Many companies manufacture models of parts of a human body, for example, the head, brain, knee and so on, which are used to guide and/or train in the preparation for surgery or as prosthesis for a patient. Jointly with the Department of Anatomy, we are working on a project to develop statistical and computational methods to check the accuracy of the manufactured models.

Consider, for example, a manufactured model of a human head. How is it created in the first place? From a cadaver, the head is surgically removed. Then its internal and external shape, structure and content are thoroughly scanned. This scanned data is then used to create a physical model by stereolithography for a human head. But how should one compare a model head with the original cadaver head? Many deep statistical, mathematical and computational problems are involved in this project. The dental surgeon and anatomist, Alan Jackson, as well as a plastic surgeon, Hiroshi Nisikawa, are participating in all aspects of modeling because ultimately one has to decide how the bones are distributed both around and inside the physical model in relation to the real head. The challenges are numerous. There are no quick or easy solutions. But, at every turn, the team members are making progress. This research at the CoMIR is supported by the Wellcome Trust.

1.18 Visiting other places

Mukhopadhyay: Please comment on some of the exciting places you have visited.

Mardia: All my visits have been exciting. Let me, however, comment briefly on some of the visits to the U.S.A., the U.S.S.R., India and Spain. In America, I have visited many places, but visiting Princeton never fails to fascinate me. The environment in Princeton stimulated my research every time I went there. It was so kind of Geoff Watson to invite me for a month every year until 1993. I recall that it started in 1985. I had the opportunity to talk to Geoff and his colleagues at any time during these visits, but the best part of the arrangement had to do with my total freedom whenever I was there. There was never any push to work with so and so or to guide me to think like so and so. I felt totally free to pursue any research project that I wanted to pursue and Geoff was always there to give the moral support and advice. I considered Princeton my second home.

I came to know Colin Goodall at Princeton. A series of collaborations took place and are still continuing between Colin and me over many years. Subsequently, he moved to Penn State. In Princeton, Colin and Geoff organized many workshops on shapes and every single one of these was productive and stimulating. One time I complained to Henry Daniel[s] that I normally got very little money from Leeds to attend these workshops and that I had to spend a large amount of money from my own pocket to take care of the expenses during each trip. Henry said, “Kanti, remember this. It is worth getting out of England for one whole month every year even if you finance the trips yourself.” (Laughter.)

Mukhopadhyay: (Laughter.) It sounds like very saintly advice!

Mardia: Saintly advice indeed! Lengthy visits to Princeton have slowly been replaced by regular visits to Ann Arbor, Michigan, for collaborations with Fred Bookstein. You may think of it as a transition from Princeton to Ann Arbor. I like visiting Ann Arbor very much. I also visited Penn State a number of times to collaborate with Colin Goodall.

Mukhopadhyay: Didn’t you visit Bloomington, Indiana, for some time?

Mardia: In 1977, I came to visit Bloomington, Indiana, for a semester on account of Madan Puri's invitation. I taught two courses and I got the opportunity to work with Madan Puri.

Mukhopadhyay: Any recollections from your trip to the U.S.S.R?

Mardia: In 1976, there was a conference, Stochastic Geometry and Directional Statistics, in Yerevan, Armenia, U.S.S.R., which was attended by a selected group of British delegates. The list of delegates included David Kendall, Brian Ripley, John Kent, Peter Jupp and me. When we arrived there, this woman (an interpreter) repeatedly asked me, "How could you be a British delegate?" You see, I looked so different from other British delegates! Eventually I replied, "Well, I am the contradiction." (Laughter.)

I think that M. Abramowitz was one of the main organizers. This was a wonderful conference and we were treated like royalty. All the facilities were there and the talks were very enjoyable too. I still remember that the hospitality was remarkable.

We used to get breakfast a little late. It used to be sort of a brunch. I am a pucca (that is, one-hundred percent) vegetarian. The local hostess knew this. So she used to put large amounts of cereal, fruits, bread, salad, etc. on my plate in order to make up for all the missed meat and fish. Everyone else used to get very small portions!

Mukhopadhyay: I believe that there is a punch line to this story. (Laughs.)

Mardia: Oh yes! Then came the conference dinner where each delegate was supposed to propose a toast. When my turn came, I got up and said, "The nice lady who has been looking after me did such a fantastic job. I am so grateful to her. But, I did not quite understand why I was given three or four times the normal portion of bread, fruits and salads during each meal." My hostess did not realize what I was saying and in the meantime she served me a large fruit plate with a lot of varieties. It was the largest fruit plate I had ever seen. It was so funny! Everyone broke into laughter. But, as soon as she realized what I had said, she replied calmly through the interpreter, "A cow must eat a lot of grass to sustain good health." What a defense! I was amazed by her spontaneity and sense of humor. It was hilarious.

Mukhopadhyay: Kanti, you have lived outside of India for nearly thirty-five years. Even though you have Indian roots and heritage, my guess is that many a time you have made trips to India as any other visiting scientist. Any recollections of your special visits to India?

Mardia: Right after I had settled in the United Kingdom, whenever I used to visit India, I made special efforts to go and visit the University of Bombay and give seminars there. This is the place where I grew up as an academic. I have always felt that bond. I was humbled to be invited for the M. C. Chakrabarti Memorial Lectureship Endowment in 1991. There I gave a series of seminars in shape analysis with applications to image processing. When I developed those lecture notes, my shape analysis book was slowly taking its shape. I was also moved and humbled by the presence of my own professor, Mr. Mehta, in the audience. It felt like a fairy tale to me.

After C. G. Khatri's untimely death, a conference in Delhi, organized in 1990, was dedicated to his memory. I felt touched when I was invited to present a paper there in memory of my long-time friend and collaborator. I read the paper on "Khatri's contribution to the matrix von Mises-Fisher distribution and its relation with shape analysis" and I genuinely felt honored.

I have visited Jaipur, India, where my career started and gave some talks in the Department of Statistics. It was wonderful to see again my advisor, Professor G. C. Patni, after many years.

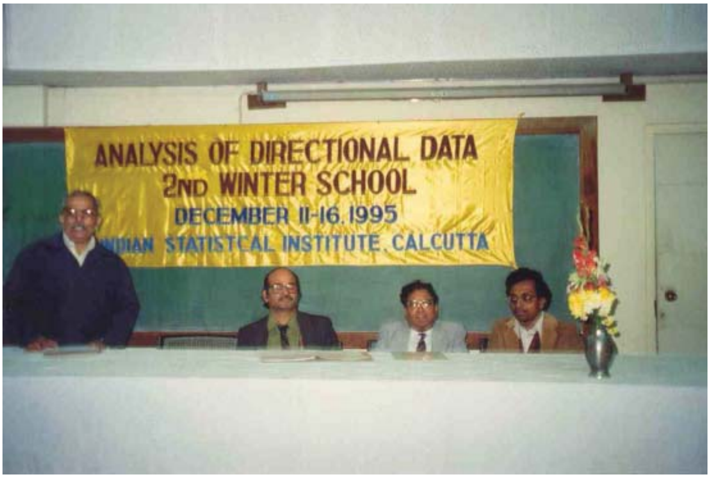

ANALYSIS OF DIRECTIONAL DATA
2ND WINTER SCHOOL
DECEMBER 11-16,1995
NDIAN STATISTCAL INSTITUTE CALCUTTA

Figure 1.15 Conference at the Indian Statistical Institute, Calcutta, 1995. From left to right: C. H. Sastri (Head, Applied Statistics Division), S. B. Rao (Director), K. Mardia and A. Sengupta.

Mukhopadhyay: Did you happen to see Professor B. D. Tikkiwal again?

Mardia: I have seen Professor Tikkiwal in large conferences, for example, at the Indian Science Congress. I do not necessarily go to visit with him when I am in India. He continues to do research on sampling. He also came to visit England and I invited him to come to Leeds, probably in 1980. He was passing through but we had some nice times together.

Mukhopadhyay: Are you going to mention your trips to Spain?

Mardia: I recall visits to the Department of Statistics in the University of Granada, Spain, for joint projects on distribution theory and spatio-temporal modeling during the last four years. I have actively collaborated with Ramon Gurtziat and José Angulo. Either I visit there once a year or someone from there visits Leeds. José visited Leeds in July, 1999.

However, the visits to the Continent do not exactly suit us since we are completely vegetarians. Even in salads, one will often find the crunchies prepared with ham! But, lately when we have visited, we have rented an apartment with kitchen facilities.

Mukhopadhyay: During a recent visit to India, you have launched long-term joint collaborations with scientists from the Indian Statistical Institute (ISI), Calcutta. Would you like to mention that?

Mardia: At one point, the Indian High Commissioner to the United Kingdom, Dr. L. M. Singhvi, became the Ambassador from India to the United Kingdom. He was very keen on creating interactions among the universities in the United Kingdom and India. He suggested looking into possible collaborations and exchange programs between Leeds and some Indian universities. I thought that ISI, Calcutta, was the right place to begin this exchange program on an experimental basis because the activities in image analysis and machine vision were strong in both places. I approached Professor Jayanta Ghosh and we made formal arrangements in 1995 to embark on the program in the next five years. The progress has been slow but several things have happened. A large conference was held in 1998–99 in ISI, Calcutta, where Ian Dryden gave a workshop, Shapes and Images. In the conference, I happened to deliver both the keynote and closing addresses. These were attended by groups of researchers in machine vision, pattern recognition, statistics, mathematics, computer science, both from within ISI, and other academic institutions as well as companies and industries. There were participants from overseas too. This was a very high profile event.

At the end of the conference, there was substantial dialogue among various groups and this was one of the objectives for initiating such a large exchange program in the first place. The former Director of ISI, Professor S. B. Rao, mentioned that this was the first time the statisticians and the staff members from the computer vision and image analysis within ISI got together on a large scale. I anticipate that there will be a reciprocative conference on shapes and images in Leeds in the year 2000 to preserve the flame. I expect a delegation of three to six researchers from India to Leeds in that event.

1.19 Collaborators, colleagues and personalities

Mukhopadhyay: Let us now hear about some influential collaborators, colleagues and personalities.

Mardia: Let me start with Geoff Watson. I first met Geoff in February or March, 1977, in Houston, Texas. The way I met him was very interesting. Earlier, Tom Downs visited us in the United Kingdom and I went to Houston to reciprocate that visit. I heard that Geoff was coming to Houston as an external examiner of one of Tom's Ph.D. students. Geoff was to stay in the university guest house. He was possibly returning from a skiing trip. He missed some connecting flights on account of bad weather and his plane was very late. He was very tired but he somehow arrived on campus around 3 a.m. Geoff knocked on the entrance to the guest house a few times, but no one came to open the door for him. He slept through the rest of the night on the "welcome mat" at the entrance.

Next morning, the Ph.D. exam was right on schedule. I was invited to observe the proceedings. Geoff arrived there with a smile on his face. Last night's episode did not bother him even the least bit. He was laughing and joking as he described what had actually happened. This was my first encounter with Geoff. He was probably the most rugged man I ever met.

Later that day, Geoff gave a seminar on genetics, and in the evening we formally met and went out to dinner together with a couple of other people. We had some informal discussions. He spoke very kindly of my book on directional data which he had reviewed earlier. Geoff was of course a pioneer in this area and his encouragement meant a lot to me. Geoff, Tom and I went to a cowboy show and Geoff quickly bought and wore a very distinctive cowboy hat. I did not anticipate this at all. He was very easy to get along with!

Mukhopadhyay: Any other recollections about Geoff Watson?

Mardia: I first visited Princeton in 1985. Henry Daniels was also visiting at that time and we were living close to each other on the campus of the Institute of Advanced Studies. That period was particularly hard for Geoff. His department was disintegrating. In fact we attended a musical evening, "On the demise of the Department of Statistics" and Henry Daniels took part in the show. Colin Goodall's wife, Lisa, had a part in this too. It was a great musical evening but the unfortunate part was that we were bidding farewell to Geoff's department.

Geoff was busy with regular teaching duties. But frequently he appeared very frustrated. He even looked depressed sometimes. So, I used to talk to him about Yoga exercises.

Figure 1.16 Kanti Mardia with Geoff Watson, in Houston, 1977.

Whenever we saw each other, we discussed what both of us were doing, but we never came up with a problem where the two of us could work together. He was always very modest about his own research. Geoff was also a great painter. He was his own man when he painted. Geoff visited Leeds perhaps four or five times. He last visited Leeds about four years ago. Whenever we came to Princeton, Geoff's wife, Shirley, was kind to take care of us. They were wonderful hosts.

Mukhopadhyay: You edited *The Art of Statistical Science* (Mardia 1992a), a seventieth birthday festschrift volume for Geoff Watson. Did you present the volume to him in person?

Mardia: In 1992, Geoff turned seventy and he was to retire. I prepared a special volume of papers in his honor. Many collaborators and admirers of Geoff participated in this volume. All contributors responded enthusiastically. In 1993, there was a conference in Princeton where many of us participated. Michael Stephens, Jim Durbin and John Kent were also in attendance. One of his daughters is a famous opera singer and she gave a recital. Geoff became very emotional and he had tears in his eyes. At the end of the conference, I presented the festschrift volume to Geoff. He could barely speak and said only a few words of appreciation. After I presented the festschrift volume, he presented to me one of his many beautiful paintings in water colors.

Mukhopadhyay: When did you last see Geoff Watson?

Mardia: Geoff was visiting Colin Goodall in New Jersey. This was 1997 when Pavan and I drove there to say hello to him. At that time he was creating a painting of the Fine Hall. He kindly gave me a print of that painting.

Mukhopadhyay: Who comes to your mind next?

Mardia: Let me give you some recollections about David Kendall. I came to know David more than twenty years ago when he invited me to Cambridge to give a seminar on directional data. My colleague John Kent was one of David's students.

From the very beginning, David and I liked each other. Subsequently, we exchanged ideas on stochastic geometry. I got to know him well during our trip to Yerevan, Armenia, U.S.S.R., on a delegation. A lot of people casually think that directional data is just another part of multivariate analysis. In Russia, when we walked together, David would emphasize that we must convince the mathematicians and statisticians that such a simplistic attitude is not correct. He argued vigorously that non-Euclidean geometry and topology actually set apart directional data and shape analysis from traditional multivariate analysis. He discussed these with unmistakable energy.

He knew my hobby of collecting editions of the *Rubaiyat of Omar Khayyam*. One day, David mentioned that he was once invited to the Omar Khayyam Club in London during a special event. He said that most of the major publishers were represented there. He went on to describe how this club was unique in its mission. At that time I did not know anything about this club. Eventually I found the club and now I am a member of the Omar Khayyam Club. It is quite a merry place. Sometimes visitors give light-hearted and hilarious lectures on Omar Khayyam and Edward FitzGerald, and other times the gathering may be quite formal. The membership consists of people from all walks of life. I presented my millennium paper (Mardia 2000) entitled "Omar Khayyam, René Descartes and solutions to algebraic equations" and put forward a thesis that Omar Khayyam's work during the twelfth century might have foreshadowed the contributions of Descartes in analytical geometry.

Figure 1.17 Kanti Mardia with David Cox in Oslo, 1977.

Mukhopadhyay: You mentioned Fred Bookstein before. Do you wish to add anything else?

Mardia: The proceedings volume of the first conference on shape, organized in Leeds in 1995, was dedicated to Fred Bookstein and David Kendall. They are both pioneers in this field. I may say that Fred complemented David's fundamental ideas and vision in his own characteristic style and created the impetus for this field's phenomenal growth.

Fred is superb in giving intuitive geometrical arguments. Frequently we have to work hard to come up with algebraic validation of Fred's original "simple" claims. But sometimes

Fred's intuitive answer and the algebraically derived answer will differ slightly, especially in higher manifolds. Then, the situation becomes serious! (Laughs.)

Mukhopadhyay: Do you now wish to give some remarks about D. R. Cox?

Mardia: With great pleasure. Let me start by saying that David R. Cox got his Ph.D. degree in statistics from the University of Leeds in 1949. He is the "jewel in the crown" of Leeds. David was jointly supervised by Henry Daniels from the Wool Industries Research Association (WIRA) and Bernard Welch from the university. Some of the early works of David Cox had to do with fiber and yarn data having long-range and serial correlations.

I first met David a long time ago, perhaps when I was in Newcastle. David is very easy to get along with and talk to for his comments and advice. When he was in London, I used to visit him for his advice and guidance on technical as well as administrative matters. When I work on some new results, I ask him for advice or related references. Regardless of the problem, whether it is statistical, mathematical, conceptual or administrative in nature, David always has something very valuable to say.

Now he is in Oxford and he is technically retired. But we all understand that retirement for David R. Cox means that he is a full-time researcher. He is a very popular person and a great, inspiring speaker. He is always in demand as an invited speaker all over the world.

Figure 1.18 Visit to Omar Khayyam Tomb in Nishabur, Iran, 1994. Standing in the garden alongside Omar Khayyam statue.

Figure 1.19 Kanti Mardia with Ulf Grenander, in Newton Institute, Cambridge, 1994.

Mukhopadhyay: So far you have not mentioned Ulf Grenander or Stu Geman of Brown University.

Mardia: In 1985 or 1986, there was a workshop or conference in Edinburgh where I first met Ulf Grenander and Stu Geman. Subsequently, I have visited Brown a number of times. I have become close to their group. I find Stu Geman extremely clever and we get along very well. When we discuss topics in image analysis, he will freely share new ideas with me. Often these ideas, once polished and tightened, lead to new concepts or measures.

Ulf Grenander came to Leeds as an invited participant at a conference, “The Art and Science of Bayesian Image Analysis,” in 1997. Of course, Ulf is mathematically very deep. Ulf is always very precise in what he says. By talking to him, one will easily discover that he is a great mathematician. At the Newton Institute in Cambridge there is a regular program of workshops and symposia to bring experts together. I have taken part in some of these workshops. So I came to know Ulf and his wife, Paj, well. I know Paj as a very outgoing person. She is an energetic bridge player. I once asked Ulf, “Your wife plays bridge in her spare time. What do you do in your spare time?” He replied, “I do not have much spare time, but when I do, I do more mathematics.”

Mukhopadhyay: So far, you have talked about some of the leading statisticians. In your research, you have met and worked with many scientists from other fields. Do you wish to mention any of these?

Mardia: I have been fortunate to meet many leading scientists in the areas of machine vision and image analysis. I may mention Joseph Kittler from Surrey and David Hogg, my colleague in the Department of Computer Science in Leeds. From the United States, I may mention some of the pioneers such as Azriel Rosenfeld and Laveen Kanal, both from the University of Maryland, and Anil Jain from Michigan State University, East Lansing. These are some of the top people in what they do.

1.20 Logic, statistics and Jain religion

Mukhopadhyay: Over the years you have done research and written extensively on the science and logical structure of the Jain religion. What is the origin of this aspect of your life?

Mardia: Sirohi is my birthplace where the predominance of Jain religion and culture is the way of life. I was nourished by the practices and philosophy of Jainism. My life has since been greatly influenced by this environment. The Jains are pure vegetarians which implies the total exclusion of meat, fish, eggs and even onions and potatoes from their diet. My involvement and passion for the Jain religion has given me a lot of excitement in life. A part of it is because I was brought up within the Jain tradition.

Mukhopadhyay: In your Inaugural Lecture delivered at Leeds in October 1975, you had talked on "Do-it-yourself statistical analysis." This lecture was a serious mix of science, philosophy, logic, statistics and Jainism (Mardia 1976).

Mardia: In Leeds up to 1980s, when one received the position of a full professor, huge official ceremonies were held for the inauguration. This was a big moment in anyone's life. The Professor would deliver a substantial inaugural address directed toward the colleagues as well as the larger community. It is an overwhelming experience. I got the Chair in Leeds in 1973, and on 13 October, 1975, I delivered the inaugural address. Lord Boyle was Vice-Chancellor of Leeds then.

In a scientific theory, one proves many results under some basic assumptions or hypothesis. In statistics, we make inferences regarding a population with the help of the information gained from a random sample. This is inductive logic. But, no assumption or hypothesis is perhaps universally true or false. In statistics, we say, "Do not reject the null hypothesis," but is it equivalent to say, "Accept the null hypothesis?" In statistics, there is a middle ground which is because in the logic of statistics, there is no place for absolutism. Simply put, a core in Jainism says, "It is wrong to assert absolutely." In fact, I really should say, "Maybe it is wrong to assert absolutely."

The idea of *nonabsolutism*, a principle which is shared by the conditional predication, was advocated by Karl Popper, one of the greatest logicians of the twentieth century. J. B. S. Haldane, a famous geneticist and statistician, also hailed nonabsolutism. My Inaugural Lecture (Mardia 1976) outlined the arguments which thread together science, philosophy, logic, statistics, Jainism and decision-making. I emphasized the utmost need for solid understanding of statistical logic and principles whenever some useful decisions are to be made. No statistical package will serve one iota of purpose for the overall good of the society unless the user of the statistical package knows both statistics and the package extremely well.

Mukhopadhyay: In your Inaugural Lecture, you pointed out that through the principle of conditional predication, one may try to justify the logical foundation of Jainism. Is there any other viewpoint?

Mardia: One may consider the holistic view called Anekāntvāda. From this viewpoint, the philosophy of Jainism requires that one must consider anything as a whole in order to understand it. Understanding some bits and pieces about something is not same as understanding the whole thing.

Mukhopadhyay: The connection between what you just said and what we normally do in statistics is very clear. Statisticians take a look at part of a population only, and by examining its features try to guess the features in the whole population. Thus the inferences made cannot be perfect. When I say that, I actually become a believer of nonabsolutism.

Mardia: Your understanding is correct.

Mukhopadhyay: You have written a very authoritative book on Jainism. Do you wish to mention anything on that?

Mardia: Jain religion is not personality based. It derives its foundations from what is called Jain-science. Once I sorted out the logical links and scientific arguments, I discussed my axiomatic theory with some scholars of Jain religion, religious leaders and leading monks both in India and abroad. My theory has been received very well. Subsequently, in Mardia (1990), I published a book, *The Scientific Foundation of Jainism*. The book has now gone into a second edition. I went to a Indian publisher in India because then the book would be more affordable to the general public. But, looking back, I realize that it was a mistake. The book is cheaper in India, but the publication quality is not very high and its copies are hardly distributed overseas. The book is essentially for members of the younger generation with scientific minds.

Mukhopadhyay: You authored a booklet, *Jain Thoughts and Prayers* (Mardia 1992b), which was prepared also for the younger generations.

Mardia: That is correct. We have to get the younger generation involved; my booklet helps in that mission. I may add that a Jain Center, including a temple, has been built in Leicester (England). I was deeply involved with the whole project for a number of years. I am Chairman of the Yorkshire Jain Foundation, formally established in April, 1987. The Foundation holds a library on Jainism and comparative religions. I am also Vice-Chairman and a Trustee of the Jain Academy which promotes educational initiatives related to Jain Studies.

1.21 Many hobbies

Mukhopadhyay: Do you have any hobbies?

Mardia: I started learning chess by myself when I was six years old. During my school and college days, I played chess extensively. My eldest brother used to play chess. I had a great fascination for this wonderful game. I became quite proficient in this game. As a student representing the University of Bombay, I played in a chess tournament, reached the college-level final, but then I lost in the final round. That loss put a damper on my pursuit of chess! I discovered later that the chap who defeated me in the final actually went on to become the national champion. I already was committed to becoming a statistician! (Laughter.)

I still play chess now and then. Everyone in my family plays chess and I hope that my four-year-old grandson, Ashwin, will pick it up too. Indeed, my grandson plays with me, but at present sometimes he makes up his own rules — he has to win! (Laughter.)

Mukhopadhyay: Do you also play bridge?

Mardia: I play bridge occasionally with my family. In the past, I have also taken part in bridge tournaments in Leeds. I collected some master points, but I was spending too much time on this to continue to play at the tournament level. Subsequently, I gave up playing in tournaments.

Mukhopadhyay: Among colleagues, have you met some exciting chess or bridge players? Do you have other hobbies?

Mardia: I am sure that I have, but I do not remember many details. I can tell you an interesting story. Probably in 1977, when I did lot of traveling from Bloomington, I went to Virginia to give a seminar. Before the seminar, I was talking to Jack Good and casually mentioned that I heard he was very good in chess and I was also interested in chess. He said, “Very good. Let us play a game before your seminar then.” It turned out to be the preseminar game. (Laughter.)

He is an extremely good chess player. He made a number of great moves. Even though I gave him a good fight, I naturally lost the game to a much better player.

My other hobby is to collect antiquarian books. When I go to conferences or visit places, whether in Europe, Asia, the U.S.A. or Canada, I must go and check out some of the best local antique bookshops. I watch whether any book fair will coincide with a conference in some city and plan my itinerary accordingly so that I can also go to the book fair while I am in that neighborhood. I have a large library of antiquarian books. When other conference delegates go to visit palaces and museums, Kanti Mardia goes to some out of the way old bookshops! (Laughter.)

In Ann Arbor, I feel at home when I walk around. One reason is that there are quite many exciting bookshops there for used or old books. I love buying and reading old books on art, religion, culture, music, society, languages, history of computing and travel. Sometimes Shirley, Geoff Watson's wife, took me to book fairs in Princeton.

Mukhopadhyay: How did this hobby get started?

Mardia: I started with the Rubaiyat of Omar Khayyam. Over the years, I have gathered a large collection. I mentioned earlier that I am a member of the Omar Khayyam Club in London.

1.22 Immediate family

Mukhopadhyay: What do you wish to mention about your immediate family?

Mardia: I consider myself very lucky to have Pavan as my wife. Pavan took care of the family's upbringing and practically all other conceivable responsibilities on top of her full-time career as a math school teacher. She is very talented. She sacrificed much more than anyone will ever know. If I am light, she is the electric current; if I am software, she is the hardware. Indeed, she ran the family and kept me in line, which is not always easy. I am actually a good-for-nothing on the domestic front. Perhaps, I should not have said that. (Laughter.)

Pavan is patient and she stays calm when suddenly some unexpected things happen. I am literally the opposite of that. She does not say much, but on the other hand I am too extrovert. Thank God, Pavan's habits and demeanor complement mine. (Laughter.)

She took early retirement about four years ago so that the two of us are now able to travel abroad together. She also takes over my general secretarial work when I travel. On the road, to have someone to talk to or share something exciting is always healthy for both minds and souls. We have many common interests; for example, we both swim, enjoy traveling and collect antiquarian books.

Mukhopadhyay: How about your son, Hemant, and daughters Bela and Neeta? Do they take after you or their mom?

Mardia: From the very beginning my children have been more attached to their mother and that is quite expected. They also inherited the best qualities from Pavan. Generally speaking, they are quite calm and patient. Our children can relate to Pavan easily. I am always there for them, but for everyday's nitty-gritty details, the children will probably have more confidence in their mom. (Laughter.)

Mukhopadhyay: (Laughter.) Kanti, there is no need to explain. I understand exactly where you are coming from. What do your daughters and son do?

Mardia: Quite early, Pavan and I decided that our children must have the freedom to pursue their own interests to build careers. We were always available for advice and guidance, but we never pushed the children to any particular profession. Our children are now grown-ups, and I am proud of their individualities and specialties.

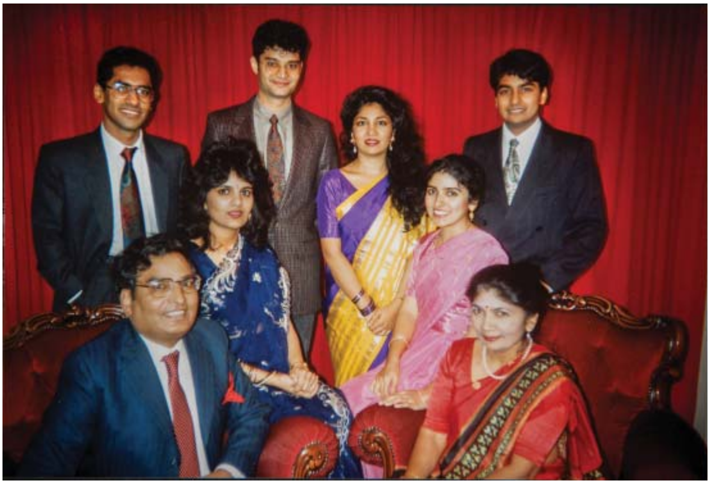


Figure 1.20 Mardia's children with spouses, 1991. Standing from left to right: Raghu (Bela's husband), Hemansu (Neeta's husband), Bela (daughter) and Hemant (son). Sitting from left to right, Kanti Mardia, Preeti (Hemant's wife), Neeta (daughter) and Pavan Mardia.

Our eldest daughter, Bela, is the only one who took some statistics courses for her degree. She has been working as a systems analyst and her husband, Raghunathan, is a medical doctor. They have given me a grandson, Ashwin, whom I mentioned earlier. Bela lives in Hull with her family.

Our son, Hemant, studied electrical engineering. He is now the director of a company in telecommunications that manufactures special types of low-frequency filters for digital technology. His wife, Preeti, is a food scientist and is now a manager of marketing. Hemant and Preeti live near our place.

Our youngest daughter, Neeta, lives more than two hundred miles away. She is a lawyer and her husband, Hemansu, is a purchasing manager in a large business complex. Raghu, Hemansu and Preeti hail, respectively, from Southern India, Gujarat and Delhi. A good cross-section of India is thus represented in my immediate family.

Mukhopadhyay: Do you get to see your children and their families frequently?

Mardia: We are very close to one another. We visit them or they come to visit us on birthdays, holidays, and other special occasions. So we all constantly keep in touch.

My children gave me a surprise sixtieth birthday party. Many of my friends from Newcastle and Hull were invited. Leeds was very well represented too. It was a big affair and the party was arranged in a cunning way! I had no idea what was about to hit me!

Mukhopadhyay: You just said you were close to your children. So what happened! (Laughter.)

Mardia: I thought all along that I knew my wife, the children, and my friend Raj very well! Behind my back, they were in this together. (Laughter.)

On a serious note, I add that I was moved and I was delighted to see so many friends and well-wishers. I enjoyed the party thoroughly.

Figure 1.21 Kanti Mardia's sixtieth birthday celebration in 1995. The birthday cake with Omar Khayyam's Rubai (adapted).

1.23 Retirement 2000

Mukhopadhyay: I hear that you are to retire in September, 2000. That is going to be an important landmark, in your career. In the professional life, are you resetting the priorities?

Mardia: Retirement is mandatory at age 65 in U.K. universities. But because I am still active in research and I have several grants, I was to become at least an emeritus professor after retirement. I have started thinking about the changes the retirement will bring. Long-term prospects are totally unknown. Things will no doubt be different. I will probably have a small desk in the corner of an office somewhere. (Laughs.)

On a serious note, I have been appointed as a full-time Senior Research Professor from 1 October, 2000, at Leeds — a special position of its type created for me. So things are going to be quite exciting.

Mukhopadhyay: Congratulations. This sounds like a great opportunity and you certainly deserve it. But let me ask you this. You have been the mover and shaker at the University of Leeds for a long time. What are some of the items on your "must do list" before retirement?

Mardia: Upon my retirement from the University of Leeds on September 30, 2000, I want to finish so many things! My top priority is to finish the book on spatial analysis with John Kent. Another top priority for me is to prepare the department in a way that in the next Research Assessment Exercise, due 2001, we receive the top grade. I want the department to march forward with solid footing. My successor will have to come aboard, along with a few other important appointments, so that the new leadership and other appointees may overlap with my administration for about a year. This is to make sure that the transition is as smooth as possible and none of the ongoing projects are affected adversely. This is a very

Figure 1.22 Kanti Mardia with his grandson, Ashwin, and the Indian High Commissioner, Dr. L. M. Singhvi, in Leeds, 1996.

difficult task to accomplish. I have prepared the department for the new leadership as best as I possibly can.

Mukhopadhyay: What will you be doing immediately after retirement from Leeds?

Mardia: I will be collaborating with the researchers at the Center of Medical Imaging Research (CoMIR) in Leeds. In the newly created position of Senior Research Professor, I will be stationed at the University of Leeds. Several universities in the United Kingdom have urged me to join their faculty as a research professor and accordingly I may visit these places perhaps a couple of times in order to stimulate their research programs. I will continue to visit abroad as I have done for a number of years.

Mukhopadhyay: For the post-retirement life, in the long haul, have you made any plans?

Mardia: Statistical research has always fascinated me and I have greatly enjoyed doing whatever I have done. If my wife's and my health cooperate, I will continue remaining active in research. Changes in our lives and careers are on the horizon, and naturally we wait with some apprehensions. But we have weathered changes in our lives so many times in the past forty years or so! Hopefully, this time around we will do all right too. I remain very optimistic.

Mukhopadhyay: You have several research grants right now. I suppose that some of these grants will continue into your retirement.

Mardia: This is right. I am all set for a number of years. A few postdoctoral fellows and four graduate students will continue to work with me. However, the Senior Research Professorship will allow additional postdoctoral fellows and graduate students.

The remit of this position is to continue to promote and lead research in the department. We already have made plans for the next five years' LASR workshops. My guess is that the research papers will flow for many years to come. A joint paper with Chris Glasbey is expected to be read to the Royal Statistical Society in the near future, I think.

I hope to continue my collaborations on projects at the CoMIR. I hope something big and something useful will evolve from the joint initiatives between Leeds and the Indian

Statistical Institute. I would like to visit the United States perhaps for two or three months a year to carry out joint research with my collaborators.

Naturally, there will be some financial constraints, but we will be all right as long as I stop spending money on those antiquarian books! (Laughs.)

Mukhopadhyay: Do you want to mention any other big plans?

Mardia: I have decided that I am not going to start the revisions of any of my earlier books. However, I may not mind an advisory role in revising the multivariate analysis book. I will not really enjoy rehashing these old materials. I have decided to take up the challenge to write two books simultaneously. I expect that one of the books will be on spatio-temporal modeling. The other one will likely be on statistics of images. I organized a number of conferences and edited special volumes in related areas. The material is there. But, the books are expected to be more self-sufficient and comprehensive. A synthesis would be the most important aspect in each book. These are two very substantial future projects. Hopefully I will succeed.

I am also seriously looking into the history of computing. I may narrow this field down to the history of statistical computing. We all know and appreciate the fact that the area of statistical computing has come a long way in the past fifty or sixty years. It will be great to compile this history of development.

I have so many serious projects planned beyond retirement! I sometimes wonder myself whether I will be able to reach my goals. An Indian proverb comes to mind: “When I had teeth, I could not afford those crispy chickpeas. But now that I can afford an unlimited supply of crispy chickpeas, to my amazement I discover that I have no teeth left.” (Laughter.)

Mukhopadhyay: Your record speaks for itself. I have all the confidence that you will indeed finish all these marvelous projects in the very near future. Thank you so much for this conversation which I have enjoyed immensely. I wish you, Pavan, and your loved ones all the health and happiness in the world.

Mardia: Many thanks, Nitis.

Acknowledgments

Nitis Mukhopadhyay expresses his gratitude to Mrs. Pavan Mardia, wife of Kanti Mardia, for her help during this project. He also thanks Professor Leon Jay Gleser, former Executive Editor of Statistical Science, for his valuable suggestions and thorough editing of the manuscript.

References

Anderson TW 1958 *Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, Inc., New York.

Besag JE 1986 On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society Series B* **48**, 259–302.

Bingham C and Mardia KV 1978 A small circle distribution on a sphere. *Biometrika* **65**, 379–389.

Bookstein FL 1986 Size and shape spaces for landmark data in two dimensions (with discussion). *Statistical Science* **1**, 181–242.

Bookstein FL 1989 Partial warps: thin plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 567–585.

Bookstein FL 1996 Landmark methods for forms with landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis* **1**, 225–243.

Cochran WG and Cox GM 1950 *Experimental Designs*. John Wiley & Sons, Inc., New York.

Cramér H 1946 *Mathematical Methods of Statistics*. Princeton University Press.

Dryden IL and Mardia KV 1991 General shape distributions in a plane. *Advances in Applied Probability* **23**, 259–276.

Dryden IL and Mardia KV 1998 *Statistical Shape Analysis*. John Wiley & Sons, Inc., New York.

Fisher RA 1953 Dispersion on a sphere. *Proceedings of the Royal Society of London Series A* **217**, 295–305.

Fisher NI, Lewis T and Embleton BJJ 1987 *Statistical Analysis of Spherical Data*. Cambridge University Press.

Geman S and Geman D 1984 Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Goodall CR and Mardia KV 1994 Challenges in multivariate spatio-temporal modelling *Proceedings XVII International Biometrika Conference*, vol. 1, pp. 1–17, Hamilton, Canada.

Goodall CR and Mardia KV 1999 Projective shape analysis. *Journal of Computational and Graphical Statistics* **8**, 143–168.

Gower J 1975 Generalized Procrustes analysis. *Psychometrika* **40**, 33–51.

Hartley HO 1950 The use of range in analysis of variance. *Biometrika* **37**, 271–280.

Hartley HO and Smith WB 1968 A note on the correlation of ranges in correlated normal samples. *Biometrika* **55**, 595–597.

Johnson NL and Kotz S 1972 *Continuous Multivariate Distributions*. John Wiley & Sons, Inc., New York.

Jupp PE and Mardia KV 1979 Maximum likelihood estimators for the matrix von Mises-Fisher and Bingham distributions. *Annals of Statistics* **7**, 599–606.

Kempthorne O 1952 *The Design and Analysis of Experiments*. John Wiley & Sons, Inc., New York.

Kendall DG 1989 A survey of the statistical theory of shape (with discussion). *Statistical Science* **4**, 87–120.

Mardia KV 1962 Multivariate Pareto distributions. *Annals of Mathematical Statistics* **33**, 1008–1015.

Mardia KV 1967a Correlation of the ranges of correlated samples. *Biometrika* **54**, 529–539.

Mardia KV 1967b Discussion of "a study of low-temperature probabilities in the context of an industrial problem," by V. D. Barnett and T. Lewis. *Journal of the Royal Statistical Society Series A* **130**, 202.

Mardia KV 1967c A nonparametric test for the bivariate two-sample location problem. *Journal of the Royal Statistical Society Series B* **29**, 320–342.

Mardia KV 1967d Some contributions to contingency-type bivariate distributions. *Biometrika* **54**, 235–249.

Mardia KV 1968 Small sample power of a nonparametric test for the bivariate two-sample location problem in the normal case. *Journal of the Royal Statistical Society Series B* **30**, 83–92.

Mardia KV 1970a *Families of Bivariate Distributions*. Griffin, London.

Mardia KV 1970b Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530.

Mardia KV 1971 The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika* **8**, 105–121.

Mardia KV 1972 *Statistics of Directional Data*. Academic Press, New York.

Mardia KV 1975 Statistics of directional data (with discussion). *Journal of the Royal Statistical Society Series B* **37**, 349–393.

Mardia KV 1976 Do-it-yourself statistical analysis. Inaugural address. *Leeds Review* **19**, 79–98.

Mardia KV 1977 Mahalanobis distances and angles In *Multivariate Analysis IV* (ed. Krishnaiah PR) North-Holland, Amsterdam.

Mardia KV 1989 Discussion of "A survey of the statistical theory of shape," by D. G. Kendall. *Statistical Science* **4**, 108–111.

Mardia KV 1990 *The Scientific Foundation of Jainism*. Motilal Banarsidass, Delhi.

Mardia KV 1991 On longevity and life-line. *Journal of Applied Statistics* **17**, 443–448.

(ed. Mardia KV) 1992a *The Art of Statistical Science. A Tribute to G. S. Watson*. John Wiley & Sons, Inc., New York.

Mardia KV 1992b *Jain Thoughts and Prayers*. Yorkshire Jain Foundation, Leeds.

Mardia KV 2000 Omar Khayyam, René Descartes and solutions to algebraic equations (abstract) *International Congress in Commemorating Hakim Omar Khayyam Neyshabouri, 900th Death Anniversary 9*, Neyshabour, Iran.

Mardia KV and Constable PDL 1992 On shape and size analysis of palmar interdigital areas. *Journal of Applied Statistics* **19**, 285–292.

Mardia KV and Dryden IL 1989a Shape distributions for landmark data. *Advances in Applied Probability* **21**, 742–755.

Mardia KV and Dryden IL 1989b The statistical analysis of shape data. *Biometrika* **76**, 271–281.

Mardia KV and Edwards R 1982 Weighted distributions and rotating caps. *Biometrika* **69**, 32–330.

Mardia KV and Gadsden RJ 1977 A circle of best-fit for spherical data and areas of vulcanism. *Journal of the Royal Statistical Society Series C* **26**, 238–245.

Mardia KV and Goodall CR 1993 Spatial temporal analysis of multivariate environmental monitoring data In *Multivariate Environmental Statistics* (ed. Patil GP and Rao CR) North-Holland Amsterdam pp. 347–386.

Mardia KV and Hainsworth TJ 1988 A spatial thresholding method for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**, 919–927.

Mardia KV and Holmes D 1980 A statistical analysis of megalithic data under elliptic pattern. *Journal of the Royal Statistical Society Series A* **143**, 293–302.

Mardia KV and Jupp PE 1999 *Statistics of Directional Data*, 2nd ed. John Wiley & Sons, Inc., New York.

Mardia KV and Kent JT 1988 Spatial classification using fuzzy membership models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**, 659–671.

Mardia KV and Marshall RJ 1984 Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135–146.

Mardia KV and Puri ML 1978 A spherical correlation coefficient robust against scale. *Biometrika* **65**, 391–395.

Mardia KV and Spurr BD 1973 Multisample tests for multimodal and axial circular populations. *Journal of the Royal Statistical Society Series B* **35**, 422–436.

Mardia KV and Sutton TW 1975 On the modes of a mixture of two von Mises distributions. *Biometrika* **62**, 699–701.

Mardia KV and Sutton TW 1978 A model for cylindrical variables with applications. *Journal of the Royal Statistical Society Series B* **40**, 229–233.

Mardia KV and Watkins AJ 1989 On multimodality of the likelihood in the spatial linear model. *Biometrika* **76**, 289–295.

Mardia KV and Zemroch PJ 1978 *Tables of the F- and Related Distributions with Algorithms*. Academic Press, New York. [Russian translation (1984). Nauka, Moscow.].

Mardia KV, Baczkowski AJ, Feng X and Hainsworth TJ 1997a Statistical methods for automatic interpretation of digitally scanned fingerprints In *Special Issue: Pattern Recognition Letters* (ed. Gelsema ES and Kanal LN) vol. 18 North-Holland Amsterdam pp. 1197–1203.

Mardia KV, Qian W, Shah D and de Souza K 1997b Deformable template recognition of multiple occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**, 1036–1042.

Mardia KV, Goodall CR and Walder AN 1996 Distributions of projective invariants and model-based machine vision. *Advances in Applied Probability* **28**, 641–661.

Mardia KV, Goodall CR, Redfern EJ and Alonso FJ 1998 The kriged kalman filter (with discussion). *Test* **7**, 217–285.

Mardia KV, Kent JT and Bibby JM 1979 *Multivariate Analysis*. Academic Press, New York.

Mardia KV, Kent JT and Briden JC 1983 Linear and planar structure in ordered multivariate data as applied to progressive demagnetization. *Geophysical Journal of the Royal Astronomical Society* **75**, 593–621.

Mardia KV, Puri ML and Edwards R 1978 Analysis of central place theory. *Bulletin of the International Statistical Institute* **47**, 93–110.

Mardia KV, Walder AN, Berry E, Sharples D, Milner PA and Dickson RA 1999 Assessing spinal shape. *Journal of Applied Statistics* **26**, 735–745.

Newrick PG, Affie E and Corrall RJM 1990 Relationship between longevity and life-line: a manual study of 100 patients. *Journal of the Royal Society of Medicine* **83**, 499–501.

Ord JK 1972 *Families of Frequency Distributions*. Griffin, London.

Pearson ES and Hartley HO 1972 *Biometrika Tables for Statisticians 2*. Cambridge University Press.

Pitt HR 1963 *Integration, Measure and Probability*. Oliver and Boyd, London.

Plackett RL 1965 A class of bivariate distributions. *Journal of the American Statistical Association* **60**, 516–522.

Rao CR 1952 *Advanced Statistical Methods in Biometric Research*. John Wiley & Sons, Inc., New York.

Ripley B 1981 *Statistical Inference for Spatial Processes*. Cambridge University Press.

Switzer P 1983 Some spatial statistics for the interpretation of satellite data (with discussion). *Bulletin of the International Statistical Institute* **50**, 962–971.

Watson GS 1973 Review of K. Mardia's "Statistics of Directional Data". *Technometrics* **15**, 935–936.

Wheeler S and Watson GS 1964 A distribution-free two-sample test on a circle. *Biometrika* **51**, 256–257.

Wishart J 1928 The generalized product moment distribution in samples from a normal multivariate population. *Biometrika* **20A**, 32–52.

2

A Conversation with Kanti Mardia: Part II

Nitis Mukhopadhyay

Department of Statistics, University of Connecticut, Storrs, CT, USA

In April 1999, Kanti Mardia and I sat down to record a comprehensive conversation at the time which later appeared in Statistical Science (Mukhopadhyay 2002). He was 64 in 1999. Mardia continues to be one of the leading researchers internationally. It is a second nature for him to break new grounds and open new horizons. He travels extensively all over the globe as a true ambassador of statistical science.

Enquiring minds will surely want to know what this amazingly colorful colleague with nonstop energy has been up to since 1999. I take this opportunity prior to Mardia's eightieth birthday celebration to record a brief but updated conversation. I am on a mission to discover how Mardia's life, work, and views of statistical science have twisted and turned during the past 15 or so years. The following conversation took place in July 2014.

2.1 Introduction

Mukhopadhyay: Kanti, first let me congratulate you on your upcoming eightieth birthday celebration in 2015. How does it feel to be almost 80?

Mardia: Thank you, Nitis. It feels wonderful to be almost 80. Life continues to be interesting, though sometimes it can be too hectic.

Mukhopadhyay: If you do not mind, shall I take this opportunity to converse with you about your life and work? How about detailing whatever you have been up to since April 1999 when we had our first comprehensive conversation (Mukhopadhyay 2002)?

Mardia: That should be fun. Please go right ahead.

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

Figure 2.1 Kanti Mardia and Nitis Mukhopadhyay at the University of Connecticut-Storrs, during the first conversation. April 19, 1999.

2.2 Leeds, Oxford, and other affiliations

Mukhopadhyay: What is your position now at Leeds? What have been your major responsibilities?

Mardia: In our Statistical Science interview, we talked about my impending retirement in October 2000. At that time, it was known that I was appointed as a full-time Senior Research Professor at the University of Leeds, though on a rolling contract. There has been increasing focus on getting grants as a precondition but I could not get a major one except one in 2002. However, the University has been renewing the contract year by year on considering my other contributions.

In particular, the post has allowed me to take PhD students and lead the Leeds Annual Statistical Research (LASR) workshops. These PhD students have been mostly involved with research on statistical bioinformatics – a subject which I did not even mention in my earlier conversation. More than ten PhD students have completed their degrees. Two are continuing, one at Leeds and one at Oxford University.

Recall that I was preparing the Department then for the Research Assessment Exercise, one of my last tasks as Chair of the Department. We came through with flying colors and this performance also continued in RAE 2008. In fact, the Department was able to include my name in the RAE 2008 and again in what is called REF 2014. REF depends very much on case studies having made an impact, and the only one from the Statistics Department is mainly from my initiative; this case study relates to FASD, which I will tell you more about later. As commented by the former Dean, Mike Wilson: this goose keeps on laying golden eggs!

Mukhopadhyay: You mentioned Oxford University. What is your affiliation at Oxford?

Mardia: I have been a Visiting Professor in Oxford from March 2013. I enjoy this appointment as the ethos of Oxford is so inspiring. I have developed successful collaborations there including supervision of a PhD student. I was awarded a Dorothy Hodgkin

Figure 2.2 25th Anniversary of the Department of Statistics, Oxford University, South Park Road, Oxford. Front row standing 4th from left: Kanti Mardia. October 4, 2013.

studentship which allowed selection of a PhD student from overseas, including India and China. Thanks go to Clive Bowman, formerly Director at Glaxo-Smith-Kline (GSK) for initiating this studentship.

Mukhopadhyay: What other positions have you held, perhaps elsewhere, especially in India and China?

Mardia: I was appointed an Adjunct Faculty in the renowned Indian Institute of Management, Ahmadabad (IIMA) from March 2012 – March 2014 though I have been visiting informally from 2008 for conference presentations and seminars. At first, visiting IIMA as Adjunct Faculty fitted very well with my travel plans as it gave me and my wife, Pavan, opportunities to reunite with our families who live there. But the IIMA framework required longer visits which I could not accommodate just yet. So my formal appointment has come to an end, though informal visits and collaborations continue such as with Professor Arnab Laha.

Mukhopadhyay: These sound definitely exciting.

2.3 Book writing: revising and new ones

Mukhopadhyay: What is the status of some of your previous books? Should I watch for new editions?

Mardia: I had said in my 1999 conversation that my top priority was to finish writing the book on spatial analysis with John Kent. Alas, the effort is still going on, and I am hopeful that it will be finished in 2015. The end is in sight. Our publisher, Wiley, keeps reminding us the phrase, "if manuscripts are not realised – they escape!"

Figure 2.3 From left to right: Ashish Nanda (Director), Kanti and Pavan Mardia, Raghuram (Dean). IIMA Director's office. January 8, 2014.

I optimistically mentioned at the time that I would write a new book on spatial temporal modeling and another one on statistical imaging. Neither has materialized yet. I also said that I was not going to start the revision of any of my books, other than the multivariate analysis book (Mardia et al. 1979) in an advisory capacity. In fact, its second edition is in preparation with Charles Taylor as a new coauthor in place of John Bibby. Ian Dryden is revising the shape analysis book (Dryden and Mardia 1998). Also there have been so many developments in directional statistics that it would be worthwhile to make a new edition of the directional statistics book with Peter Jupp (Mardia and Jupp 2000).

Mukhopadhyay: Please tell me about your new book initiatives since 1999.

Mardia: There has been more concentration on research articles and the spatial analysis book.

Since 2000 through 2013, I have jointly edited LASR Proceedings, produced annually. I am pleased that we took the important step of making these proceedings available online. One may visit: www1.maths.leeds.ac.uk/statistics/workshop. Another important book (Hamelryck et al. 2013) is the volume on *Bayesian Methods in Structural Bioinformatics*, jointly edited with Thomas Hamelryck (Copenhagen University) and Jesper Ferkinghoff-Borg (Technical University of Denmark, Lyngby).

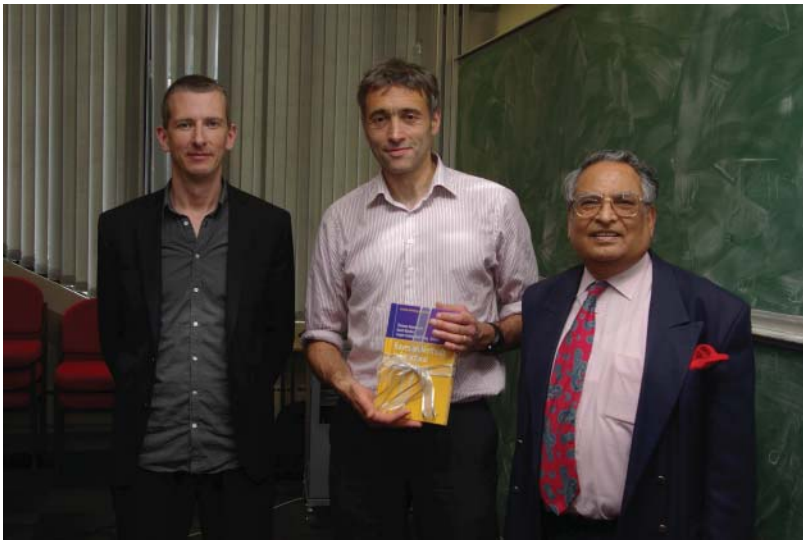


Figure 2.4 Thomas Hamelryck (left), David Westhead (center), and Kanti Mardia (right). Launch of ‘Bayesian Methods in Structural Bioinformatics’. LASR in July 2012.

Mukhopadhyay: I understand that you feel excited about the Foreword specially prepared for this edited volume.

Mardia: The Foreword was written by Gerard Bricogne (Global Phasing Ltd., Cambridge, UK) who summarized our objective succinctly in the opening paragraph: “The publication of this ground-breaking and thought-provoking book in a prestigious Springer series will be a source of particular pleasure and of stimulus for all scientists who have used Bayesian methods in their own specialized area of Bioinformatics, and of excitement for those who have wanted to understand them and learn how to use them but have never dared ask.” The 2013 Hamelryck et al. book, together with my paper (Mardia 2013) in the Journal of Royal Statistical Society, Series C (JRSS, C) would give a good start for a new researcher interested in this area.

2.4 Research: bioinformatics and protein structure

Mukhopadhyay: Please tell me about your JRSS, C paper of 2013.

Mardia: Proteins are the workhorses of all living systems, and protein bioinformatics deals with analysis of protein sequences (one-dimensional) and their structures (three-dimensional). This paper reviews statistical advances in three major active areas of protein structural bioinformatics: structure comparison (alignment), Ramachandran plots, and structure prediction. These topics play a key role in understanding one of the greatest unsolved problems in biology – that is, how proteins fold from one dimension to three dimensions and have relevance to protein functionality, drug discovery, and evolutionary biology.

In each area, the paper gave the biological background and reviewed one of the main bioinformatics solutions to a specific problem in that area. It then presented statistical tools recently developed to investigate these problems, consisting of Bayesian alignment, directional distributions, and hidden Markov models. It illustrated each problem with a new case study and described what statistics can offer to these problems. It also highlighted challenges facing these areas and concluded with an overall discussion. I feel that this unique format makes the paper accessible to statisticians as well as bioinformaticians.

I am proud of this paper, which by the way I had hoped to be a “Discussion Paper” for the Royal Statistical Society, but this format did not materialize. It turned out that way, perhaps because those who could referee it happened to be mostly my collaborators. Hence, the editorial board had problems in getting it refereed for a discussion paper!

Mukhopadhyay: By the way, what is protein sequence and protein structure? The readers will benefit from your brief explanations.

Mardia: A protein is a sequence of amino acids, of which there are 20 types, and each amino acid has a one-letter code: A, C, D, ...; for example, DYMQKREVDLHN represents a protein subsequence (in contrast to A, C, T, G for DNA). Broadly speaking, a protein structure is concerned with its three-dimensional atomic configuration. Details might include the location of particular atoms, such as the C_α carbon atom which is present in every amino acid. Other important features of protein structure include so-called elements of secondary structure, in particular, α-helices and β-sheets. An α-helix is a helix which contains 3.6 amino acids per turn and a β-sheet is composed of aligned strands of amino acids, called β-strands. These secondary structures are largely summarized by dihedral angles between certain atoms of amino acids, of which the most important are the ϕ- and ψ-angles that occur in alternation along the protein.

Mukhopadhyay: Will you please explain the phrases such as α-helix, β-sheet, and the ϕ- and ψ-angles in simple terms for the general readership?

Mardia: A protein has two parts: one part is known as the backbone (main chain) which has a repeated sequence of three atoms, carbon (C), nitrogen (N), and carbon (C_α). The second part is the side chain which is attached to the carbon atom C_α in the backbone and it is different for each of 20 amino acids. The backbone plays a major role in understanding protein function and, in view of the physicochemical properties, this can be summarized in terms of the dihedral angle ϕ between the four consecutive atoms (C, N, C_α, C) and the dihedral angle ψ between the next four atoms (N, C_α, C, N) of the backbone (imagine these five atoms C, N, C_α, C, and N in a sequence).

When a protein folds, it has two main repeated patterns (secondary structure): α-helix, β-sheet. In fact, the late William Astbury of Leeds University found that there were repeated patterns in a protein and he called these α and β patterns. Subsequently, the details of the two shapes – one as a helix and the other as a sheet with strands were discovered. See http://www.leeds.ac.uk/heritage/Astbury/Molecular_models/index.html.

Mukhopadhyay: Will you care to explain what is a Ramachandran plot and what does it do? I am sure that a brief explanation will help the general readership.

Mardia: The angles ϕ and ψ lie between 0 and 2π, and these angles can be shown in a scatter plot – now known as a Ramachandran plot which was invented in 1963 by Ramachandran jointly with his colleagues in Chennai.

The plot is an unwrapped version on a plane of the points on a torus. Such a plot indicates which areas are allowed for the angles to cover so that there are no-go (forbidden) areas. It also shows clusters for different shapes such as α-helix or β-sheet so that it is viewed as a classification map. Such plots are used in assessing the quality of new protein structures as one of their main applications, namely in assessing how many points may go into the no-go areas.

Mukhopadhyay: What research interests have been predominant in your deliberations since 1999? Please explain their importance and novelty.

Mardia: Since 1999, I have been mainly focusing on new statistical methodology required for structural bioinformatics. It turns out that it needs advances in directional statistics and shape analysis, among other areas.

In particular, I have been focusing on protein structure and functions, which has applications in drug discovery, medicine, and evolutionary biology. Proteins are extremely important for all living systems, but there is still a lot of mystery in their functions. A malfunction or misfold leads to diseases, such as Alzheimer's and cancer.

Mukhopadhyay: Handling protein folding is tough, is it not?

Mardia: One of the hardest problems in biology is that of protein folding, which affects protein function, that is, how protein from amino acid sequences (one dimension) folds into three dimensions. What I am working on is, in a tangential way, moving toward solving this puzzle.

Mukhopadhyay: How did you come into the area of protein bioinformatics?

Mardia: I came into this field by chance. In 1999, the late Harshinder Singh from the National Institute for Occupational Safety and Health, West Virginia University, Morgantown, invited me to collaborate on a problem in protein bioinformatics.

The problem involved deriving entropy of molecules such as methanol which reduces to an application of directional distribution. In fact we published the first paper on this in LASR 2001 (Demchuk et al. 2001). The paper has also to do with a particular protein TNF-beta (which has 707 dihedral angles) and is one of the key mediators of AIDS pathogenesis.

Then, we started working on multivariate von Mises distribution since the atomic structure of protein can be described by a set of dihedral angles. This paper (Mardia et al. 2008) was published after Singh passed away.

Mukhopadhyay: Did you form a critical mass of researchers in Leeds and elsewhere dedicated to protein bioinformatics?

Mardia: Yes, I feel I succeeded in forming a critical mass slowly and steadily. To start off, I looked for a collaborator to carry the subject forward in Leeds and was lucky to come to know Dave Westhead (Professor of Bioinformatics, Leeds University) and his PhD student Nicola Gold, who helped us to get into deeper aspects of the subject. Fortunately, then to create a critical mass, we had several good PhD students jointly with my colleagues in the Department and with biologists in Leeds; the biologists also actively participated in the thematic LASR workshops. In particular, it led to the development of unlabeled shape analysis, which was important for aligning proteins (Green and Mardia 2006). Also we developed how to detect biomarkers using a statistical model and EM algorithm protein gel data of renal cancer which appeared in the *Annals of Applied Statistics* (Mardia et al. 2012b). Another paper of great potential interest in drug discovery appeared in *Biometrics* (Mardia et al. 2011); it is on modeling what are called pharmacophores; a pharmacophore characterizes the physicochemical properties common to all active molecules, called ligands.

Mukhopadhyay: Will you please highlight some of your other major contributions in this area?

Mardia: One of the major contributions has been with Thomas Hamelryck, solving the probabilistically local structural prediction problem. Given a sequence of amino acids, we predicted what will be, for example, secondary structure such as the helix, the β-sheet and so on. The paper (Boomsma et al. 2006) appeared in the *Proceedings of National Academy of Sciences* (*PNAS*), and the valuable software which has been used by biologists is available in the public domain.

2.5 Research: not necessarily linked directly with bioinformatics

Mukhopadhyay: This may be a good time to summarize succinctly some of your highly influential publications since 1999 in areas which are not necessarily linked directly with bioinformatics. Kanti, will you please elaborate?

Mardia: Some key papers I would like to mention include the *PNAS paper and the Significance paper* (Mardia et al. 2013) on Foetal Alcohol Spectrum Disorder (FASD). While much of my work during this period has been motivated by protein bioinformatics, I have also published extensively in other highly visible areas of research.

An important paper, following my old interest in image analysis, was the discussion paper (Glasbey and Mardia 2001) in JRSS B with Chris Glasbey. There were a number of good discussants included from the image analysis community as well as statisticians. This has led to some new work by others, but I stopped working in this area to concentrate more on statistical bioinformatics.

Other important work has been with Fred Bookstein with whom I have been visiting for many years starting from 1996. Our paper (Mardia et al. 2006) in intrinsic random field has been important in application to FASD. A recent paper in *Significance* (Mardia et al. 2013) had a substantial impact in the sense that Fred Bookstein has served as an expert witness on FASD at murder trials.

Mukhopadhyay: Who else have you collaborated with?

Mardia: I have collaborated with Sujit Sahu and Giovanna Jona Lasinio on spatial temporal modeling, extending my Kriged Kalman filter (Sahu and Mardia 2005; Sahu et al. 2005). I also worked on projective shape with Vic Patrangenaru (Mardia and Patrangenaru 2005). This area still needs more attention (Kent and Mardia 2012). Another collaboration is with Eulogio Pardo-Igúzquiza and related to spatial statistics on maximum likelihood estimator of a spatial model. The methodology is implemented in the software MATERN (Pardo-Igúzquiza et al. 2009). Further work in image analysis includes a paper on image deformation with Miguel Angelo (Mardia et al. 2006). It has been fun to continue to collaborate so widely.

Mukhopadhyay: Currently which group or groups are you collaborating with?

Mardia: I am collaborating with the group in Copenhagen headed by Thomas Hamelryck and with another group in Oxford University headed by Charlotte Deane. I hope that these will resolve various cutting-edge problems.

Mukhopadhyay: How did some of these collaborations begin?

Mardia: Let me start with my collaboration with Thomas Hamelryck – my work with Thomas started with one query which he had sent to me and John Kent. John took the opportunity to collaborate with Thomas, and then I came to know Thomas well during his visit to LASR. We started working on using bivariate distributions into his hidden Markov model for local structural prediction for protein. Wouter Boomsma was his PhD student at that time, and he imaginatively pursued the challenge.

The work with Thomas still continues, having worked with quite a number of his colleagues and researcher collaborators. One aim is to produce a global structural predictor of proteins, and we have been working on this theme by using a reference ratio method (Mardia et al. 2012a; Mardia and Hamelryck 2012) in conjunction with our local predictor.

Mukhopadhyay: Please explain what may entail a global structural predictor and a reference ratio method.

Figure 2.5 LASR 2011 team welcoming Vice-Chancellor Michael Arthur. From left to right: Kanti Mardia, Michael Arthur, Pavan Mardia, Jochen Voss, and Arief Gusnanto.

Mardia: The proteins I have been talking about are called globular proteins. These have compact spherical shape, that is, these may fold. Roughly speaking, the reference ratio method allows the combination of a probability distribution of local structure (which would lead to predicted protein as noncompact) with the probability distribution of some global variables, which will make the predicted protein compact. The problem is still open for the latter distribution. We do not know fully what global variables would do the trick.

Mukhopadhyay: Could you tell me about your collaborations with Peter Green?

Mardia: Collaboration with Peter Green started in a curious way when I was giving a seminar in Bristol in 2003 on the alignment method which we had been developing at Leeds. After the seminar, Peter said something like, “We can jointly improve the method using Bayesian methods.” The improved paper (Green and Mardia 2006) was published in *Biometrika*. We still collaborate, but with Peter’s simultaneous appointment in Sydney, our once vigorous collaborations are now slowing down. Our good friendship continues.

Mukhopadhyay: When speaking about your position in Oxford, you mentioned your new collaboration with Charlotte Deane. That must be exciting.

Mardia: Oh, yes. Since joining Oxford University, I have been lucky to start collaborating with Charlotte Deane who leads a very large group of researchers on bioinformatics.

I already have one joint PhD student. We discuss the works of her other PhD students and see if I can provide important feedback. In particular, I have worked with Henry Wilman who is working on a challenging problem for drug discovery. But, interestingly, it boils down to analyzing the geometry of a helix (Deane et al. 2013). It was nice of Charlotte to nominate me to be a Fellow of Kellogg College, which has been a new experience.

Mukhopadhyay: Other selected collaborators you may mention?

Mardia: Other kinds of work have been pursued with many collaborators during these years, including Dave Westhead and Richard Jackson (Leeds University Bioinformatics

Figure 2.6 Kanti Mardia with James Watson in Cold Spring Harbor, March 30, 2006.

Group), Douglas Theobald (Brandeis University, Massachusetts), and Luigi Ippoliti (University of Chieti, Italy).

2.6 Organizing centers and conferences

Mukhopadhyay: Given your high profile presence in the forefront of bioinformatics research, why have you not created a new center dedicated to bioinformatics in Leeds?

Mardia: Since I saw the development in bioinformatics, I thought of channeling these activities through a Centre of Statistical Bioinformatics which was founded in 2006 with Wally Gilks from Cambridge. My dream to make it the world center has not yet materialized. However, its influence on our MSc course and PhD intake, plus follow-up and publications has been evident.

In 2006, I was still dreaming and I went to see James Watson to assess what was his view of the importance of statistics in bioinformatics. I recall a very vague answer, but he remembered with some fondness the late William Astbury (Leeds University) who pioneered molecular biology.

Mukhopadhyay: In your assessment, why has such a dream of the world center not become a reality? Has a less-than-ideal level of funding been a sore point?

Mardia: One of the saddest parts of my last 15 years is that I did not succeed in getting any research grant to hire a postdoctoral fellow or a junior researcher in a field motivated by bioinformatic applications. I have consistently received one negative report whereas the others would rate my proposals “outstanding.” The negative reports have carried more weight, perhaps due to the change of structure of the panel at EPSRC. Previously, there was a full statistical panel, but now it is a part of the mathematics panel. All my research in the last 15 years in this area has been with PhD students and collaborators. With additional funding, I am sure that the subject of protein bioinformatics, in particular, would have had more cutting-edge statistical methodologies in its forefront.

In spite of the claim by the funding authorities to support interdisciplinary projects, mostly it has been a difficult concept to execute, partly hampered by the reviewing process of the grants. For example, in reviews of my interdisciplinary grant applications a mathematician thinks it is too "applied," whereas a biologist thinks it is too "mathematical." My last successful major EPSRC application was in mining with Peter Dowd in 2002. The topic was "Stochastic Modeling of Fractures in Rock Masses," which is directly relevant to depositing nuclear waste.

Mukhopadhyay: All corners of this world around us are now more accessible than ever and there exist established as well as upcoming bioinformatics centers. University of California, Berkeley, University of Louisville, Kentucky, and others come to mind. So, the situation is not entirely hopeless. What are your global assessments about some of these existing leading centers?

Mardia: The centers you have mentioned are mainly working on DNA, micro-arrays, and so on. I have been working with my collaborators (Copenhagen, Oxford, and Brandeis University) on Bayesian inference of protein structure. As far as I know, the other "Laboratories" I think of would include those in Seattle, Stanford, Ann Arbor, and Duke.

Mukhopadhyay: In our last published conversation (Mukhopadhyay 2002), we discussed at length the creation of the Center of Medical Imaging Research (CoMIR) and many wonderful prospects. What is the present status of CoMIR? What is your role there now? Would you say that your original vision behind CoMIR has largely been attained? How so? Please explain.

Mardia: Unfortunately the main Director, Mike Smith, moved to another university. So, the progress with the Center became very limited eventually and it slowly died a natural death. We simply could not attract input and financial support from industries. My experience says that computer scientists play a key role in converting methodological advances into a usable resource. This experience has helped me in finding the right type of collaborators in statistical bioinformatics.

Mukhopadhyay: I am sorry to hear that CoMIR has nearly folded with time. The severe economic downturn felt all over the globe since 2007-2008 is probably one reason for reduced funding that led to CoMIR's untimely demise. On a more positive note, will you share the present status of LASR?

Mardia: LASR has been flourishing and it has kept pace with the times; see http://www1.maths.leeds.ac.uk/statistics/workshop. In the first decade, we started emphasizing geosciences, but in the second decade we focused on image analysis. In the third decade, beginning with the year 2000, the LASR has mainly focused on statistical bioinformatics. Since life-science consists of such a large area, I believe that this theme will continue: there is plenty of big data in this area.

The LASR workshops have provided a template on how to bring statisticians together and make them think on new emerging areas of science, especially by creating the workshop's format as a mixture of invited and contributed presentations plus posters. It has created a tradition of being informal and relaxed by encouraging interactions and discussions among participants and to have a good mix of young and experienced researchers representing both genders.

Mukhopadhyay: Kanti, what is your role in organizing LASR workshops now?

Mardia: My role has been continuing to be the Chairman of the workshop. The detailed task of organizing the workshop has become very smooth with key supports from my colleagues, including Robert Ackroyd, Paul Baxter, Stuart Barber, John Kent, and Arief Gusnanto. Also, there has been support from PhD students, including Jochen Voss, Chris Fallaize, and Anthony Riley in particular.

We have invaluable support from Dave Westhead, Professor of Bioinformatics. The funding situation has always been a sticky point, though LASR has been supported occasionally by EPSRC, LMS, RSS plus a select group of industries, including GSK.

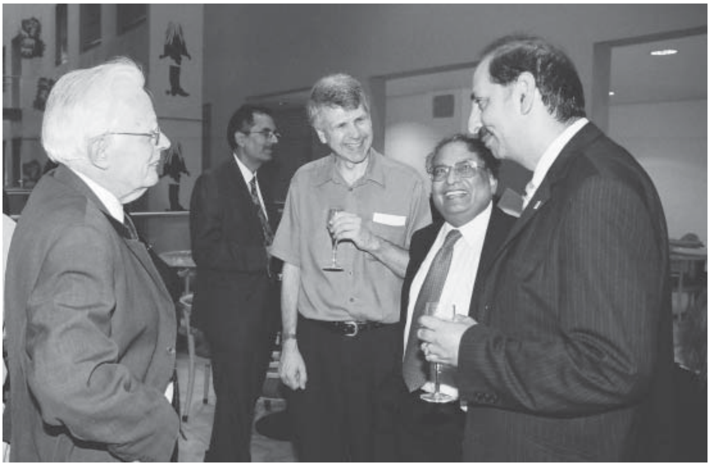


Figure 2.7 25th Anniversary celebration at the Royal Armouries, Leeds. In the foreground, from left to right: David Cox, John Kent, Kanti Mardia, and Councillor Mohammed Iqbal (Lord Mayor of Leeds).

Mukhopadhyay: Would you say that your original vision behind LASR has largely been attained? Please feel free to elaborate.

Mardia: For a number of years, we have focused partly on statistical bioinformatics in which the LASR workshops have been instrumental in bringing together different scientific communities under a single roof. However, importance of shapes, images, directional statistics, spatial statistics in other scientific areas still continue to be emphasized. I may mention that one theme in this year's LASR workshops falls squarely on "Non-Euclidean Statistical Models and Methods."

These Workshops started more than three decades ago as an annual event to foster interdisciplinary research in statistics in an emerging area of science. A special event was LASR's 25th anniversary celebration in 2006. We held the conference dinner at the Leeds Armouries. Most of these workshops were held in the idyllic setting at Hinsley Hall, Leeds, which was an original seventeenth-century monastery. We also produced a comprehensive leaflet describing the achievements of LASR and highlighting historical bioinformatics in Leeds; see http://www1.maths.leeds.ac.uk/statistics/workshop/LASRwebleaflet.pdf.

During these workshops, I had been reminding the gathering that some of the original pictures of X-ray similar to those used by Crick and Watson were available to William Astbury. Visit: http://www.leeds.ac.uk/heritage/Astbury/From_Wool_Fibres_to_DNA/index.html. He had the model for DNA as well as full protein, but these were wrong and, as it has been said, if he had collaborated with the School of Mathematics here, perhaps Leeds would have been the first to discover the models for DNA and protein. See also Hall (2014, Chapter 1).

Mukhopadhyay: Will you mention some of the distinguished visitors to LASR?

Mardia: There have been many distinguished bioinformaticians including Michael Levitt from Stanford University, California who was awarded a Nobel Prize in 2013 in chemistry. Statisticians included Terry Speed, David Cox, Peter Green, John Kingman, Bernard Silverman, David Hand, Wilfred Kendall, John Haslett, and the late Julian Besag. Thus, we have been very well supported by the statistical community.

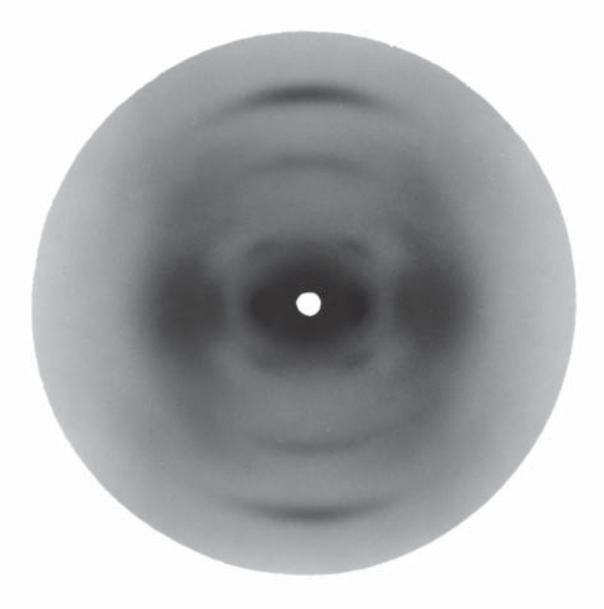


Figure 2.8 Very early X-ray picture (1930s) of collagen fiber (frog's toe tendon) from Astbury's laboratory, Leeds. Such X-ray images were instrumental in DNA discoveries by Watson and Crick (*Source:* from the personal collection of K. V. Mardia).

2.7 Memorable conference trips

Mukhopadhyay: Your work, accomplishments, and visibility have certainly continued to take you to visit many corners of the world since 1999. Please mention some of these trips.

Mardia: There have been visits to several conferences at different venues to give plenary lectures. The list includes visits to the University of Peradeniya (Kandy, Sri Lanka), Copenhagen University (Aarhus), University of Rome (Italy), Hong Kong, Research Triangle Institute (SAMSI, North Carolina), Banff Conference Center (Alberta, Canada), Granada and Valladolid (Spain), Beijing and Shanghai (People's Republic of China), Montreal (Canada), Florence (Italy), Berlin (Germany), and Brussels (Belgium).

Most of my talks have been on protein bioinformatics in order to make statisticians aware of this new and challenging area of research. I have also given a talk on the pleasure and pain of interdisciplinary research so that young researchers may become aware of both aspects.

Mukhopadhyay: How about most memorable trips and why were they memorable?

Mardia: Memorable trips — some of my trips have included sightseeing. Memorable trips were largely those which took me to new and exciting places for the first time such as Sri Lanka, Banff, Rome, and Beijing and Shanghai. It was fantastic to see the Great Wall of China. When I last saw David Kendall in October 2006 (a year before his death), he regarded the Great Wall of China as one of his memorable places. So, I always wanted to visit the Great Wall of China.

Figure 2.9 From left to right: Nobel Laureate Michael Levitt, Kanti Mardia, Fred Bookstein, and Clive Bowman. LASR 2008, Hinsley Hall, Leeds.

Figure 2.10 From left to right: Terry Speed, Philippa Burdett (PhD student), Arthur Lesk, Thomas Hamelryck, Kanti Mardia, and Chris Fallaize in a poster session during LASR 2010.

In Beijing IAMG conference, I selected my topic (Mardia 2007) as “Should Geostatistics Be Model-Based?” The audience was mostly from geosciences (nearly 500 people) and I decided to bring the model-based approach to geostatistics to their attention. Often, there are two streams in statistical research – one developed by practitioners and other by mainstream statisticians. Development of geostatistics is a very good example where pioneering work under realistic assumptions came from mining engineers (French School led by Matheron) whereas it is only now that the statistical framework is getting more transparent. Indeed, the subject with statistical emphasis has been maturing fast, as seen by various excellent books from the statistical side. The model-based approach is mainly based on my formulation as the spatial linear model. I presented first time the basic ideas of the model in 1980 to a geological conference in Paris followed by my full joint paper in *Biometrika* in 1984 (Mardia and Marshall 1984). These maximum likelihood estimates in the 1980s generated some debate but now through new research work by many, there is better understanding of their behavior and consequently their practical importance as I discussed in this paper (Mardia 2007). Of course, there will be a good coverage of these developments in my forthcoming book on Spatial Statistics with John Kent.

Also, a place like Valladolid was interesting because I could visit a town which was full of bookshops, more like Hay-on-Wye in the United Kingdom. Banff was especially great because it was so scenic with snow and ice on the hills. I particularly enjoyed the layout of the workshop held in Banff.

2.8 A select group of special colleagues

Mukhopadhyay: Kanti, after 1999, both Robin Plackett and David Kendall have passed on. I know that they influenced you greatly. Would you like to add your remarks in memory of their legacy?

Mardia: David Kendall pioneered in particular shape analysis and I will always remember him as someone tackling hard problems. His way of combining geometry in statistics and computation has been an influential pathway. He had a rigorous way of dealing with his submission of papers – if I recall correctly, once he submitted a paper to JRSS, B (Kendall 1984) but it was too long for the journal. So, David was asked to cut his paper's size which I am glad to say he refused. I also recall his paper for Geoff Watson's festschrift volume (Mardia 1992) which I was editing: I asked David to show more details on his key differential geometry ideas in the paper, but he politely declined saying that the details given there were adequate. I should mention that David Kendall was the PhD supervisor of my colleague, John Kent.

Robin Plackett was my PhD supervisor and mentor. I remember that his work always had been motivated by practical problems and he gave hints that one had to keep a watch on what other prominent researchers were working on such as that of David Cox. Unlike David Kendall, Plackett was not keen on doing his own computation and I recall helping with one of his computational problems in the 1960s. I think he was somewhat shy and the last I saw him was when he came to my seminar in the 1990s that I gave in Newcastle, but he left straight after my seminar. Before the seminar, he said how delighted he was to see my work in the emerging areas.

Mukhopadhyay: Will you mention a thing or two about some of your other special colleagues or collaborators?

Mardia: These years have been exciting from this point of view. I have already mentioned a few special collaborations. Coming back to Thomas Hamelryck, our combination has been good in that I provide, in some cases, appropriate statistical methodology to his problem in bioinformatics; he has then been able to incorporate the method in software in the public domain (PHAISTOS). In the process, I have collaborated with many of his postdocs and the strong collaboration continues.

I should mention my collaboration with my departmental colleagues, John Kent, Charles Taylor, Stuart Barber, and Jochen Voss. Most of the collaboration has been in statistical bioinformatics, though with John Kent our collaboration which started as early as 1977 still thrives. It is very difficult to clearly differentiate our style, but John has a special talent for getting to the heart of a problem analytically. If I recall correctly, the late Julian Besag once said to me that we are a formidable team.

Mukhopadhyay: How about some of your students?

Mardia: I have written a large body of papers with my PhD students who really helped in building the subject in spite of no substantive external funding. Since 2005, my PhD students who worked on statistical bioinformatics in Leeds include Mani Subramanium, Vysaul Nyirongo, Gareth Hughes, John Davies, Emma Petty, Kerstin Hommola, Zhengzheng Zhang, Chris Fallaize, and Anthony Riley. Pip Burdett (Leeds University) and Jinwoo Leem (Oxford University) are continuing PhD students as we speak.

Mukhopadhyay: Your thoughts on C. R. Rao, Ulf Grenander, David Cox, Fred Bookstein?

Mardia: This is a hard question since these scientists are all pioneers and original thinkers. I treat C. R. Rao, Ulf Grenander, and David Cox as my role models, and I hope I could continue working as they have been doing. Fred Bookstein has enormous energy and talent to cross boundaries. He is very prolific, dynamic, and quick on sharing his thoughts, and besides he is also approachable. Fred has been a regular supporter of LASR and he has attended the workshops continuously from 1991. He has found the LASR workshops to be a platform to air his unorthodox ideas in statistical science.

Peter Green stands out among new collaborators. He is extraordinarily quick to grasp new ideas and come up with new approaches for solutions. Besides, Peter also has a great talent in computational statistics. I believe that he is a superstar in statistical science.

2.9 High honors

Mukhopadhyay: Please tell me about the high honors bestowed upon you since 1999. The Guy Medal in Silver from the RSS and S. S. Wilks medal from the American Statistical Association (ASA) come to mind. Congratulations for those great honors. What were the corresponding citations?

Mardia: I was awarded the Silver Medal of the RSS in 2003. The citation for the award read: “The Guy Medal in Silver for 2003 is awarded to Professor Kanti Mardia for his many path-breaking contributions to statistical science, including two fundamental papers read to the Society on ‘Statistics of directional data’ (1975) and ‘A penalized likelihood approach to image warping’ (with C. A. Glasbey, 2001), his highly acclaimed monographs and his lasting leadership role in interdisciplinary research”.

Mukhopadhyay: Did the Guy Medal consider your full body of work too or other highly influential publications?

Mardia: This award depended mainly upon the discussion papers as cited.

Mukhopadhyay: How about the S. S. Wilks award?

Mardia: In August, 2013, the Samuel S. Wilks Memorial Medal was awarded to me by the ASA during the Joint Statistical Meetings held in Montreal, Canada. I was the fiftieth recipient.

The citation for the award read: “For extensive work covering a wide span of applied and theoretical research, including seminal results in shape analysis, spatial statistics, multivariate analysis, directional data analysis and bioinformatics with special applications to geostatistics, image analysis and protein structure; for the international dissemination of

Figure 2.11 Peter Green, the RSS President, presenting Silver Guy Medal of the Royal Statistical Society to Kanti Mardia, June18, 2003.

statistical thought and innovative ideas through research publications, presentations, books, monographs, the establishment and running of annual research workshops and interdisciplinary centers; and for his insightful guidance of future generations of statisticians".

Mukhopadhyay: What were the requirements for the nomination of the S. S. Wilks award?

Mardia: The requirements for the nominations were: "The Wilks Memorial Award is bestowed upon a distinguished individual who has made statistical contributions to the advancement of scientific or technical knowledge, ingenious application of existing knowledge, or successful activity in the fostering of cooperative scientific efforts that have been directly involved in matters of national defense or public interest ...".

Mukhopadhyay: It seems to me that this was the first time the S. S. Wilks medal went to someone living in United Kingdom. How so in your view?

Mardia: It is difficult to pin-point exactly why I got this award, but my guess is that it is given for the cumulative research contribution with possibly my kind of special brand of leadership in interdisciplinary research, and for moving Leeds and the UK forward in geosciences, image analysis, and bioinformatics. Also, I may add that I began working when these subjects were just emerging so that the LASR workshops helped tremendously in crossing the research boundaries at the right time.

Mukhopadhyay: What is your unique philosophy on honors and life in general?

Mardia: I wrote a parody on a verse from Thomas Gray's "Elegy written in a Country Churchyard":

"The pomp of professorship, boast of medal, All that publication and breakthrough e'er gave, Await alike th'inevitable hour. The paths of glory (also) lead but to the grave."

This is also complemented by the famous quote from Jain thinking:

"Kashaya muktih kil muktirev,"

meaning freedom from destructive emotion is in reality the only way to true enlightenment. This line becomes clearer when seen in the context of the transliteration of the full stanza as follows:

Figure 2.12 Samuel S. Wilks Memorial Medal presented in Montreal, Canada to Kanti Mardia by Marie Davidian, President of the American Statistical Association, August 4, 2013.

"Neither by wearing white robes nor by wearing nothing, Neither by logical discussion nor by metaphysical discourse, Nor is there liberation by adopting a particular theology Liberation comes only by liberating oneself from Kashaya".

2.10 Statistical science: thoughts and predictions

Mukhopadhyay: Where is statistical science going?

Mardia: I think the most important development has been the computational power which has changed attitudes of statistical scientists to move from small sample statistics to large-scale statistics. What was not feasible before as a practical methodology has now become a reality.

There is a significant rise of Bayesian methods. I think the next few years will still see the impact of statistics on new data coming from advances in technology. The greatest challenge for statisticians is to remain "ahead" of computer scientists, that is, to have substantial computational skills combined with sound statistical principles and techniques.

Mukhopadhyay: Kanti, you attend numerous international conferences. How do many nonstatisticians tend to address statistical aspects and do you feel comfortable with present status? What is your mission?

Mardia: Let me mention a quote related to our image problem (Mardia and Gilks 2005):

"In conferences we have attended where statistics is not the main topic, such as image analysis and bioinformatics, one sees the use of terms like data processing, data analysis, prediction, estimation, hypothesis, significance, etc., but the discipline of statistics is rarely acknowledged. Upon asking presenters why they have not made use of statistical methods, we have received such bizarre replies as: 'Statistics deals only with small datasets, but our problem is for a very large amount of data,' or 'Oh yes, we use statistics only when we want to calculate a measure of uncertainty,' or 'Statistical models deal only with observables, but we need models which consider also unobserved variables.' "

I have not taught now for nearly 15 years except supervising PhD students. Before then, as Head of Department, my attitude was taken from Edward Boyle, the Vice-Chancellor here, when I joined Leeds University. Boyle's maxim was to do teaching in the atmosphere of research so that they fed each other, especially at MSc level. Most of my research problems arose from interdisciplinary projects, and the LASR workshops have been a great platform in which to have interdisciplinary dialogue, and to keep up with the changing nature of statistical science itself. I believe that LASR has been a big force in many ways. To tell you the truth, the LASR embodies my mission: "Statistics without science is incomplete, Science without statistics is imperfect."

Mukhopadhyay: Where do you believe statistical science will be in 10 or 20 years from now?

Mardia: Nitis, since my last published conversation with you, research life has already become easier when I consider accessibility to literature and pdf files, e-books, and substantially more available data on the web.

It is not possible for me to exactly foresee what will be the development of statistics in the next 20 years. However, I may make some predictions based on the developments that have happened in the last decade or so.

In recent years, new methods of acquiring "big" data have become available in many fields such as medicine, genetics, engineering, management, and these have led to inception of new statistical methods for analysis of data.

Mukhopadhyay: Please share some of your personal experiences.

Mardia: When I began working on image analysis, those images were very coarse. But, over time with improvements in technology, images have become significantly sharper. The analysis of images over the years has given rise to a large number of problems which required development of new statistical methods.

For example, problems of object recognition, classification, and discrimination with image data have led to development of exciting new statistical methods. For the purpose of object recognition, obviously shape is an important attribute. After David Kendall and Fred Bookstein gave initial ideas of how to quantify shape (Bookstein 1986; Kendall 1984; Kendall 1989), a large body of work has been created by me and others, especially regarding the distributions of shapes and analysis of shape data. The Mardia–Dryden distribution (Kendall 1989) is now regarded as an important probability model for the distribution of shapes.

Mukhopadhyay: Kanti, did you have similar personal experiences when you turned your research toward proteomics?

Mardia: Nitis, yes, of course. Very little statistical work had been done in this field before the turn of the century as there was very little communication between the biologists and the statisticians. With better communication now in place, statisticians have been able to participate in dealing with challenging problems and studying huge amounts of data – mostly multivariate and high dimensional in nature – emanating from this field. Again, analysis of these data is triggering development of new and substantial methods of statistical analysis.

Mukhopadhyay: Kanti, is there anything else you want to add on this topic?

Mardia: Real data arising from any field, where statisticians' expertise is crucial, are often complex and large in size. They are often difficult to analyze since they contain outliers, have multimodality, show skewness, and consist of a mixture of discrete, continuous, ordinal, and nominal measurements. Customary statistical methods may not always answer questions arising from such data.

Further, I feel that statisticians should go beyond writing purely mathematical papers. Instead, attention may be diverted into development of applicable solutions, algorithms, plus computer programs and software. They should make the new methods so developed easily accessible to scientists and practitioners of other disciplines.

I think good statistical science will be more computational and the new research methodologies will have available web-based software, more libraries in R, and more data in the public domain. More support for interdisciplinary teams will be essential for new breakthroughs.

Mukhopadhyay: But, now, if everyone moves away from pure mathematics, there is clear danger that we will be faced with numerous methodologies without solid theoretical foundations which will surely create a kind of anarchy. To be frank, we are already seeing some of that in the horizon. Are we not? So, let me ask you to clarify your stance. Will you please?

Mardia: My paradigm is: New Questions ↔ New Data ↔ New Methods. I think that the future trend in statistics will be a hybrid of model-based statistics and algorithmic statistics.

Walter Gilks and I wrote the article (Mardia and Gilks 2005) entitled "Meeting the statistical needs of 21st-century science" in *Significance*, December 2005 issue, which said:

"... we propose a holistic approach to statistics. Holistic medicine ('alternative' medicine) treats the patient as a whole rather than targeting just the affected part. The same philosophy applied to statistical practice suggests that one should set clients' problems in the context of their priorities, available data and methods, current scientific knowledge, computational considerations, risk assessment and required end-product. Bayesian or frequentist, exploratory or predictive, all provide different types of cure!"

It concludes with: "Through our brief account, we have identified three themes. First, statistics should be viewed in the broadest possible way for scientific explanation or prediction of any phenomenon. Second, the future of statistics lies in a holistic approach to interdisciplinary research. Third, a change of attitude is required by statisticians – a paradigm shift – for the subject to go forward."

Mukhopadhyay: In 1999, you told me that you were a pragmatic Bayesian. What did that mean? Are you more (or less) of a Bayesian now than you were in 1999?

Mardia: Still I remain open-minded and whatever statistical tool works, I go for it. I will remain a pragmatic Bayesian.

2.11 Immediate family

Mukhopadhyay: Kanti, shall we talk about your immediate as well as extended family?

Mardia: Inevitably, both sad and happy events have taken place in the family over the past few years. One of my elder brothers (Jawerchand) died. The eldest brother (Mangeshkumar) reached the age of 90 this year, and my mother-in-law has reached the wonderful age of 103. My wife, Pavan, and I are both fortunate to have good health so far. We celebrated our Golden Wedding anniversary in 2008. We both travel together to most conferences and when I spend time on my visiting professorships. Incidentally, we complete 50 years in this country in September this year.

Other great family news since 2000 is that my eldest daughter has Sashin, our second grandson (now 11 years old), and my first grandson Ashwin has just started a degree in economics at York University. There is almost zero probability of more grandchildren but we look forward to great-grandchildren.

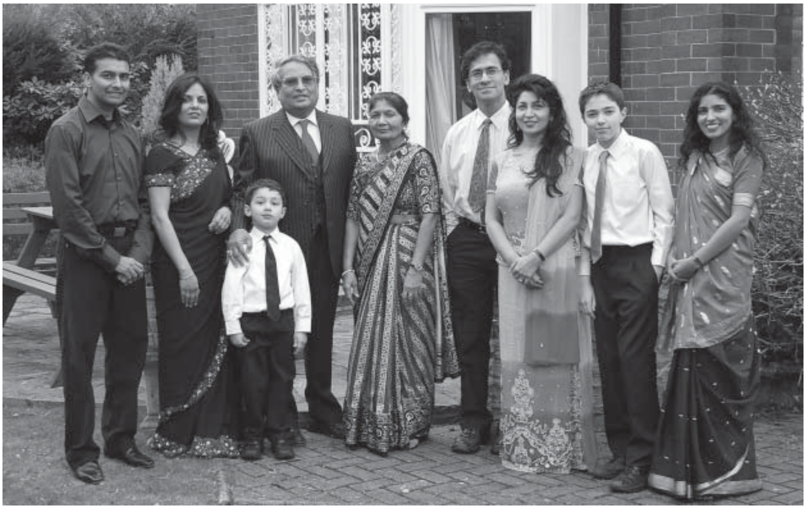


Figure 2.13 The Golden Wedding Anniversary 2008: Family photo (from left to right) shows Hemant, Preeti, Sashin (in front), Kanti, Pavan, Ragunath, Bela, Ashwin, and Neeta.

Figure 2.14 Sashin reciting four noble truths during the launch of "Living Jainism," Jain Temple, Leicester, July 15, 2013.

Mukhopadhyay: How about your son, Hemant, and daughters, Bela and Neeta? Where are they now and what are they doing?

Mardia: My son, who was living earlier very near to us, was the CEO of a publicly quoted company (Filtronic) until 2013. He has moved approximately nine months ago to Norway as the CEO of a new company (Idex), specializing in fingerprint imaging and fingerprint recognition technology. His wife, Preeti, has a key role in the company, while also doing an M.Sc. in Business Management.

My younger daughter, Neeta, had a tough time in heading her legal firm and she took a sabbatical from her profession to fulfill successfully her dream of inventing and marketing "sweet samosas" (under the banner of "Sweet Karma"). She is joined in this by her partner, Jon Handley, and is growing it successfully in her spare time as she is back in her professional work as a lawyer.

Currently, my elder daughter, Bela (who manages her several properties) and her husband Raghunath (who is a medical doctor), live in Hull, meaning about one hour's drive to and from Leeds. Neeta and Hemant live far away, so that in an emergency, we have to rely on each other (my wife and myself) and our friends, Raj, Meena and Tilak in Leeds.

Mukhopadhyay: What are some of your hobbies now?

Mardia: There has been no time to play either chess or bridge except sometimes I play chess with my grandchildren.

My previous hobby of collecting antiquarian books has come to an end. More concentration has been on books which I read including those related to Jainism. Yoga exercises and health club visits are new additions to my hobbies since we spoke last.

2.12 Jain thinking

Mukhopadhyay: How is your work in Jain thinking going on?

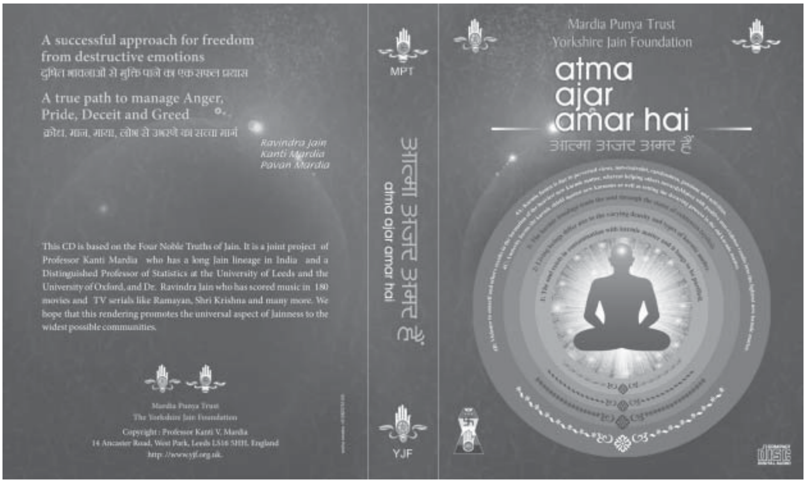
Mardia: My "The Scientific Foundations of Jainism" has become a classic. It has now been translated into Hindi (2004) and Gujarati (2012). The axiomatic system of the Four Noble Truths of Jains devised by me is described in this book. In 2013, a Bollywood music director, Ravindra Jain, composed songs based on Four Noble Truths, and this album was launched in Mumbai, India in January 2014 under the media glare (Jain et al. 2014).

These also provided a basis for my new book *Living Jainism* (Mardia and Rankin 2013) written with Aidan Rankin. In it, we addressed the unique nature and teachings of Jainism, the unity of life with which Jainism began, the implications of that unity, and the need to reorient our behavior accordingly.

Mukhopadhyay: How about other related activities in the north of England?

Mardia: I continue as the Chairman of the Yorkshire Jain Foundation (YJF), which we founded in 1987. One of the major contributions has been having a Jain temple (part of a Hindu Temple) in Leeds from 2001; the Jain temple is a focus for Jains in the north of England but receives many visitors including from schools learning comparative religions. As there are a very few Jains around in Yorkshire, the activities of the YJF have been also at the national and international levels.

During the last 20 years when I found time, I wrote on the subject of "kashayas," which approximately stands for destructive emotions. The book will highlight chronological thinking of Jains on this topic starting from time immemorial. Indeed we hope to spend more time on the YJF and Mardia Punya Trust. In 2004, the Vanik Association (a community of


A successful approach for freedom from destructive emotions
दूषित भावनाओं से मुक्ति पाने का एक सफल प्रयास
A true path to manage Anger, Pride, Deceit and Greed
क्रोध, मान, माया, लोभ से उभरने का सच्चा मार्ग
Ravindra Jain
Kanti Mardia
Pavan Mardia
MPT
आत्मा अजर अमर हैं
atma ajar amar hai
YJF
Mardia Punya Trust
Yorkshire Jain Foundation
atma ajar amar hai
आत्मा अजर अमर हैं

This CD is based on the Four Noble Truths of Jain. It is a joint project of Professor Kanti Mardia who has a long Jain lineage in India and a Distinguished Professor of Statistics at the University of Leeds and the University of Oxford, and Dr. Ravindra Jain who has scored music in 180 movies and TV serials like Ramayan, Shri Krishna and many more. We hope that this rendering promotes the universal aspect of Jainness to the widest possible communities.

Mardia Punya Trust
The Yorkshire Jain Foundation
Copyright : Professor Kanti V. Mardia
14 Ancaster Road, West Park, Leeds LS16 5HH, England
http://www.yjf.org.uk.

Figure 2.15 Cover of an album launched in January 2014, in Mumbai. Lyrics are based on the Four Noble Truths formulated by Kanti Mardia in 1978. Musical compositions: Ravindra Jain, Bollywood celebrity.

professionals in the UK) honored in appreciation of my contribution in education as well as in Jainism. This was during their silver jubilee anniversary celebration year.

Mukhopadhyay: In January 2007, I recall that you met the President of India, His Excellency Dr. Abdul Kalam, and discussed with him the topic of “Statistics, Science, and Spirituality.” Could you please elaborate?

Mardia: Nitis, thanks for bringing this up. Someone arranged a brief interview with the President of India, His Excellency Dr. Abdul Kalam, for 5 minutes. However, our discussion continued for half hour.

Mukhopadhyay: What did you two discuss?

Mardia: One topic was the Four Noble Truths that I had produced for Jainism. Of course, he was aware that Jainism influenced Mahatma Gandhi in his principle of nonviolence. It is claimed to be a religion and Kalam knew Jain philosophy and its pre-Vedic origin, but obviously he was not aware of my work.

Also, we discussed how Jain ideas foreshadowed statistics in relation to the principle of inference from samples to population, the idea of meta-analysis (anekantvad = many-sided view), nonabsolutism in scientific discovery (syadvad). Dr. Kalam himself is a scientist interested in philosophy. In fact, he recently wrote a joint book with a very important Jain guru (the late Mahapragna). We both believe that divine values are necessary in science and religion. Einstein summarized this beautifully: “Science without Religion is lame, Religion without Science is blind”.

2.13 What the future may hold

Mukhopadhyay: What are some of your small or big plans in the future?

Mardia: I want to see our spatial statistics book finished and the second editions of my three previous books done. I hope to keep up-to-date with at least my own areas of expertise. I need to learn more R.

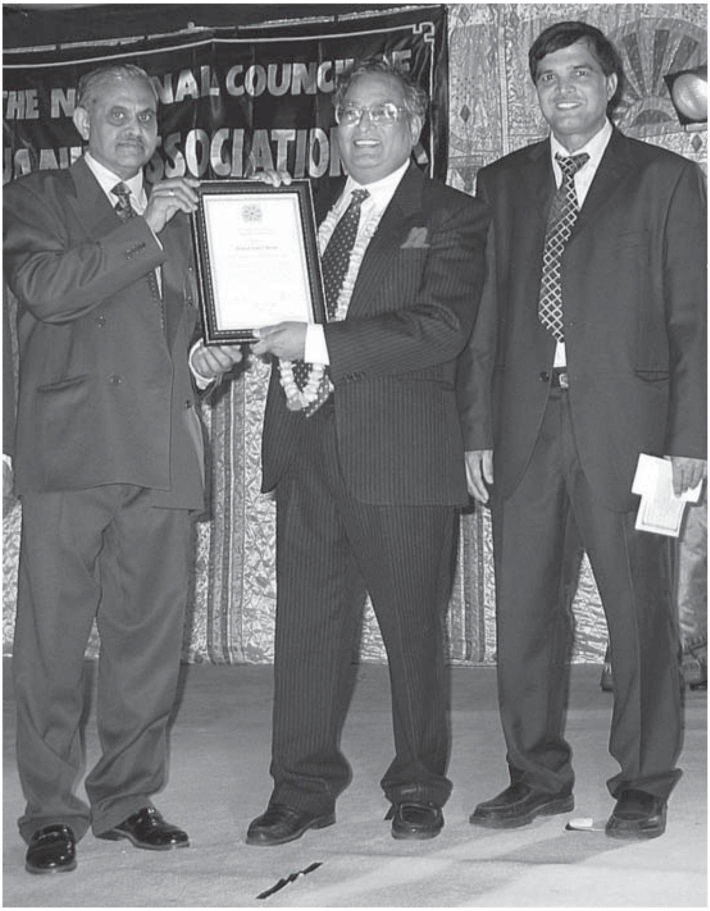


Figure 2.16 Best professional award to Kanti Mardia (center) by the National Council of Vanik Associations, UK, during its silver jubilee celebration, April 25, 2004.

I would like to lend my helping hands to support "LASR-type" workshops. I intend to push geometry-driven statistics with real problems arising from life sciences. Statistics on manifolds is now recognized as a very important area of applications and it is growing fast – curves, surfaces, and non-Euclidean data. A key problem is to invent the right type of models for such manifolds. As in directional statistics and shape analysis, here again it is expected that the normalization constants would be intricate, as will be the associated inference.

Mukhopadhyay: What else may be on your plate?

Mardia: I hope to spend more time in Oxford and IIMA although my base will remain in Leeds, especially since we have been here since 1973 and our very close community life is here. As long as my brain and body will allow, I will continue to be active.

Mukhopadhyay: You have been a truly optimal role model for me. I certainly wish wholeheartedly that your wonderful wishes will all come true.

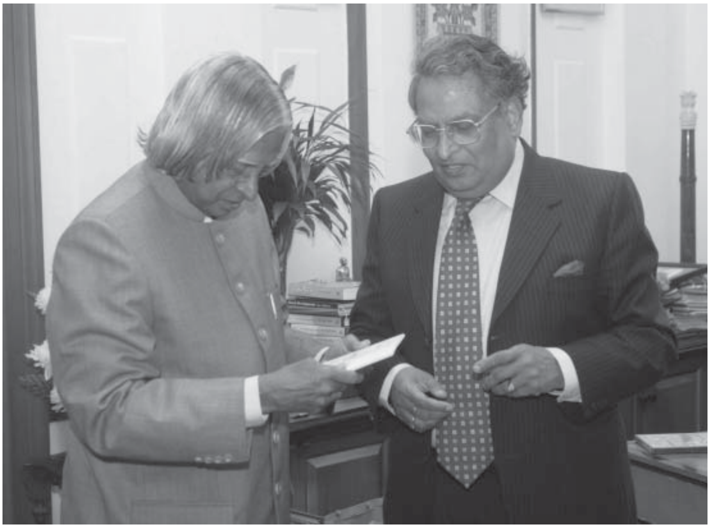


Figure 2.17 Kanti Mardia met the President of India, His Excellency Dr. Abdul Kalam. The Presidential Palace in Delhi, January 3, 2007.

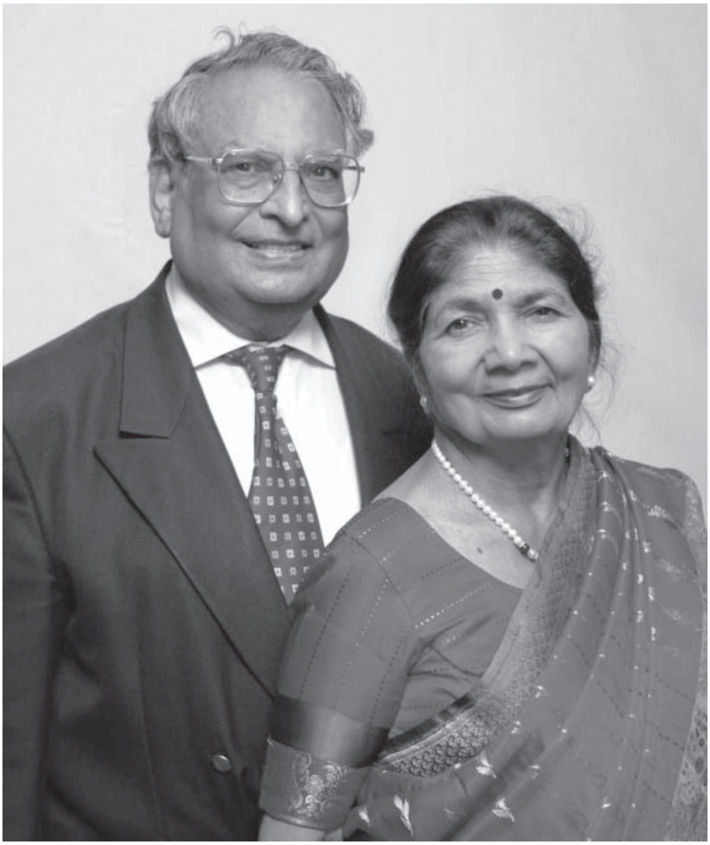


Figure 2.18 Kanti and Pavan Mardia. Leeds, August 19, 2014.

Mardia: Nitis, thanks a lot, but everything also depends on Pavan's health as I depend on her completely in day-to-day life. I mentioned in my last published conversation that my wife is the hardware if I am the software. This is more so now.

Mukhopadhyay: Kanti, I have to close this conversation with some final heartfelt words of appreciation. Thank you so much for giving me the rare opportunity to chat you not once, but twice, in 1999 and again in 2014. I pray for continued best of health and happiness for you, Pavan and the rest of your family. I have no idea how you preserve your youth and bubbling energy. Whatever is your secret, please do not change a thing.

Wish you the happiest 80th birthday in 2015. Rest assured, I will be ready to have a chat with you yet again in 2024-25 prior to your 90th birthday celebration. God bless, my friend.

Acknowledgment

Thanks to the Mardia family for kindly providing all the photos that I have used in this article.

References

Bookstein FL 1986 Size and shape spaces for landmark data in two dimensions (with discussion). *Statistical Science* **1**, 181–242.

Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A and Hamelryck T 2006 A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 8932–8937.

Deane CM, Dunbar J, Fuchs A, Mardia KV, Shi J and Wilman HR 2013 Describing protein structure geometry to aid in functional understanding In *LASR2013 Proceedings — Statistical Models and Methods for non-Euclidean Data with Current Scientific Applications* (ed. Mardia KV, Gusnanto A, Riley AD and Voss J), pp. 49–51. Leeds University Press.

Demchuk E, Hnizo V, Mardia KV, Sharp DS and Singh H 2001 Statistics and molecular structure of biological macromolecules In *LASR2001 Proceedings — Functional and Spatial Data Analysis* (ed. Mardia KV and Aykroyd RG), pp. 9–14. Leeds University Press.

Dryden IL and Mardia KV 1998 *Statistical Shape Analysis*. John Wiley & Sons, Ltd, Chichester.

Glasbey CA and Mardia KV 2001 A penalized likelihood approach to image warping (with discussion). *Journal of the Royal Statistical Society, Series B* **63**, 465–514.

Green PJ and Mardia KV 2006 Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika* **93**, 235–254.

Hall KT 2014 *The Man in the Monkeynut Coat: William Astbury and the Forgotten Road to the Double-Helix*. Oxford University Press.

Hamelryck T, Mardia KV and Ferkinghoff-Borg J 2013 *Bayesian Methods in Structural Bioinformatics*. Springer-Verlag, New York.

Jain R, Mardia KV and Mardia PK 2014 *Atma Agar Amar Hai*. D.R. Productions, Mumbai.

Kendall DG 1984 Shape manifolds, procrustean metrics and complex projective spaces. *Bulletin of London Mathematical Society* **16**, 81–121.

Kendall DG 1989 A survey of the statistical theory of shape. *Statistical Science* **4**, 87–120.

Kent JT and Mardia KV 2012 A geometric approach to projective shape and the cross-ratio. *Biometrika*, **99**, 833–849.

Mardia KV 1992 *The Art of Statistical Science, in Honor of G. S. Watson*. John Wiley & Sons, Inc., New York.

Mardia KV 2007 Should geostatistics be model-based? *Proceedings of the IAMG 2007 Conference: Geomathematics and GIS Analysis of Resources, Environment and Hazards, Beijing, China*, pp. 4–9.

Mardia KV 2013 Statistical approaches to three key challenges in protein structural bioinformatics. *Journal of the Royal Statistical Society, Series C* **62**, 487–514.

Mardia KV and Gilks WR 2005 Meeting the statistical needs of 21st-century science. *Significance* **2**, 162–165.

Mardia KV and Hamelryck T 2012 Discussion to "Constructing summary statistics for Approximate Bayesian Computation: semi-automatic approximate Bayesian computation" by P. Fearnhead and D. Prangle.. *Journal of the Royal Statistical Society, Series B* **74**, 462–463.

Mardia KV and Jupp PE 2000 *Directional Statistics*. John Wiley & Sons, Ltd, Chichester.

Mardia KV and Marshall RJ 1984 Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135–146.

Mardia KV and Patrangenaru V 2005 Directions and projective shapes. *Annals of Statistics* **33**, 1666–1699.

Mardia KV and Rankin AD 2013 *Living Jainism, An Ethical Science*. Mantra Books, Winchester.

Mardia KV, Bookstein FL and Kent JT 2013 Alcohol, babies and the death penalty: saving lives by analyzing the shape of the brain. *Significance* **10**, 12–16.

Mardia KV, Bookstein FL, Kent JT and Meyer CR 2006 Intrinsic random fields and image deformations. *Journal of Mathematical Imaging and Vision* **26**, 59–71.

Mardia KV, Borg M, Ferkinghoff-Borg J and Hamelryck T 2012a Towards a general probabilistic model of protein structure: the reference ratio method In *Bayesian Methods in Structural Bioinformatics* (ed. Hamelryck T, Mardia KV and Ferkinghoff-Borg J) Springer-Verlag New York pp. 125–134.

Mardia KV, Hughes G, Taylor CC and Singh H 2008 A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics* **36**, 99–109.

Mardia KV, Kent JT and Bibby JM 1979 *Multivariate Analysis*. Academic Press, New York.

Mardia KV, Nyirongo VB, Fallaize CJ, Barber S and Jackson RM 2011 Hierarchical Bayesian modelling of pharmacophores in bioinformatics. *Biometrics* **67**, 611–619.

Mardia KV, Petty EM and Taylor CC 2012b Matching markers and unlabeled configurations in protein gels. *Annals of Applied Statistics* **6**, 853–869.

Mukhopadhyay N 2002 A Conversation with Kanti Mardia. *Statistical Science* **17**, 113–148.

Pardo-Igúzquiza E, Mardia KV and Chica-Olmo M 2009 MLMATERN: a computer program for maximum likelihood inference with the spatial Matérn covariance model. *Computers & Geosciences* **35**, 1139–1150.

Sahu SK and Mardia KV 2005 A Bayesian Kriged Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society, Series C* **54**, 223–244.

Sahu SK, Jona Lasini G, Orasi A and Mardia KV 2005 A comparison of spatio-temporal Bayesian models for reconstruction of rainfall fields in a cloud seeding experiment. *Journal of Mathematics and Statistics* **1**, 273–281.

3

Selected publications

K V Mardia

1. Monographs

Families of Bivariate Distributions (K.V. Mardia). Charles Griffin, London, 1970.

Statistics of Directional Data (K.V. Mardia). Academic Press, New York, 1972. Translated in Russian.

Tables of the F- and Related Distributions with Algorithms (K.V. Mardia and P.J. Zemroch). Academic Press, London and New York, 1978.

Multivariate Analysis (K.V. Mardia, J.T. Kent and J.M. Bibby). Academic Press, New York, 1979. Translated in Persian.

Statistical Shape Analysis (I.L. Dryden and K.V. Mardia). John Wiley & Sons, Ltd, Chichester, 1998.

Directional Statistics, 2nd edition (K.V. Mardia and P.E. Jupp). John Wiley & Sons, Ltd, Chichester, 2000.

2. Edited Volumes

Statistics in Earth Sciences (Ed. K.V. Mardia). A special issue of the Communications in Statistics **A10**(15), 1989.

Statistical Methods in Image Processing (Ed. K.V. Mardia). A special issue of the Journal of Applied Statistics **16**(2), 1989.

The Art of Statistical Science (Ed. K.V. Mardia). In honor of G.S. Watson. John Wiley & Sons, Inc., New York, 1992.

Statistics and Images: Vol. I, (Eds. K.V. Mardia and G. Kanji). Carfax Publishing, Abingdon, Oxfordshire, 1993.

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

Statistics and Images: Vol. II. (Eds. K.V. Mardia and G. Kanji) Carfax Publishing, Abingdon, Oxfordshire, 1994.

Current Issues in Statistical Shape Analysis, (Eds. K.V. Mardia and C.A. Gill). Proceedings: Leeds University Press, 1995.

Image Fusion and Shape Variability Techniques, (Eds. K.V. Mardia, C.A. Gill and I.L. Dryden), Leeds University Press, 1996.

The Art and Science of Bayesian Image Analysis, (Eds. K.V. Mardia, C.A. Gill and R.G. Aykroyd). Leeds University Press, 1997.

Medical Image Understanding and Analysis '98, (Eds. K.V. Mardia, E. Berry, D.C. Hogg and M.A. Smith). British Machine Vision Association Publication, 1998.

Spatial Temporal Modelling and its Applications, (Eds. K.V. Mardia, R.G. Aykroyd and I.L. Dryden). Proceedings: Leeds University Press, 1999.

Functional and Spatial Data Analysis, (Eds. K.V. Mardia and R.G. Aykroyd). Proceedings: Leeds University Press, 2001.

Statistics of Large Datasets, (Eds. K.V. Mardia, R.G. Aykroyd and P. McDonnell). Proceeings: Leeds University Press, 2002.

Stochastic Geometry, Biological Structure and Images, (Eds. K.V. Mardia, R.G. Aykroyd and M.J. Langdon). Proceedings: Leeds University Press, 2003.

Bioinformatics, Images, and Wavelets, (Eds. K.V. Mardia, R.G. Aykroyd and S. Barber). Proceedings: Leeds University Press, 2004.

Quantitative Biology, Shape Analysis, and Wavelets, (Eds. K.V. Mardia, S. Barber, P.D. Baxter and R.E. Walls). Proceedings: Leeds University Press, 2005.

Interdisciplinary Statistics and Bioinformatics, (Eds. K.V. Mardia, S. Barber, P.D. Baxter and R.E. Walls). Proceedings: Leeds University Press, 2006.

Systems Biology & Statistical Bioinformatics, (Eds. K.V. Mardia, S. Barber and P.D. Baxter). Proceedings: Leeds University Press, 2007.

The Art and Science of Statistical Bioinformatics. LASR 2008 Proceedings. (Eds. S. Barber, P.D. Baxter, A. Gusnanto and K.V. Mardia), Leeds University Press, 2008, pp. 46–46.

Statistical Complexity in Protein Bioinformatics. LASR 2009 Proceedings. (Eds. A. Gusnanto, K.V. Mardia and C.J. Fallaize), Leeds University Press, 2009, pp. 9–20.

High Throughput Sequencing, Proteins and Statistics. LASR 2010 Proceedings. (Eds. A. Gusnanto, K.V. Mardia, C.J. Fallaize and J. Voss), Leeds University Press, 2010.

Next Generation Statistics in Biosciences. LASR 2011 Proceedings. (Eds. K.V. Mardia, A. Gusnanto, A.D. Riley and J. Voss), Leeds University Press, 2011.

New Statistics and Modern Natural Sciences. (Eds. K.V. Mardia, A. Gusnanto, A.D. Riley and J. Voss) (2012) Leeds University Press.

Bayesian Methods in Structural Bioinformatics. (Eds. T. Hamelryck, K.V. Mardia and J. Ferkinghoff-Borg). Springer-Verlag.

Statistical Models and Methods for non-Euclidean Data with Current Scientific Applications. LASR 2013 Proceedings. (Eds. K.V. Mardia, A. Gusnanto, A.D. Riley and J. Voss). Leeds University Press, 2013.

3. Journal Research Papers

1962 Mardia, K.V. Multivariate Pareto distributions. *Annals of Mathematical Statistics* **33**, 1008–1015.

1964 Mardia, K.V. Some results on the order statistics of the multivariate normal and Pareto-Type 1 populations. *Annals of Mathematical Statistics* **35**, 1815–1818.

1967 Mardia, K.V. Correlation of the ranges of correlated samples. *Biometrika* **54**, 529–539.

1967 Mardia, K.V. A non-parametric test for the bivariate two-sample location problem. *Journal of the Royal Statistical Society Series B* **29**, 320–342.

1967 Mardia, K.V. Some contributions to contingency-type bivariate distributions. *Biometrika* **54**, 235–249. [Corrections in Biometrika, 1968, Vol.55, p597].

1969 Mardia, K.V. On Wheeler-Watson's two-sample test on a circle. *Sankhyā A* **31**, 177–190.

1969 Mardia, K.V. The performance of some tests of independence for contingency-type bivariate distributions. *Biometrika* **56**, 449–451.

1970 Mardia, K.V. A translation family of bivariate distributions and Frechet bounds. *Sankhya A* **32**, 119–122.

1970 Mardia, K.V. Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530.

1971 Mardia, K.V. The effect of non-normality on some multivariate tests and robustness to non-normality in the linear model. *Biometrika* **58**, 105–121.

1972 Mardia, K.V. A multisample uniform scores test on a circle and its parametric competitor. *Journal of the Royal Statistical Society Series B* **34**, 102–113.

1973 Phatarfod, R.N. and Mardia, K.V. Some results for dams with Markovian inputs. *Journal of Applied Probability* **10**, 166–180.

1974 Mardia, K.V. Applications of some measures of multivariate skewness and kurtosis in testing multinormality and robustness studies. *Sankhyā B* **36**, 115–126.

1975 Mardia, K.V. and Zemroch, P.J. Circular statistics. AS81. *Journal of the Royal Statistical Society Series C* **24**, 147–150.

1975 Mardia, K.V. and Zemroch, P.J. Spherical statistics. AS80. *Journal of the Royal Statistical Society Series C* **24**, 144–146.

1975 Mardia, K.V. and Zemroch, P.J. The von Mises distribution function. AS86, *Journal of the Royal Statistical Society Series C* **24**, 268–272.

1975 Mardia, K.V. Statistics of directional data (with Discussion). Paper presented to the Royal Statistical Society. *Journal of the Royal Statistical Society Series B* **37**, 349–393.

1975 Mardia, K.V. and Sutton, T. On the modes of a mixture of two von Mises distributions. *Biometrika* **62**, 699–701.

1976 Mardia, K.V. and El-Atoum, S.A.M. Bayesian inference for the von Mises-Fisher distribution. *Biometrika* **63**, 203–206.

1976 Mardia, K.V. Do-it-yourself statistical analysis. Inaugural Address. Leeds Review, pp. 79–98.

1976 Mardia, K.V. Linear-circular correlation coefficients and rhythmometry. *Biometrika* **63**, 403–405.

1976 Mardia, K.V. and Spurr, B.D. On some tests for the bivariate two-sample location problem. *Journal of the American Statistical Association* **71**, 994–995.

1977 Mardia, K.V. and Gadsden, R.J. A small circle of best-fit for spherical data and areas of vulcanism. *Journal of the Royal Statistical Society Series C* **26**, 238–245.

1977 Khatri, C.G. and Mardia, K.V. The von Mises-Fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society Series B* **34**, 95–106.

1977 Mardia, K.V. and Khatri, C.G. The uniform distribution on Stiefel manifold. *Journal of Multivariate Analysis* **7**, 468–473.

1978 Mardia, K.V. and Sutton, T. A model for cylindrical variables with applications. *Journal of the Royal Statistical Society Series B* **40**, 229–233.

1978 Mardia, K.V. and Puri, M.L. A spherical correlation coefficient robust against scale. *Biometrika* **65**, 391–395.

1978 Jupp, P.E. and Mardia, K.V Density preserving statistics and densities for sample means. *Annals of Probability* **6**, 688–694.

1978 Bingham, C. and Mardia, K.V. A small circle distribution on the sphere. *Biometrika* **65**, 379–389.

1978 Mardia, K.V. Some properties of classical multi-dimensional scaling. *Communications in Statistics - Theory and Methods* **47**, 1233–1241.

1979 Kent, J.T., Mardia, K.V. and Rao, J.S. A characterization of the uniform distribution on the circle. *Annals of Statistics* **7**, 882–889.

1979 Jupp, P.E. and Mardia, K.V. Maximum likelihood estimators for the matrix von Mises-Fisher and Bingham distributions. *Annals of Statistics* **7**, 599–606.

1980 Mardia, K.V. and Holmes, D. A statistical analysis of megalithic data under elliptic pattern. *Journal of the Royal Statistical Society Series A* **143**, 293–302.

1980 Jupp, P.E. and Mardia, K.V. A general correlation coefficient for directional data and related regression problems. *Biometrika* **67**, 163–174. [Amendments and Corrections in Biometrika, 1981, vol. 68, p738].

1982 Jupp, P.E. and Mardia, K.V. A characterization of the multivariate Pareto distribution. *Annals of Statistics* **10**, 1021–1024.

1982 Mardia, K.V. and Edwards, R. Weighted distributions and rotating caps. *Biometrika* **69**, 323–330.

1983 Jupp, P.E. and Mardia, K.V. A note on the maximum-entropy principle. *Scandinavian Journal of Statistics* **10**, 45–47.

1983 Kent, J.T., Briden, J.C. and Mardia, K.V. Linear and planar structure in ordered multivariate data as applied to progressive demagnetization of paleomagenetic remanence. *Geophys. Journal of the Royal Astronomical Society* **75**, 593–621.

1984 “spatial discrimination and classification maps”. *Communications in Statistics* **13**, 2181–2197. Corrections in “Spatial discrimination & classification maps” *Communications in Statistics* (1987) **A, 16**.

1984 Mardia, K.V. and Marshall, R.J. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135–146.

1984 Mardia, K.V., Holmes, D. and Kent, J.T. A goodness-of-fit test for the von Mises-Fisher distribution. *Journal of the Royal Statistical Society Series B* **46**, 72–78.

1985 Mardia, K.V. and Marshall, R.J. Some minimum norm quadratic estimators of the components of spatial covariance. *Journal of the International Association for Mathematical Geology* **17**, 517–525.

1988 Mardia, K.V. and Hainsworth, T.J. A spatial thresholding method for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**, 919–927.

1988 Kent, J.T. and Mardia, K.V. Spatial classification using fuzzy membership models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**, 659–671.

1988 Mardia, K.V. Multi-dimensional multivariate Gaussian Markov random fields. *Journal of Multivariate Analysis* **24**, 265–284.

1989 Mardia, K.V. and Dryden, I.L. Shape distribution for landmark data. *Advances in Applied Probability* **21**, 742–755.

1989 Mardia, K.V. and Watkins, A.J. On multimodality of the likelihood in the spatial linear model. *Biometrika* **76**, 289–295.

1989 Mardia, K.V. Shape analysis of triangles through directional techniques. *Journal of the Royal Statistical Society Series B* **51**, 449–458.

1989 Mardia, K.V. and Dryden, I. The statistical analysis of shape data. *Biometrika* **76**, 271–281.

1989 Mardia, K.V. Some contributions to shape analysis. *Statistical Science* **4**, 108–111.

1989 Jupp, P.E. and Mardia, K.V. Unified view of the theory of directional statistics, 1975-1985. *International Statistical Review* **57**, 261–294.

1991 Mardia, K.V. and Kent, J.T. Rao score tests of goodness-of-fit and independence. *Biometrika* **78**, 2, 355–363.

1991 Dryden, I. and Mardia, K.V. General shape distributions in a plane. *Advances in Applied Probability* **23**, 259–276.

1991 Goodall, C.R. and Mardia, K.V. A geometrical derivation of the shape density. *Advances in Applied Probability* **23**, 496–514.

1992 Goodall, C.R. and Mardia, K.V. The noncentral Bartlett decomposition and shape densities. *Journal of Multivariate Analysis* **40**, 94–108.

1992 Dryden, I.L. and Mardia, K.V. Size and shape analysis of landmark data. *Biometrika* **79**, pp 57–68.

1992 Mardia, K.V., Li, Q. and Hainsworth, T.J. On the penrose hypothesis on fingerprint patterns. *IMA Journal of Mathematics Applied in Medicine and Biology* **9**, pp 289–294.

1993 Goodall, C.R. and Mardia, K.V. Multivariate aspects of shape theory. *Annals of Statistics* **21**, 848–866.

1994 Mardia, K.V. and Dryden, I.L. Shape averages and their bias. *Advances in Applied Probability* **26**, 334–340.

1994 Dryden, I.L. and Mardia, K.V. Multivariate shape analysis. *Sankhyā*: Special Volume 55, Series A, Part 3, 460–480. Dedicated to the memory of P.C. Mahalanobis.

1994 Mardia, K.V., Dryden, I.L., Hurn, M.A., Li, Q., Millner, P.A. and Dickson, R.A. Familial spinal shape. *Journal of Applied Statistics* **21**, 623–642.

1994 Mardia, K.V. and Walder, A.N. Size-shape distributions for paired landmark data. *Advances in Applied Probability* **26**, 894–905.

1995 Prentice, M.J. and Mardia, K.V. Shape changes in the plane for landmark data. *Annals of Statistics* **23**, 1960–1974.

1996 Mardia, K.V., Kent, J.T., Goodall, C.R. and Little, J.A. Kriging and splines with derivative information. *Biometrika* **83**, 207–221. [Amendments and Corrections in Biometrika, 1998, vol 85, p505]

1996 Mardia, K.V., Coombes, A., Kirkbride, J., Linney, A. and Bowie, J.L. On statistical problems with face identification from photographs. *Journal of Applied Statistics* **23**(6), 655–675.

1996 Mardia, K.V., Goodall, C. and Walder, A.N. Distributions of projective invariants and model-based machine vision. *Advances in Applied Probability* **28**, 641–661.

1996 Hurn, M.A., Mardia, K.V., Hainsworth, T.J., Kirkbride, J. and Berry, E. Bayesian fused classification of medical images. *IEEE Transactions on Medical Imaging* **15**, 850–858.

1996 Mardia, K.V., Kent, J.T. and Walder, A.N. Conditional cyclic Markov random fields. *Journal of Advances in Applied Probability* **28**, 1–12.

1996 Kent, J.T. and Mardia, K.V. Spectral and circulant approximations to the likelihood for stationary Gaussian random fields. *Journal of Statistical Planning and Inference* **50**, 379–394.

1997 Kent, J.T. and Mardia, K.V. Consistency of Procrustes estimators. *Journal of the Royal Statistical Society Series B* **59**, 281–290.

1997 Mardia, K.V., Qian, W., Shah, D. and de Souza, K.M.A. Deformable template recognition of multiple occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**, 1036–1042.

1997 Diaz Garcia, J.A., Jaimez, R.G. and Mardia, K.V. Wishart and pseudo-Wishart distributions and some applications to shape theory. *Journal of Multivariate Analysis* **63**, 73–87.

1998 Mardia, K.V., Goodall, C.R., Redfern, E.J. and Alonso, F.J. The kriged Kalman filter. Presented to the Spanish Statistical Society. *TEST 7*, pp. 217–276. (Discussion paper.)

1999 Mardia, K.V., Southworth, H.R. and Taylor, C.C. On bias in maximum likelihood estimators. *Journal of Statistical Planning and Inference* **76**, 31–39.

1999 Mardia, K.V., Morris, R.J., Walder, A.N. and Koenderink, J.J. Estimation of torsion. *Journal of Applied Statistics* **26**, 373–381.

1999 de Souza, K.M.A., Kent, J.T. and Mardia, K.V. Stochastic templates for aquaculture images and a parallel pattern detection. *Journal of the Royal Statistical Society Series C* **48**, 211–227.

1999 Goodall, C. and Mardia, K.V. Projective shape analysis. *Journal of Computational and Graphical Statistics* **8**, 143–168.

1999 Mardia, K.V. and Dryden, I.L. Complex Watson distribution and shape analysis. *Journal of the Royal Statistical Society Series B* **61**, 913–926.

2000 Mardia, K.V., Bookstein, F.L. and Moreton, I.L. Statistical assessment of bilateral symmetry. *Biometrika* **87**, 285–300.

2000 Southworth, R., Mardia, K.V. and Taylor, C.C. Transformation- and label-invariant neural network for the classification of landmark data. *Journal of Applied Statistics* **27**, 205–215.

2001 de Souza, K.M.A., Jackson, A.L., Kent, J.T., Mardia, K.V. and Soames, R.W. An assessment of the accuracy of stereolithographic skull models. *Clinical Anatomy* **14**, 296.

2001 Kent, J.T. and Mardia, K.V. Shape, Procrustes tangent projections and bilateral symmetry. *Biometrika* **88**, 469–485.

2001 Glasbey, C.A. and Mardia, K.V. A penalized likelihood approach to image warping (with discussion). *Journal of the Royal Statistical Society Series B* **63**, 465–514.

2002 Duta, N, Jain, A.K. and Mardia, K.V. Matching of palm prints. *Pattern Recognition Letters* **23**, 477–485.

2002 Downs, T.D. and Mardia, K.V. Circular regression. *Biometrika* **89**, 683–697.

2004 Bookstein, F.L., Mardia, K.V. and Kirkbride, J. Statistics of shape, direction and cylindrical variables. *Journal of Applied Statistics* **31**, 465–479.

2005 Sahu, S.K. and Mardia, K.V. A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. *Applied Statistics* **54**, 223–244.

2005 Mardia, K.V. and Patrangenaru, V. Directions and projective shapes. *Annals of Statistics* **33**, 1666–1699.

2005 Mardia, K.V. and Gilks, W.K. Meeting the statistical needs of 21st-century science. *Significance* **2**, 162–165.

2005 Sahu, S.K., Jona Lasini, G., Orasi, A. and Mardia, K.V. A comparison of spatio-temporal Bayesian models for reconstruction of rainfall fields in a cloud seeding experiment. *Journal of Mathematics and Statistics* **1**, 273–281.

2006 Mardia, K.V., Angulo, J.M. and Goitia, A. Synthesis of image deformation strategies. *Image and Vision Computing* **24**, 1–12.

2006 Xu, C., Dowd, P.A., Mardia, K.V. and Fowell, R.J. A connectivity index for discrete fracture networks. *Mathematical Geology* **38**, 611–634.

2006 Xu, C., Dowd, P.A., Mardia, K.V. and Fowell, R.J. A flexible true Pluri-Gaussian code for spatial facies simulations. *Computers & Geosciences* **32**, 1629–1645.

2006 Green, P.J. and Mardia, K.V. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika* **93**, 235–254.

2006 Mardia, K.V., McDonnell, P. and Linney, A.D. Penalised image averaging and discriminations with facial and fishery applications. *Journal of Applied Statistics* **33**, 339–371.

2006 Mardia, K.V., Bookstein, F.L., Kent, J.T. and Meyer, C.R. Intrinsic random fields and image deformations. *Journal of Mathematical Imaging and Vision* **26**, 59–71.

2006 Kent, J.T., Mardia, K.V. and McDonnell, P. The complex Bingham quartic distribution and shape analysis *Journal of Royal Statistical Society Series B* **68**, 747–765.

2006 Micheas, A.C., Dey, D.K. and Mardia, K.V. Complex elliptical distributions with application to shape analysis. *Journal of Statistical Planning and Inference* **136**, 2961–2982.

2007 Mardia, K.V., Taylor, C.C. and Subramaniam, G.K. Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* **63**, 505–512.

2007 Mardia, K.V., Nyirongo, V.B., Walder, A.N., Xu, C., Dowd, P.A., Fowell, R.J. and Kent, J.T. Markov chain Monte Carlo implementation of rock fracture modelling. *Mathematical Geology* **39**, 355–381.

2007 Mardia, K.V., Nyirongo, V.B., Green, P.J., Gold, N.D. and Westhead, D.R. Bayesian refinement of protein functional site matching. *BMC Bioinformatics* **8**, 257.

2007 Davies, J.R., Jackson, R.M., Mardia, K.V. and Taylor, C.C. The Poisson index: a new probabilistic model for protein-ligand binding site similarity. *Bioinformatics* **23**, 3001–3008.

2007 Dowd, P.A., Xu, C., Mardia, K.V., and Fowell, R.J. A comparison of methods for the stochastic simulation of rock fractures. *Mathematical Geology* **39**, 697–714.

2008 Mardia, K.V., Hughes, G., Taylor, C.C. and Singh, H. A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics* **36**, 99–109.

2008 Boomsma, W., Mardia, K.V., Taylor, C.C., Ferkinghoff-Borg, J., Krogh, A. and Hamelryck, T. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 8932–8937.

2008 Mardia, K.V. and Nyirongo, V.B. Simulating virtual protein C_α traces with applications. *Journal of Computational Biology* **15**, 1209–1220.

2009 Pardo-Igúzquiza, E., Mardia, K.V. and Chica-Olmo, M. MLMATERN: a computer program for maximum likelihood inference with the spatial Matérn covariance model. *Computers & Geosciences* **35**, 1139–1150.

2009 Frellsen, J., Moltke, I., Thiim, M., Mardia, K.V., Ferkinghoff-Borg, J. and Hamelryck, T. A probabilistic model of RNA conformational space. *PLoS Computational Biology* **5**, 1–11.

2009 Dowd, P.A., Martin, J.A., Xu, C., Fowell, R.J. and Mardia, K.V. A three-dimensional fracture network data set for a block of granite. *International Journal of Rock Mechanics and Mining Sciences* **46**, 811–818.

2009 Mardia, K.V., Kent, J.T., Hughes, G. and Taylor, C.C. Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika* **96**, 975–982.

2010 Mardia, K.V. Bayesian analysis for bivariate von Mises distributions. *Journal of Applied Statistics* **37**, 515–528.

2011 Mardia, K.V., Nyirongo, V.B., Fallaize, C.J., Barber, S. and Jackson, R.M. Hierarchical Bayesian modelling of pharmacophores in bioinformatics. *Biometrics* **67**(2), 611–619.

2011 Hommola, K., Gilks, W.R and Mardia, K.V. Log-linear modelling of protein dipeptide structure reveals interesting patterns of side-chain-backbone interactions. *Statistical Applications in Genetics and Molecular Biology* **10**, Article 8, 1–29.

2012 Mardia, K.V., Petty, E.M. and Taylor, C.C. Matching markers and unlabeled configurations in protein gels. *Annals of Applied Statistics* **6**, 853–869.

2012 Mardia, K.V. and Cooper, S.B. Alan Turing and enigmatic statistics. *Bulletin Brazil International Society for Bayesian Analysis* **5**, 2–7.

2013 Mardia, K.V., Fallaize, C.J., Barber, S., Jackson, R.M. and Theobald, D.L Bayesian alignment of similarity shapes. *Annals of Applied Statistics* **7**, 989–1009.

2013 Mardia, K.V., Bookstein, F.L. and Kent, J.T. Alcohol, babies and the death penalty: saving lives by analysing the shape of the brain. *Significance* **10**, 12–16.

2013 Mardia, K.V. Statistical approaches to three key challenges in protein structural bioinformatics, *Journal of the Royal Statistical Society Series C* **62**, 487–514.

2013 Valentin, J., Andreetta, C., Boomsma, W., Bottaro, S., Ferkinghoff-Borg, J., Frellsen, J., Mardia, KV, Tian, P. and Hamelryck, T. Formulation of probabilistic models of protein structure in atomic detail using the reference ratio method. *Proteins* **82**, 288–299.

2013 Olsson, S., Frellsen, J., Boomsma, W., Mardia, K.V. and Hamelryck, T. Inference of structure ensembles of flexible biomolecules from sparse, averaged data. *PLoS ONE* **8**, e79439.

4. Articles in Edited Volumes (other than edited by Mardia)

1972 Mardia, K.V. “Percentage Points of the F-distribution for Fractional Degrees of Freedom” *Biometrika Table for Statisticians* **2**, E.S. Pearson and H.O. Hartley. Table 4, 170–174.

1975 Mardia, K.V. “Characterization of directional distributions” *Statistical Distributions in Scientific Work*, Characterizations and Applications, Vol. 3. G.P. Patil, S. Kotz and J.K. Ord (eds). D. Reidel Publishing Co., Dordrecht, Holland, pp. 365–386.

1975 Bingham M.S. and Mardia, K.V. “Maximum likelihood characterizations of the von Mises distribution” *Statistical Distributions in Scientific Work*, Characterizations and Applications, Vol. 2. G.P. Patil, S. Kotz and J.K. Ord (eds). D. Reidel Publishing Co., Dordrecht, Holland, pp. 387–398.

1977 Mardia, K.V., Edwards, R. and Puri, M.L. “Analysis of Central Place Theory” *Proceedings of the 41st Conference of the International Statistical Institute.* Vol. 47, pp. 93–110.

1980 Mardia, K.V. ”Some statistical inference problems in Kriging II: Theory” *Proceedings of the 26th International Geology Congress*, pp. 113–131. Sciences de la Terre: Advances in Automatic Processing and Mathematical Models in Geology, Series Informatique Geologie, number 15.

1982 Mardia, K.V. “Directional distributions” *Encyclopaedia of Statistical Science*, Vol. II S. Kotz and T. Johnson (eds), John Wiley & Sons, Ltd London and New York, pp. 381–385.

1983 Mardia, K.V. “Mardia’s test for multivariate normality, skewness and kurtosis tables” *Statistical Tables for Multivariate Analysis in Kres, H*. Springer-Verlag, New York, pp. 426–431.

1985 Mardia, K.V. “Mardia’s test of multinormality” *Encyclopaedia of Statistical Sciences*, Vol. 5, John Wiley & Sons, Inc., New York, pp. 217–221.

1990 Mardia, K.V. “Maximum likelihood estimation for spatial models” *Proceedings Spatial Statistics: Past, Present and Future* D.A. Griffith (ed.), Institute of Mathematical Geology. Michigan Document Service, pp. 203–225.

1991 Mardia, K.V., Kent, J.T. and Walder, A.N. “Statistical shape models in image analysis” *Proceeding of Interface’91*, pp. 550–557.

1992 Mardia, K.V. and Kent, J.T. “Statistical shape methodology in image analysis” NATO Conference. *Proceedings of “Shape in Pictures”* O. Ying-Lie, A. Toet, H.J.A.M. Heijmans, D.H. Foster, P.M. Driebergen (eds), Springer-Verlag Heidelberg, The Netherlands, 1993, pp. 443–452.

1993 Mardia, K.V. and Goodall, C.R. Spatial-temporal analysis of multivariate environmental monitoring data *Multivariate Environmental Statistics*, Vol. 6 G.P. Patil and C.R. Rao (eds), pp. 347–385.

1993 Mardia, K.V., Rabe, S. and Coombes, A.M. “Statistical analysis of 3D range images” *Proceedings of 8th Scandinavian Conference on Image Analysis*, Vol. I K.A. Hogda, B. Braathen and K. Heia (eds), Nobin, Norway, pp. 17–19.

1994 Kent, J.T., Mardia, K.V. and Rabe, S. “Face description from laser range data”. *SPIE, The International Society for Optical Engineering. San Diego 1994*, Vol. 2299, pp. 32–45.

1994 Kent, J.T. and Mardia, K.V.) “Link between kriging and thin plate splines” *Festschrift Volume to P. Whittle. Probability, Statistics and Optimization* F.P. Kelly (ed), John Wiley & Sons, pp. 325–339.

1994 Goodall, C. and Mardia, K.V. “Challenges in multivariate spatio-temporal modelling” Invited Session “Spatial-Temporal Model in Environmental Statistics” *Proceedings of the XVII International Biometric Conference Hamilton*, Ontario, Canada, 1994, Vol. 1, pp. 1–17.

1996 Kent, J.T., Mardia, K.V. and West, J. “Ridge curves and shape analysis” *Proceedings of the 7th BMVC’96* R.B. Fisher and E. Trucco (eds), pp. 43–52.

1997 de Souza, K.M.A., Kent, J.T. and Mardia, K.V. “Estimation of objects in highly-variable images using Markov Chain Monte Carlo” *BMVC’97 Proceedings*. 2 A.F. Clark (ed.), pp. 460–469.

1997 Statistical methods for automatic interpretation of digitally scanned fingerprints. (Mardia, K.V., Baczkowski, A.J., Feng, X. and Hainsworth, T.J.) *Pattern Recognition Letters* **18**, 1197–1203.

1998 Mardia, K.V., Kent, J.T., Lee, D. and de Souza, K.N.A. “Using a Bayesian approach on a network of quads to track tagged cardiac MR images” *Proceedings MIUA’98*, pp. 1–4.

1999 Mardia, K.V. “Landmark data” *Encyclopaedia of Statistical Science: Update*, Vol. 3 S. Kotz, C.B. Read and D.L. Banks (eds), Wiley-Interscience Publication, pp. 391–402.

2000 Kent, J.T., Mardia, K.V., Morris, R.J. and Aykroyd, R.G. “Procrustes growth models for shape” *First Joint Statistical Meeting*, IISA India, 236–238.

2002 Kent, J.T. and Mardia, K.V. “Modelling strategies of spatial-temporal data”. *Spatial Cluster Modelling* A.B. Lawson and D.G.T. Denison (eds), Chapman and Hall/CRC, London, pp. 213–226.

2002 Mardia, K.V. “Why is directional statistics pivotal to geosciences?” *Opening Address at the 8th Annual Conference IAMG Proceedings*, Vol. 1, pp. 41–47.

2007 Lasinio, G.J., Sahu, S.K. and Mardia, K.V. “Modeling rainfall data using a Bayesian Kriged-Kalman model”. *Bayesian Statistics and its Applications* Upadhaya, S.K., Singh, U. and Dey, D.K. (eds) Anamaya Publishers, New Delhi, India, pp. 301–318.

2007 Mardia, K.V. “Why statistical shape analysis is pivotal to the modern pattern recognition” *Proceedings of the 6th International Conference on Advances in Pattern Recognition*, Pal, L. (ed.), The World Scientific Publishing Co., Singapore, pp. 3–11.

2007 Mardia, K.V. “Should geostatistics be model-based?” *Proceedings of the IAMG 2007 Conference: Geomathematics and GIS Analysis of Resources, Environment and Hazards*, Beijing, China, pp. 4–9.

2008 Boomsma, W., Borg, M., Frellsen, J., Harder, T., Stovgaard, K., Ferkinghoof-Borg, J., Krogh, A., Mardia, K.V. and Hamelryck, T. “Protein structure prediction using a probabilistic model of local structure”. *8th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction Cagliari*, Sardinia, Italy December 3-7, 2008, pp. 82–83.

2010 Green, P.J., Mardia, K.V., Nyirongo, V.B. and Ruffieux, Y. “Bayesian modeling for matching and alignment of biomolecules *The Oxford Handbook of Applied Bayesian Analysis”* O’Hagan, A. and West, M. (eds), Oxford University Press, pp. 27–50.

Part II

DIRECTIONAL DATA ANALYSIS

4

Some advances in constrained inference for ordered circular parameters in oscillatory systems

Cristina Rueda[1], Miguel A. Fernández[1], Sandra Barragán[1] and Shyamal D. Peddada[2]

[1]*Department of Statistics and O.R., Universidad de Valladolid, Valladolid, Spain*

[2]*National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA*

4.1 Introduction

Constraints on parameters arise naturally in many applications. Statistical methods that honor the underlying constraints tend to be more powerful and result in better interpretation of the underlying scientific data. In the context of Euclidean space data, there exists over five decades of statistical literature on constrained statistical inference and at least four books on the subject (e.g. Robertson et al. 1988; Silvapulle and Sen 2005). However, it was not until recently that these methods have been used extensively in applied research. For example, constrained statistical inference is gaining considerable interest among applied researchers in a variety of fields, such as, for example, toxicology (Peddada et al. 2007), genomics (Hoenerhoff et al. 2013; Perdivara et al. 2011; Peddada et al. 2003), epidemiology (Cao et al. 2011; Peddada et al. 2005), clinical trials (Conaway et al. 2004), or cancer trials (Conde et al. 2012, 2013).

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

While Euclidean space data are commonly encountered in applications, there are numerous instances where the underlying data and the parameters of interest reside on a unit circle. Statistical theory and methodology for analyzing such angular data has a long history (Fisher 1993; Mardia and Jupp 2000) and, as witnessed through his publications and his highly referenced book (Mardia and Jupp 2000), Professor Mardia was one of the chief architects and pioneers of this important research area. His work has a wide range of applications in fields such as, for example, geosciences, spatial data, image analysis, and bioinformatics.

In comparison to over fifty years of statistical literature on constrained inference for Euclidean space data, constrained statistical inference for circular data is almost nonexistent although constraints on unit circle were encountered by applied researchers such as, for example, social psychologists and neuroscientists (Schlosberg 1952; Russell 1980; Forgas 1998; Mechsner et al. 2001; Oullier et al. 2002; Posner et al. 2005) or molecular biologists (Whitfield et al. 2002; Peng et al. 2005; Hughes et al. 2009).

Parameters on a unit circle are often the result of an oscillatory system. Oscillatory systems arise naturally in many applications, such as, among others, sales of seasonal products, regulation of hormones in humans, circadian clock, or periodic expression of genes participating in cell division cycle. Often there are several components (or variables) involved in such oscillatory systems that act in a well-coordinated manner such as an orchestra for the system to function. The system can be disrupted if one or more components go out of order. Researchers are often interested in detecting such components. For example, large scale genomic studies are routinely conducted to identify genes/proteins that have a periodic expression in a given biological system. Depending on the underlying scientific question of interest, researchers are often interested in, for example, correlating the phases of periodic genes across different experimental conditions or species or tissues. Thus, the statistical problem of interest is to draw inferences regarding the relative order among parameters on a unit circle.

Just as one cannot trivially extend standard methods developed for unconstrained statistical inference in the Euclidean space to circle, constrained statistical inference for the Euclidean space cannot be extended to constraints on a unit circle (cf. Rueda et al. 2009). Since constrained statistical inference on a unit circle is a relatively new topic and yet has numerous applications, the purpose of this paper is threefold. First, we describe recent theoretical and methodological advances in this field, next we shall describe some applications of the methodology in cell biology and lastly we shall present several open research problems and potential applications. While the methodology described here is a review of what has already appeared in our previous papers, the applications are new. More specifically, in Section 4.2 we introduce the framework and the problem of interest. In Section 4.3, we describe the problem of estimating ordered parameters on a unit circle using circular isotonic regression. Analogous to the isotonic regression estimator in the Euclidean space, circular isotonic regression estimator (CIRE) obtains ordered estimates of circular parameters under a prespecified order among them. Using these ordered point estimators, under suitable distributional assumptions, in Section 4.4 we describe conditional tests for order among circular parameters. In Section 4.5, the problem of estimation of a global order among a set of circular objects using data from multiple experiments is described. Statistical methodology described in this paper is illustrated in Section 4.6 using data obtained from cell biology. We conclude the paper by presenting several open research problems and potential applications in Section 4.7.

4.2 Oscillatory data and the problems of interest

Time course data are commonly obtained in many applications. However, in some applications such as, among others, marketing research, cell biology, endocrinology, and psychology, researchers are interested in studying various characteristics (or parameters) of the time course pattern. Although the raw data itself may reside in the Euclidean space, the underlying parameters of interest may be points on a unit circle. To illustrate this, consider data provided in the toy example described in Figure 4.1. To promote tourism to its summer resort in an island in the Pacific, suppose a travel agency runs an advertisement campaign several months before each summer. The advertisement costs in dollars over time are plotted in Figure 4.1 (dashed curve). The travel agency tracks the sales of airline tickets to the island over the same period (dotted curve) as well as the income revenues on the island due to tourism (solid curve). One of the parameters of interest to the travel agency is to determine the time of peak advertisement to maximize its impact on the overall sales. Thus, the parameters of interest are the times that correspond to the peaks of the curve (location of the vertical lines in Figure 4.1). Since these curves are periodic, they can be mapped onto a unit circle and the time to peak value of any given curve can be thought of as an angular parameter on the circle (see Figure 4.2). Thus in this example, the angular parameters are ordered with the dashed value followed by the dotted value, which is followed by the solid one in the anticlockwise direction. Focus of this paper is to draw inferences regarding the relative order among these angular parameters on the unit circle. As noted in the introduction, similar examples arise in a wide range of settings and the application of interest in this paper is cell biology, which is explained in greater detail in the illustration section.

In Liu et al. (2004), a nonlinear model called the Random Period Model (RPM) was introduced for such time course data. Although their motivation was to describe the time course expression of cell cycle genes, their model can be used for any such time course data. The model is given by $Y_g(t) = f(t, \eta_g) + \varepsilon_g(t)$, where t is the time, and $\varepsilon_g(t)$ is a zero mean error term with no additional distributional assumptions made. The expected response $f(t, \eta_g)$ is modeled as,

$$f(t, \eta_g) = a_g + b_g t + \frac{K_g}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos\left(\frac{2\pi t}{T\exp(\sigma z)} + \phi_g\right) \exp\left(\frac{-z^2}{2}\right) dz,$$

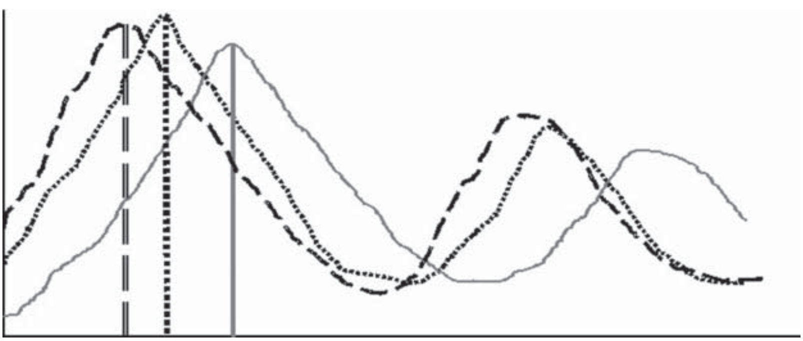


Figure 4.1 Advertisement costs (dashed curve), sales of airline tickets (dotted curve) and income revenues (solid curve) in dollars over time.

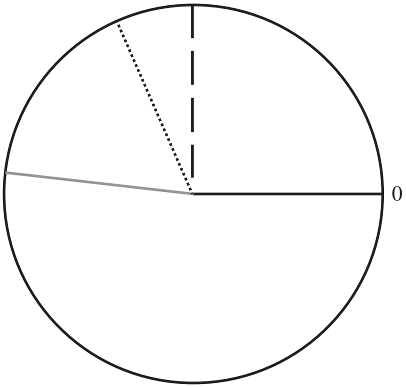


Figure 4.2 Peak costs plotted on a circle. Angle from 0 to the dashed line for advertisement costs, from 0 to the dotted line for sales of airline tickets and from 0 to the solid line for the income revenues.

for all $t = 1, \ldots, n_g$ and $g = 1, \ldots, k$ and where $\eta_g = (K_g, T, \sigma, \phi_g, a_g, b_g)$ is the parameter vector. Parameters of the model are interpreted as follows. The parameters T and σ are the same for all cells and genes in the population. The parameter T governs the duration of the cell cycle, while σ measures the rate of attenuation in amplitude with each cycle (the larger σ the faster the decay in amplitude). The parameter ϕ_g is the angle of peak expression of gene g in the cell cycle with $\phi = 0$ being the point when cells are released. Parameter K_g is the amplitude of the first period and parameters a_g and b_g take into account possible drifts in the gene background expression level. The unconstrained estimators of all parameters of RPM, including the angular parameter ϕ_g, are obtained using nonlinear least squares methodology. Throughout this paper, we shall refer to the angular parameter ϕ_g as the phase angle due to its biological relevance.

Suppose we have k oscillatory variables (in the aforementioned tourism example we had three) and suppose for the ith variable the phase angle is denoted by ϕ_i, $i = 1, 2, \ldots, k$. Then, using the unconstrained estimators θ_i, $i = 1, 2, \ldots, k$ obtained from the RPM model, our goal is to conduct inference regarding the relative order of $\phi_1, \phi_2, \ldots, \phi_k$ around the unit circle. Suppose we travel around the circle in an anticlockwise direction and suppose the angle ϕ_1 is followed by ϕ_2, which is followed by $\phi_3 \cdots$ followed by ϕ_k which is finally followed by angle ϕ_1. Then we shall adopt the following notation (cf. Rueda et al. 2009; Fernández et al. 2012) to describe the relative order:

$$\phi_1 \preceq \phi_2 \preceq \cdots \preceq \phi_k \preceq \phi_1.$$

It is important to note that the aforementioned order is invariant of the location of the pole of the circle. Alternatively, the aforementioned order is rotation invariant. For this reason, Rueda et al. (2009) referred to the aforementioned order as an isotropic order. The focus of this paper is to discuss recent developments in the literature on the following problems: (i) estimation of $\phi_1, \phi_2, \ldots, \phi_k$ under the aforementioned order constraint using the unconstrained estimators of ϕ_i, $i = 1, 2, \ldots, k$, obtained from RPM; (ii) testing the hypothesis that the aforementioned relative order is satisfied for a set of angular parameters; and

(iii) testing whether the relative order among a set of phase angles is conserved using data from multiple experiments performed under different conditions.

4.3 Estimation of angular parameters under order constraint

We begin by discussing the problem of estimating the phase angles ϕ_i, $i = 1, 2, \ldots, k$ under the order constraint $\phi_1 \preceq \phi_2 \preceq \cdots \preceq \phi_k \preceq \phi_1$ using the unconstrained estimators θ_i, $i = 1, 2, \ldots, k$ obtained from the RPM. The general idea of estimation resembles the analogous problem in the Euclidean space. Let $\mathcal{C} = \{\phi \in [0, 2\pi)^k : \phi_1 \preceq \phi_2 \preceq \ldots \preceq \phi_k \preceq \phi_1\}$. Suppose $\mathcal{C}^i = \{\phi \in [0, 2\pi)^k : 0 \leq \phi_i \leq \phi_{i+1} \leq \cdots \leq \phi_{i-1} \leq 2\pi\}$, thus the pole of the unit circle is between the parameters ϕ_{i-1} and ϕ_i. Then we have $\mathcal{C} = \bigcup_{i=1}^{k} \mathcal{C}^i$.

For an estimator $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$ of a parameter $\phi = (\phi_1, \phi_2, \ldots, \phi_k)'$, the distance between the two is defined as the sum of circular errors (SCE) given by:

$$\mathrm{SCE}(\theta, \phi) = d(\theta, \phi) = \sum_{i=1}^{k} r_i\{1 - \cos(\theta_i - \phi_i)\},$$

where r_i represents a measure of concentration of θ_i about its modal direction (see Mardia and Jupp 2000, p. 17). Consequently, using the unconstrained estimator θ, the estimator of ϕ under the constraint $\phi \in \mathcal{C}$ is obtained by solving the following minimization problem:

$$\min_{\phi \in \mathcal{C}} \mathrm{SCE}(\theta, \phi) = \min_{\phi \in \mathcal{C}} \sum_{i=1}^{k} r_i\{1 - \cos(\theta_i - \phi_i)\}. \tag{4.1}$$

In the case of Euclidean space data where θ has a known diagonal covariance matrix, the corresponding restricted cone is the simple order cone given by $\phi_1 \leq \phi_2 \leq \cdots \leq \phi_k$, SCE is replaced by the Euclidean distance (i.e., sum of squared errors) and the corresponding minimization problem is called the isotonic regression. Typically the problem is solved using the pool adjacent violator algorithm (PAVA). The basic underlying idea of PAVA is that components of θ that violate the underlying relative order are pooled or averaged so that the overall order is satisfied. To illustrate this, we consider the following toy example in the Euclidean space.

Example 4.3.1 *Suppose $\phi = (\phi_1, \phi_2, \phi_3) \in \mathbb{R}^3$ with $\phi_1 \leq \phi_2 \leq \phi_3$. Suppose the unconstrained sample means are given by $\theta_1 = 0.6$, $\theta_2 = 2.5$, and $\theta_3 = 1.5$. Since $\theta_2 > \theta_3$, therefore, the order $\phi_1 \leq \phi_2 \leq \phi_3$ is violated. The PAVA would average the last two coordinates, yielding the $\widetilde{\phi}_1 = 0.6$, $\widetilde{\phi}_2 = \widetilde{\phi}_3 = (1.5 + 2.5)/2 = 2$ as the constrained estimates.*

In the case when $\phi \in \mathcal{C}$, the unit circle, the solution to the minimization problem (4.1) is more complicated as noted in Rueda et al. (2009). Since (4.1) resembles the usual isotonic regression estimation for Euclidean space data, Rueda et al. (2009) refer to the solution of (4.1) as circular isotonic regression estimator (CIRE). More precisely, the CIRE, denoted as $\widetilde{\phi}$, is given by:

$$\widetilde{\phi} = \arg\min_{\phi \in \mathcal{C}} \mathrm{SCE}(\theta, \phi). \tag{4.2}$$

Before we formally describe CIRE, we consider the following toy example to describe the calculation of CIRE geometrically. We remark that when considering angular data, the arithmetic means are not always appropriate for describing the average direction between a pair of angles. Instead, one should use the angular mean direction (cf. Mardia and Jupp 2000; Rueda et al. 2009).

Example 4.3.2 *Suppose $k = 3$ with $\phi_1 \preceq \phi_2 \preceq \phi_3 \preceq \phi_1$. Suppose the unconstrained estimates using the RPM are given (in radians) by $\theta_1 = 6$, $\theta_2 = 2.5$ and $\theta_3 = 1.5$ (see Figure 4.3). Clearly these estimates do not satisfy the desired order. In the Euclidean space example described earlier, it was easy to identify the violator of the order and one could accordingly deal with it. However, in this case, since the data wrap around the circle, the violator may not be unique and one needs to explore all possibilities. If the violation is due to θ_1 and θ_3, then one would average these two and leave θ_2 as is. This would result in the constrained estimates given in the top left circle in Figure 4.4 with an SCE of 0.74. However, if the violation is due to θ_1 and θ_2 then one would average these two and leave θ_3 as is. This would result in the constrained estimates given in the top right circle in Figure 4.4 with an SCE of 1.64. Or the last possibility could be that θ_2 and θ_3 are in violation of the order. In which case, we pool the estimates θ_2 and θ_3 and leave θ_1 as is, resulting in an SCE of 0.24. See the bottom circle in Figure 4.4. Since this SCE is the smallest, it is the CIRE.*

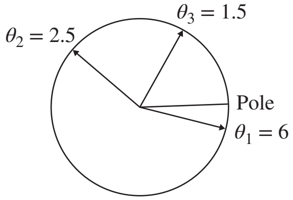


Figure 4.3 Unconstrained estimates.

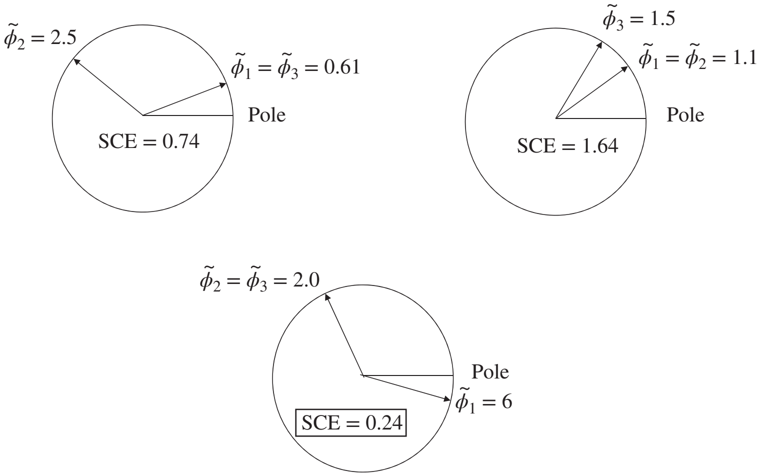

$\tilde{\phi}_2 = 2.5$
$\tilde{\phi}_1 = \tilde{\phi}_3 = 0.61$
Pole
SCE = 0.74
$\tilde{\phi}_3 = 1.5$
$\tilde{\phi}_1 = \tilde{\phi}_2 = 1.1$
Pole
SCE = 1.64
$\tilde{\phi}_2 = \tilde{\phi}_3 = 2.0$
Pole
$\tilde{\phi}_1 = 6$
SCE = 0.24

Figure 4.4 Possible constrained estimates with the CIRE appearing at the bottom circle.

An important observation to make from the aforementioned example is that in the case of circular data it is not enough to consider the adjacent violators of the order. This makes the problem computationally challenging. The main reason for this is that, unlike the arithmetic mean in the Euclidean space, the circular mean does not verify the Cauchy Mean Value property. Rueda et al. (2009) provide a general algorithm to derive CIRE and demonstrated that their algorithm is exact and computationally efficient, especially as the number of parameters increases. CIRE is implemented in the `R` package **isocir** (Barragán et al. 2013). The solution to their algorithm is characterized in the following theorem.

Theorem 4.3.3 *The CIRE exists, is almost sure unique, and can be obtained from circular means of adjacent angles as,*

$$\widetilde{\phi}_g = Ave(S_{(i)}) \textit{ for } g = 1, ..., k, \; i = 1, ..., m,$$
$$\textit{with } 0 \leq Ave(S_{(1)}) < Ave(S_{(2)}) < \cdots < Ave(S_{(m)}) < 2\pi,$$

where $(i)_{i=1}^m$ is a partition of $\{1, \ldots, k\}$, $Ave(S_{(i)})$ are the circular mean directions for angles in $S_{(i)} = \{\theta_g, g \in (i)\}$, $(1), .., (m)$ are the so called level sets (cf. Robertson et al. 1988), $n_{(i)} = \#(i)$ and $\sum_{i=1}^m n_{(i)} = n$.

In some situations, especially in cell biology, one may be interested in partial orders of the following type:

$$\{\phi_1, \phi_2, \ldots, \phi_{r_1}\} \preceq \{\phi_{r_1+1}, \ldots, \phi_{r_2}\} \preceq \cdots \preceq \{\phi_{r_s+1}, \ldots, \phi_k\} \preceq \{\phi_1, \phi_2, \ldots, \phi_{r_1}\}. \tag{4.3}$$

In the aforementioned notation, angles within each set are not ordered but the angles in one set precede the angles in the next set. Thus, all angles in $\{\phi_1, \phi_2, \ldots, \phi_{r_1}\}$ precede all the angles in $\{\phi_{r_1+1}, \phi_{r_1+2}, \ldots, \phi_{r_2}\}$ and so on. This occurs when a biologist may hypothesize that, as a group, genes in a given set have to function before the genes in the next set function for the cell division cycle to proceed. He/she may not know the order of expression of genes within each set. Barragán et al. (2013) extended the CIRE methodology of Rueda et al. (2009) to estimate parameters under the aforementioned order constraint.

4.4 Inferences under circular restrictions in von Mises models

When dealing angular data, analogous to normal distribution on the real line, one typically uses the von Mises distribution for performing inferences regarding the angular parameter (cf. Mardia and Jupp 2000). Accordingly, in this section, we shall make a simplifying assumption that the unconstrained estimators $\theta_i, i = 1, 2, \ldots, k$ are mutually independently distributed with $\theta_i \sim \mathrm{VM}(\phi_i, \kappa)$, where VM stands for von Mises distribution, ϕ_i denotes the angular mean direction and κ is the concentration parameter of the distribution. The probability density function (pdf) is given by

$$g(x, \phi_i, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(x - \phi_i)} \qquad x \in [0, 2\pi),$$

where I_0 is the modified Bessel function of first class and order zero. As noted earlier, there exists a large body of literature on statistical tests for angular data, especially under the von Mises distribution (cf. Mardia and Jupp 2000). However, until Fernández et al. (2012) and Barragán et al. (2013), there did not exist any formal literature on testing for order among angular parameters. Motivated by various applications, that is, in social psychology, neurology, cell biology, one may be interested in testing the following hypotheses:

$$\begin{array}{ll} H_0: & \phi_i, i=1,...,k \text{ follow a known order } O, \\ H_1: & H_0 \text{ is not true.} \end{array} \tag{4.4}$$

For example, O may be the circular order described earlier, that is, $\phi_1 \preceq \phi_2 \preceq \cdots \preceq \phi_k \preceq \phi_1$. Under the aforementioned distributional assumptions, the CIRE of $(\phi_1, \phi_2, \ldots, \phi_k)'$ is the restricted maximum likelihood estimator (RMLE) of $(\phi_1, \phi_2, \ldots, \phi_k)'$ (Rueda et al. 2009). From Theorem 4.3.3, we see that CIRE partitions the estimates into m level sets of consecutive coordinates on which $\widetilde{\phi}_i$ is constant.

Assuming κ is known, one may derive the likelihood ratio test (LRT) statistic T for hypotheses (4.4) as the angular distance between the unconstrained maximum likelihood estimator $(\theta_1, \theta_2, \ldots, \theta_k)'$ and the RMLE $(\widetilde{\phi}_1, \widetilde{\phi}_2, \ldots, \widetilde{\phi}_k)'$ which is given by:

$$T = 2\kappa \sum_{i=1}^{k} \left(1 - \cos\left(\theta_i - \widetilde{\phi}_i\right)\right).$$

Since in practice it is not easy to implement the LRT, Fernández et al. (2012) derived a conditional test (CT) by conditioning on the number of level sets m. Conditional tests have been well studied in the case of order restricted inference for normal models (Robertson et al. 1988) but unknown until Fernández et al. (2012) for von Mises populations. The CT of Fernández et al. (2012) rejects the aforementioned null hypothesis whenever $T \geq c(m)$, where m is the number of level sets for $(\widetilde{\phi}_1, \widetilde{\phi}_2, \ldots, \widetilde{\phi}_k)'$ and $c(m)$ is chosen so that $P(\chi^2_{k-m} \geq c(m)) = \frac{\alpha}{1-\frac{1}{(n-1)!}}$. Fernández et al. (2012) demonstrated that for large values of (κ, k), CT is an α level test (see the following theorem).

Theorem 4.4.1 *Let $\phi^I = (\phi_1, ..., \phi_k)$, with $\phi_I = \pi/2, \phi_g = 3\pi/2$ for any $g \neq I$. Denote also as $(1), ..., (m)$ the level sets of $\widetilde{\Phi}$ and $R^k_m = \left\{\theta \in [0, 2\pi)^k : \widetilde{\phi} \text{ has } m \text{ level sets}\right\}$.*

(i) If $\phi = \phi^I$, then $P_{\phi^I}\left(T \geq c\,/R^k_m\right) \underset{\kappa\longrightarrow\infty}{\longrightarrow} P(\chi^2_{k-m} \geq c)$.

(ii) For large κ, the level α of the conditional test is attained at ϕ^I:
$P_{\phi^I}\left(T \geq c(m)\right) \underset{\kappa\longrightarrow\infty}{\longrightarrow} \alpha$.

(iii) If ϕ verifies the order O:
$P_{\phi}\left(T \geq c(m)\right) \underset{\kappa\longrightarrow\infty}{\longrightarrow} b$ with $b \leq G(k)\alpha$ and $G(k) \underset{k\longrightarrow\infty}{\longrightarrow} 1$.

In practice, κ is usually unknown. In this case, κ can be replaced by a consistent estimator $\widehat{\kappa}$, and accordingly T can be modified. By appealing to Mardia and Jupp (2000), pp. 87–89, $\widetilde{\phi}_i, i = 1, 2, \ldots, k$, and $\widehat{\kappa}$ are approximately independent and furthermore

$$k\frac{\kappa}{\widehat{\kappa}} \overset{\text{approx.}}{\sim} \chi^2_{k-1}.$$

As a consequence, we may approximate the distribution of CT by the central F distribution instead of the chi-squared distribution. The proof of the theorem and other theoretical details of CT are given in Fernández et al. (2012).

Barragán et al. (2013) extended the aforementioned methodology to test hypotheses regarding partial orders. More precisely, they extended the conditional test to test the following hypotheses:

$$\begin{array}{ll} H_0: & \phi_i, i = 1, ..., k \text{ follow a known partial order } O^*, \\ H_1: & H_0 \text{ is not true,} \end{array} \tag{4.5}$$

where O^* may be the partial order appearing in equation (4.3).

4.5 The estimation of a common circular order from multiple experiments

Often data are available from multiple experiments or multiple sources and researchers are interested in estimating the common order among circular parameters. For example, using data obtained from multiple experiments on fission yeast (*Schizosaccharomyces pombe*), the yeast used in brewing alcohol, researchers are not only interested in identifying periodically expressed genes but also interested in estimating their order of peak expression (see Oliva et al. 2005; Rustici et al. 2004; Peng et al. 2005).

More precisely, our problem of interest is to determine the true relative order among k angular parameters $\phi_1, \phi_2 \ldots, \phi_k$ using the corresponding unconstrained estimators $\Theta_j = (\theta_{1j}, \theta_{2j}, \ldots, \theta_{kj})'$, $j = 1, 2, \ldots, p$, from p independent experiments. Stacking these estimators, we obtain the a $k \times p$ matrix $\Theta = (\Theta_1, ..., \Theta_p)$.

As for Euclidean space data, combining data from multiple experiments to estimate a common parameter requires one to take into account variability between and within studies. However, since the underlying time course data are usually based on a large number of time points, one may assume that the variability within experiments is negligible compared to variability between experiments. Also it is important to recognize that in addition to estimating $\phi_1, \phi_2 \ldots, \phi_k$, we are more importantly interested in estimating their relative order.

The problem at hand resembles the classical problem of determining the “true” order or ranks among k objects using the ranks assigned by p independent “judges”. For example, suppose there are k gymnasts competing in an event and there are p judges assigning ranks to each of the contestants. The goal is to estimate the true rank among the k contestants using the ranks assigned by the p judges. This is a well-studied problem in the Euclidean space (cf. Diaconis and Graham 1977; Borda 1781; Condorcet 1785; Schalekamp and Zuylen 2009) and known to be NP-hard (see Bartholdi et al. 1989). Again, due to the underlying geometry, the Euclidean space-based methods cannot be directly applied here. Barragán (2014) and Barragán et al. (2014) took the first step in addressing this problem for circular data as follows.

Let $\mathfrak{O}$ denote the set of all possible orders among k objects on a unit circle. Using data from the jth experiment, let $\widetilde{\Phi}_j^{(O)} = (\widetilde{\phi}_{1j}^{(O)}, \widetilde{\phi}_{2j}^{(O)}, \ldots, \widetilde{\phi}_{kj}^{(O)})'$ denote the CIRE under the circular order O. Then the distance between Θ_j and $\widetilde{\Phi}_j^{(O)}$ is given by:

$$d(\Theta_j, O) = \mathrm{SCE}(\Theta_j, \widetilde{\Phi}_j^{(O)}) = \sum_{i=1}^{k} \left(1 - \cos\left(\theta_{ij} - \widetilde{\phi}_{ij}^{(O)}\right)\right).$$

The average distance between Θ and the estimator of $\phi_1, \phi_2 \ldots, \phi_k$ based on the p independent experiments, called mean sum of circular errors (MSCE), is given by

$$d^*(\Theta, O) = \mathrm{MSCE}(\Theta, \widetilde{\Phi}^{(O)}) = \sum_{j=1}^{p} \omega_j d(\Theta_j, O), \tag{4.6}$$

where ω_j is the weight associated with jth experiment, which is related to the precision of the experiment j. For instance, assuming $\theta_{ij} \sim \mathrm{VM}(\phi_{ij}, \kappa_j)$ with κ_j known, the weights may be defined as $\omega_j = \frac{\kappa_j}{\sum_{j=1}^{p} \kappa_j}$.

With this notation, Barragán (2014) and Barragán et al. (2014) restated the problem of estimating the optimum circular order $O^* \in \mathfrak{O}$ as the following minimization problem:

$$O^* = \arg \min_{O \in \mathfrak{O}} d^*(\Theta, O) = \arg \min_{O \in \mathfrak{O}} \sum_{j=1}^{p} \omega_j d(\Theta_j, O). \tag{4.7}$$

As done in the case of Euclidean space data (cf. Dwork et al. 2001a; 2001b), the methodology of Barragán (2014) and Barragán et al. (2014) consists of two steps as briefly outlined subsequently. For more details, one may refer to the aforementioned references. In the first step (Step E1), an initial approximate solution to the problem is obtained. This approximate solution is refined in the second step (Step E2) by smoothing out local “bumps” in the order.

Step E1 ($\widehat{O}^0$): In this step, we cast the aforementioned optimization problem as a Traveling Salesman Problem (TSP) to obtain an approximate solution to (4.7). The TSP is well studied in the graph theory literature (cf. Hahsler and Hornik 2011; Reinelt 1994; Lawler et al. 1985), and is often used in numerous applications. Starting from a particular city, a salesman is required to visit each of the remaining $k - 1$ cities in his tour exactly once and then return to the city he started. The goal for the salesman is to determine the order in which he tours the cities so that total distance traveled by the salesman is the shortest among all possible paths he can take. Even though this problem is considered to be computationally difficult, a large number of heuristics and exact methods are available in the literature. Some of these methods provide exact solutions when the number of cities is in tens of thousands and provide good approximations when the number of cities is in millions (Reinelt 1994).

In our application, each experiment is represented by a graph where the objects are the cities/nodes (or estimated angles) and the length of the edges among nodes are the angular distances between the corresponding estimated angles in the experiment. There is a correspondence between tours in the graph an circular orders within the objects. For each experiment, we have a distance matrix. We then aggregate (using means) the p matrices to summarize all the information in a single matrix. Finally, the heuristic algorithms implemented in `R` in the TSP package (Hahsler and Hornik 2011), are used to obtain the minimum length tour among nodes. The TSP solution results in an approximate circular order $\widehat{O}^0$. Not only does this strategy results in a very good approximate solution but it is also computational fast and efficient (see Barragán 2014; Barragán et al. 2014).

Step E2 ($\widehat{O}^*$): In this second step, Barragán (2014) and Barragán et al. (2014) fine-tune the solution obtained in Step E1 by performing local smoothing to reduce the MSCE (4.6). Their solution is a modification of the *Local Kemenization* algorithm that was originally developed by Dwork et al. (2001a) for Euclidean data. This modification is called *Circular*

Local Minimization. It consists of checking possible permutations in each consecutive triple of adjacent cities in the order determined in $\widehat{O}^0$. The MSCE between the new order with the permutation and the data is computed. If the MSCE for the new circular order is smaller the candidate order is appropriately updated. Each time a triple is permuted, the previous ones are checked back again to ensure that no further improvement in the order is possible.

4.6 Application: analysis of cell cycle gene expression data

A cell division cycle in a normal eukaryotic cell consists of four phases, namely, G1, S, G2, and M phases. In G1 phase, the cell rests and grows. This is also the first check point phase where any DNA damage is detected. G1 phase is followed by the S phase where DNA replication occurs. Following S phase, cells go through a second check point called G2 phase to detect damage. In a normal setting, cells that cannot be repaired are not allowed to proceed to mitosis (M phase) where the cells divide. Genes involved in cell division cycle (called cell cycle genes) have a periodic expression consistent with the duration of cell division cycle. Such genes attain peak expression just before their biological function (Jensen et al. 2006). For a given organism, biologists are typically interested in (i) identifying cell cycle genes, (ii) identifying the time to peak expression (i.e., phase angle ϕ) of a cell cycle gene, (iii) comparing the phase angles of cell cycle genes across different experimental conditions or different organisms (cf. Bähler 2005; Jensen et al. 2006; Fernández et al. 2012). A useful database containing results from various cell cycle microarray experiments is available at www.cyclebase.org, henceforth referred as cyclebase. Cyclebase provides estimates of the peak expressions by using a simple mathematical model and data from a single experiment.

To answer questions such as the aforementioned, researchers conduct long series time course gene expression studies measuring gene expressions of thousands of genes over several time points, long enough to include at least one full cell division cycle (if not more). We illustrate the methodology described in this paper using the 34 cell cycle genes *S. pombe* genes and their *Saccharomyces cerevisiae* orthologs/paralogs described in Fernández et al. (2012). We used time course data available on 10 experiments conducted on *S. pombe* in three laboratories (five by Rustici et al. (2004), three by Oliva et al. (2005) and two by Peng et al. (2005)) and six experiments conducted on *S. cerevisiae* (one experiment each by Cho et al. (1998) and de Lichtenberg et al. (2005), and two experiments each by Pramila et al. (2006) and Spellman et al. (1998)). For each gene i, $i = 1, 2, \ldots, 34$ within the jth experiment, $j = 1, 2, \ldots, 16$, we fitted the RPM to obtain the unconstrained phase angle estimates θ_{ij} for the 34 genes in the 16 experiments. Results of the estimated phase angles for the 34 genes for *S. pombe* and their *S. cerevisiae* orthologs/paralogs for the 16 experiments considered are not provided here in order to save space but can be obtained from the authors on request.

We assumed that $\theta_{ij} \sim^{\text{independent}} \text{VM}(\phi_{ij}, \kappa_j)$, where ϕ_{ij} is the true unknown phase angle for the ith gene in the jth experiment. Note that κ_j is experiment specific and not gene specific. Thus, κ_j reflects the uncertainty associated with the j^{th} experiment and phase angles of all genes within that experiment are estimated with same uncertainty. As noted earlier, since for each gene its phase angle is estimated using RPM with a reasonably large number of time points, we assume that uncertainty associated within gene is ignorable compared to the overall uncertainty associated with the experiment. The parameter κ_j is estimated using the random effects model for circular data described in Fernández et al. (2012). Using the methodology of Barragán (2014) and Barragán et al. (2014) described in

Section 4.4 and the phase angle estimates θ_{ij} of 34 genes obtained earlier for Rustici et al. (2004), Oliva et al. (2005) and Peng et al. (2005) data, we obtained the common global order among the phase angles of the 34 *S. pombe* genes. Using the estimated order, we obtained the constrained estimates of the phase angles using CIRE for the 34 *S. pombe* genes. These estimates along with the estimates according to cyclebase are given in Table 4.1. Similarly, using the phase angle estimates of the 34 *S. cerevisiae* orthologs/paralogs, based on the data from Cho et al. (1998), de Lichtenberg et al. (2005), Pramila et al. (2006) and Spellman et al. (1998), we estimated their global order along with their constrained estimates using CIRE (Table 4.1).

Using the conditional test CT, we shall compare the global order of phase angles of the aforementioned 34 genes determined by our methodology with the order described in cyclebase for the two species of yeast. The order given by cyclebase has several ties as some genes are given the same phase angle in this database. In order to determine a simple order to be compared with the one we have estimated, we broke the ties following the simple order given by our estimation process. Within each species, for each experiment we tested the null hypothesis that the global order holds against the alternative that the null is not true using the CT. Thus for each experiment, we obtain one p-value based on the CT. Within each species, we then combined p-values from all the experiments (i.e., $p = 10$ in the case of *S. pombe* and $p = 6$ in the case of *S. cerevisiae*) using Fisher's method to obtain $L = -\sum_{j=1}^{p} \log(p\text{-value}_j)$, where p-value$_j$ is the p-value obtained for experiment j. If the p-values are independently and uniformly distributed in the interval $(0, 1)$, then $2L$ is distributed as a central χ^2 random variable with p degrees of freedom. Then, if l is the observed value for L, the final p-value, p-value$_F = \text{pr}(\chi^2_p > 2l)$, yields a single value to test the null hypothesis. The resulting p-values for each species and the orders considered for each species are given in Table 4.2. From the table, we see that the orders estimated using the methodology proposed in Section 4.4 have a much higher p-value than those appearing in cyclebase. This happened not only for the global p-value$_F$ but for the almost all the p-value$_j$ values, suggesting that the global order provided by the cyclebase for the two species should be rejected and that the order derived by the methodology of Barragán (2014) and Barragán et al. (2014) described in Section 4.4 is plausible for the two species.

The disagreement between the order specified by the cyclebase and the order specified by the methodology of Barragán (2014) and Barragán et al. (2014) can partly be explained by noting that there are some major differences in the estimates of the phase angles between cyclebase and CIRE for some genes as seen in Table 4.1. Among them, the noticeable ones (identified in bold face) are the *S. pombe* gene *mcp1* and the *S. cerevisiae* gene *SST2*. According to cyclebase, both genes have a very high periodicity rank (i.e., have a poor periodic expression) and hence are likely to have less precise estimates of phase angles and hence not surprising that the two methods disagree in their phase angle estimates. (For this reason, these genes are dropped from any further study.) Since our estimator of the global order uses information from all experiments, while taking into consideration the uncertainties associated with each experiment, we believe that our estimator of the global order is more reliable.

Since the CIRE estimators have common values for some genes (those appearing in the same level set), they also yield a partial order among the genes. The partial orders given by cyclebase and by the CIRE estimator for *S. cerevisiae* appear in Table 4.3. In that table, we can see that there is no big discrepancy among the two partial orders. The most noticeable one is perhaps that of gene MOB1, which also has a high periodicity rank.

Table 4.1 Cyclebase and CIRE phase angles estimates for the two Species.

CEREVISIAE	CycleB	CIRE	POMBE	CycleB	CIRE
HTZ1	0.57	0.03	pht1	6.09	6.09
HHF1	6.16	0.03	htb1	6.22	6.22
HTA2	0.00	0.03	hta2	0.00	0.00
HTB2	0.00	0.03	hhf1	0.00	0.00
HHT2	6.09	0.03	hht3	0.06	0.06
HHT1	6.22	0.03	h3_3	0.06	0.06
KIP3	0.38	0.38	klp5	4.78	4.76
FKH1	0.63	0.63	fkh2	4.59	4.76
SWI5	1.57	1.57	ace2	4.71	4.76
BUD4	1.57	1.57	mid2	5.40	5.07
CDC5	1.57	1.78	plo1	4.27	4.76
CHS2	1.88	1.78	chs2	4.71	4.76
MYO1	1.88	1.78	myo3	4.65	4.76
HOF1	1.95	1.88	cdc15	4.71	4.76
MOB1	1.82	1.88	mob1	5.03	5.07
ASE1	1.88	1.88	**mcp1**	**3.83**	**4.76**
CDC20	2.26	2.26	slp1	4.65	4.76
KIN3	2.58	2.58	fin1	4.96	5.07
DBF2	2.70	2.70	sid2	4.78	4.76
CDC6	3.58	3.83	cdc18	4.90	5.07
PST1	3.77	3.83	SPAC1705_03C	4.65	4.76
DSE4	4.15	3.83	eng1	5.15	5.07
SST2	**3.14**	**5.01**	rgs1	4.78	4.76
RFA1	4.96	5.01	ssb1	5.22	4.76
MRC1	5.03	5.01	mrc1	5.09	4.76
SMC3	5.03	5.01	psm3	5.09	4.76
RNR1	5.03	5.01	cdc22	5.22	4.76
MSH6	5.03	5.01	msh6	5.09	5.07
POL1	5.03	5.01	pol1	5.09	5.07
RAD51	5.09	5.01	rhp51	4.96	4.76
MCD1	5.09	5.01	rad21	4.96	5.07
POL2	5.15	5.01	pol2	4.65	4.76
CLN2	5.15	5.01	cig2	5.09	4.76
SWE1	5.47	5.01	mik1	5.03	5.07

Now, we illustrate the methodology to determine a common partial order among the two species of yeasts by using a subset of orthologs/paralogs by dropping genes that have either poor periodicity in at least one of the two species (cdc18 and eng1) or by dropping genes that were considered to violate the common order according to Fernández et al. (2012) (mid2, myo3, mob1, fin1, rhp51) and the corresponding *S. cerevisiae* ortholog/paralogs appearing in bold in Table 4.3). The partial orders obtained from cyclebase and the CIRE for the remaining 25 *S. pombe* genes are summarized in Table 4.4. According to cyclebase {msh6, pol1, rad21, mik1} are activated before {ssb1, cdc22}; however, based on our methodology, {ssb1, cdc22} are activated before {msh6, pol1, rad21, mik1}. It is interesting to note from Tables 4.3 and 4.4 that the partial order derived by our methodology is satisfied by both species of yeast. Furthermore, this order is also satisfied by other previously published results (see Fernández et al. 2012).

Table 4.2 MSCE and Fp-values for the 34 core set genes considered.

Species	Order	MSCE	p-value$_F$
POMBE	Estimated order	0.06168913	0.8443571
POMBE	Cyclebase	0.08702997	4.954466e-06
CEREVISIAE	Estimated order	0.0281629	0.1659825
CEREVISIAE	Cyclebase	0.08564166	1.067372e-27

Table 4.3 Partial orders for *S. cerevisiae* Genes.

Cyclebase partial order	
Phase	Genes
G1/S	{HISTONES}$\preceq$
S/G2	{KIP3}$\preceq${FKH1}$\preceq${SWI5,**BUD4**}$\preceq${CDC5}$\preceq${**MOB1**}$\preceq$ {CHS2,**MYO1**}$\preceq${HOF1}$\preceq$
G2/M	{CDC20}$\preceq${**KIN3**}$\preceq${DBF2}$\preceq$ {**CDC6**}$\preceq${PST1}$\preceq$
M/G1	{**DSE4**}$\preceq${RFA1}$\preceq${MRC1,SMC3,RNR1,MSH6,POL1}$\preceq$ {**RAD51**,MCD1} $\preceq$ {POL2,CLN2}$\preceq${SWE1}$\preceq$
G1/S	{HISTONES}
CIRE partial order	
Phase	Genes
G1/S	{HISTONES}$\preceq$
S/G2	{KIP3}$\preceq${FKH1}$\preceq${SWI5,**BUD4**}$\preceq${CHS2,CDC5,**MYO1**}$\preceq$ {HOF1,**MOB1**}$\preceq$
G2/M	{CDC20}$\preceq${**KIN3**}$\preceq${DBF2}$\preceq$
M/G1	{**CDC6**,PST1,**DSE4**}$\preceq$ {RFA1,MRC1,SMC3,RNR1,**RAD51**,POL2,CLN2,MSH6,... ...POL1,MCD1,SWE1}$\preceq$
G1/S	{HISTONES}

Table 4.4 Partial orders for *S. pombe* Genes.

Cyclebase partial order
{HISTONES}$\preceq$ {plo1}$\preceq${fkh2}$\preceq${slp1,SPAC1705_03C,pol2}$\preceq$ {ace2,chs2,cdc15}$\preceq${klp5,sid2}$\preceq${rad21}$\preceq${mik1}$\preceq$ {mrc1,psm3,cig2, msh6,pol1}$\preceq${ssb1,cdc22}$\preceq$ {HISTONES}
CIRE partial order
{HISTONES}$\preceq$ {klp5,kfh2,ace2,plo1,chs2,cdc15,slp1,... ...sid2,SPAC1705_03C,ssb1,mrc1,psm3,cdc22,pol2,cig2}$\preceq$ {msh6,pol1,rad21,mik1}$\preceq$ {HISTONES}

The methodology developed in Barragán (2014) and Barragán et al. (2014) is also useful to study phases of genes across multiple species. In fact, those papers provide a general methodology to discover order among cell cycle genes and subsequently allow biologists to explore new hypotheses regarding functional relationships and interactions among various cell cycle genes.

In general, the circular order restricted inference methods developed in Rueda et al. (2009), Fernández et al. (2012), Barragán et al. (2013), Barragán (2014), and Barragán et al. (2014) provide a general framework and tools for cell biologists to discover new biology.

4.7 Concluding remarks and future research

In this paper, we discussed the current and ongoing research on the estimation and testing hypotheses regarding ordered parameters on a unit circle using data from multiple experiments. Although we illustrated these methods using data from cell biology, as described in the introduction, these methods are broadly applicable in a variety of contexts including, among others, evolutionary psychology (Russell 1980; De Quadros-Wander and Stokes 2007), motor behavior (Baayen et al. 2012), and circadian biology.

System biologists are often interested in developing gene networks to describe interrelationships among various genes. Much commercial software such as QIAGEN's Ingenuity$^{®}$ Pathway Analysis (IPA$^{®}$, QIAGEN Redwood City, www.qiagen.com/ingenuity) attempts to provide such networks using curated data. However, most of those networks are based on static data. They do not take into account the temporal component in the data. However, cell division cycle is a dynamic process where at each time point a collection of cell cycle genes (and others) interact and they impact on the genes that express at a later time point. The methodologies summarized in this paper describe temporal order among cell cycle genes , but it would be useful to develop dynamic networks among a collection of cell cycle genes based on the order information provided by the methods described here.

Constrained inference methods will have a natural role in other applications involving circular data, such as regression models for angular data described in Fisher and Lee (1992), Lund (1999), Downs and Mardia (2002), Kato et al. (2008), or Kato and Jones (2010). In an ongoing research project with Professor Mardia, we are exploring piecewise circular-circular regression model under constraints which may have applications in cell biology. For instance, such models would be useful to relate phase angles of cell cycle genes from different species or experimental groups.

All the methodology presented here is available in the `R` language (R Core Team 2014). Barragán et al. (2013) have developed a package called **isocir** (isotonic inference for circular data). The last version released contains `CIRE` and `cond.test` as principal functions. `CIRE` executes the algorithm developed in Rueda et al. (2009) to find the CIRE (4.2). The `R` objects called `SEXP` are used in `C++` to improve efficiency and execution time. The function `cond.test` executes the conditional test described in Fernández et al. (2012) for the hypotheses (4.5). The methodology proposed to deal with the minimization problem (4.7) has also been implemented in the `R` language as part of the new version of the **isocir** package.

Acknowledgment

This work was supported by Spanish Ministerio de Ciencia e Innovación grant (MTM2012-37129 to S.B., M.A.F. and C.R.) and Junta de Castilla y León, Consejería de Educación and the European Social Fund (within the Programa Operativo Castilla y León 2007-2013 to S.B.) and the Intramural Research Program of the National Institute of Environmental Health Sciences (Z01 ES101744-04 to S.D.P.).

References

Baayen C, Klugkist I and Mechsner F 2012 A test for the analysis of order constrained hypotheses for circular data. *Journal of Motor Behavior* **44**(5), 351–363.

Bähler J 2005 Cell-cycle control of gene expression in budding and fission yeast. *Annual Review of Genetics* **39**, 69–94.

Barragán S 2014 *Procedimientos estadísticos para modelos circulares con restricciones de orden aplicados al análisis de expresiones de genes*. Universidad de Valladolid. Ph.D. Dissertation.

Barragán S, Fernández M, Rueda C and Peddada S 2013 isocir: an R package for constrained inference using isotonic regression for circular data, with an application to cell biology. *Journal of Statistical Software* **54**(4), 1–17.

Barragán S, Rueda C, Fernández M and Peddada S 2014 Statistical framework for determining the temporal program in an oscillatory system. *Preprint*.

Bartholdi J, Tovey C and Trick M 1989 Voting schemes for which it can be difficult to tell who won the election. *Social Choice Welfare* **6**, 157–165.

Borda J 1781 *Memorie sur les elections au scrutin*. Historie de l'Academie.

Cao Y, Chen A, Jones R, Radcliffe J, Dietrich K, Caldwell K, Peddada S and Rogan W 2011 Efficacy of sucimer chelation of mercury at background exposures in toddlers: a randomized trial. *Journal of Pediatrics* **158**(3), 480–485.

Cho R, Campbell M, Winzeler E, Steinmetz L, Conway A, Wodicka L, Wolfsberg T, Gabrielian A, Landsman D, Lockhart D and Davis R 1998 A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **2**(1), 65–73.

Conaway M, Dunbar S and Peddada S 2004 Designs for single or multiple agent phase I trials. *Biometrics* **60**, 661–669.

Conde D, Fernández M, Rueda C and Salvador B 2012 Classification of samples into two or more ordered populations with application to a cancer trial. *Statistics in Medicine* **31**(28), 3773–3786.

Conde D, Salvador B, Rueda C and Fernández M 2013 Performance and estimation of the true error rate of classification rules built with additional information. An application to a cancer trial. *Statistical Applications in Genetics and Molecular Biology* **12**(5), 583–602.

Condorcet MJ 1785 *Essai sur l application de l'analyse á la probabilité des décisions rendues á la pluralité des voix*.

de Lichtenberg U, Wernersson R, Jensen T, Nielsen H, Fausboll A, Schmidt P, Hansen F, Knudsen S and Brunak S 2005 New weakly expressed cell cycle-regulated genes in yeast. *Yeasts* **22**(5), 1191–1201.

De Quadros-Wander S and Stokes M 2007 The effect of mood on opposite-sex judgments of males' commitment and females' sexual content. *Evolutionary Psychology* **4**, 453–475.

Diaconis P and Graham R 1977 Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society Series B* **39**(2), 262–268.

Downs T and Mardia K 2002 Circular regression. *Biometrika* **89**(3), 683–697.

Dwork C, Kumar R, Naor M and Sivakumar D 2001a Rank aggregation methods for the web. *Proceedings of the 10th International World Wide Web Conference*, pp. 613–622.

Dwork C, Kumar R, Naor M and Sivakumar D 2001b Rank aggregation revisited. *Manuscript*.

Fernández M, Rueda C and Peddada S 2012 Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Research* **40**(7), 2823–2832.

Fisher N 1993 *Statistical Analysis of Circular Data*. Cambridge University Press.

Fisher N and Lee A 1992 Regression models for an angular response. *Biometrics* **48**, 665–677.

Forgas J 1998 On being happy and mistaken: mood effects on the fundamental attribution error. *Journal of Personality and Social Psychology* **75**(2), 318–331.

Hahsler M and Hornik K 2011 *Traveling Salesperson Problem (TSP).* R package version 1.0-6.

Hoenerhoff M, Pandiri A, Snyder S, Hong H, Ton T, Peddada S, Shockley K, Chan P, Rider C, Kooistra L, Nyska A and Sills R 2013 Hepatocellular carcinomas in B6c3F1 mice treated with Ginkgo biloba extract for two years differ from spontaneous liver tumors in cancer gene mutations and genomic pathways. *Toxicologic Pathology* **41**(6), 826–841.

Hughes M, DiTacchio L, Hayes K, Vollmers C, Pulivarthy S, Baggs J, Manda S and Hogenesch J 2009 Harmonics of circadian gene transcription in mammals. *PLoS Genetics* **5**(4), e1000442.

Jensen J, Jensen T, Lichtenberg U, Brunak S and Bork P 2006 Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* **443**, 594–597.

Kato S and Jones M 2010 A family of distributions on the circle with links to, and applications arising from, Möbius transformation. *Journal of the American Statistical Association* **105**(489), 249–262.

Kato S, Shimizu K and Shieh G 2008 A circular-circular regression model. *Statistica Sinica* **18**, 633–645.

Lawler E, Lenstra J, Rinnooy Kan A and Shmoys D (eds) 1985 *The Traveling Saleman Problem*. John Wiley & Sons.

Liu D, Umbach D, Peddada S, Li L, Crockett P and Weinberg C 2004 A random periods model for expression of cell-cycle genes. *Proceedings of the National Academy of Sciences of the United States of America* **101**(19), 7240–7245.

Lund U 1999 Least circular distance regression for directional data. *Journal of Applied Statistics* **26**(6), 723–733.

Mardia K and Jupp P 2000 *Directional Statistics*. John Wiley & Sons.

Mechsner F, Kerzel D, Knoblich G and Prinz W 2001 Perceptual basis of bimanual coordination. *Nature* **414**, 69–73.

Oliva A, Rosebrock A, Ferrezuelo F, Pyne S, Chen H, Skiena S, Futcher B and Leatherwood J 2005 The cell-cycle-regulated genes of Schizosaccharomyces pombe. *PLOS Biology* **3**, 1239–1260.

Oullier O, Bardy B, Stoffregen T and Bootsma R 2002 Postural coordination in looking and tracking tasks. *Human Movement Science* **21**, 147–167.

Peddada S, Dinse G and Kissling G 2007 Incorporating historical control data when comparing tumor incidence rates. *Journal of the American Statistical Association* **102**, 1212–1220.

Peddada S, Dunson D and Tan X 2005 Estimation of order-restricted means from correlated data. *Biometrika* **92**, 703–715.

Peddada S, Lobenhofer L, Li L, Afshari C, Weinberg C and Umbach D 2003 Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* **19**, 834–841.

Peng X, Karuturi R, Miller L, Lin K, Jia Y, Kondu P, Wang L, Wong L, Liu E, Balasubramanian M and Liu J 2005 Identification of cell cycle-regulated genes in fission yeast. *American Society for Cell Biology* **16**, 1026–1042.

Perdivara I, Peddada S, Miller F, Tomer K and Deterding L 2011 Mass spectrometric determination of IgG subclass-specific glycosylation profiles in siblings discordant for myositis syndromes. *Journal of Proteome Research* **10**, 2969–2978.

Posner J, Russell J and Peterson B 2005 The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology* **17**, 715–734.

Pramila T, Wu W, Miles S, Noble W and Breeden LL 2006 The forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes & Development* **22**(16), 2266–2278.

R Core Team 2014 *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0.

Reinelt G 1994 *The Traveling Salesman. Computational Solutions for TSP Applications*. Springer-Verlag.

Robertson T, Wright F and Dykstra R 1988 *Order Restricted Statitical Inference*. John Wiley & Sons.

Rueda C, Fernández M and Peddada S 2009 Estimation of parameters subject to order restrictions on a circle with application to estimation of phase angles of cell-cycle genes. *Journal of the American Statistical Association* **104**(485), 338–347.

Russell J 1980 A circumplex model of affect. *Journal of Personality and Social Psychology* **39**(6), 1161–1178.

Rustici G, Mata J, Kivinen K, Lio P, Penkett C, Burns G, Hayles J, Brazma A, Nurse P and Bahler J 2004 Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics* **36**, 809–817.

Schalekamp F and Zuylen A 2009 Rank aggregation: together we are strong. In *Proceedings of the 11th ALENEX*, pp. 38–51.

Schlosberg H 1952 The description of facial experssions in ternos of two dimensions. *Journal of Experimental Psychology* **44**, 229–237.

Silvapulle M and Sen P 2005 *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. John Wiley & Sons.

Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D and Futcher B 1998 Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9**(12), 3273–3297.

Whitfield M, Sherlock G, Saldanha A, Murray J, Ball C, Alexander K, Matese J, Perou, C.M., Hurt M, Brown P and Botstein D 2002 Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell* **13**, 1977–2000.

5

Parametric circular–circular regression and diagnostic analysis

Orathai Polsen[1] and Charles C. Taylor[2]

[1]*Department of Applied Statistics, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand*

[2]*Department of Statistics, University of Leeds, Leeds, UK*

5.1 Introduction

Circular regression has been used and applied in many areas. For example, in medicine, it is interesting to know the relationship between the peak times for two successive measurements of diastolic blood pressure (Downs 1974). In studying earthquakes, it may be useful to know whether the direction of ground movement is related to the direction of the steepest descent (Rivest 1997). In a marine biology study, it is often of interest to observe a relationship between spawning time and time of low tide (Lund 1999).

Jammalamadaka and Sarma (1993) proposed and explored a regression model for the case of a circular response variable and a circular explanatory variable. Their (bivariate) regression model is expressed by trigonometric polynomial functions of degree m and their suggested method for estimating parameters is based on least squares. Rivest (1997) presented a circular–circular regression model for decentred predictor. Downs and Mardia (2002) proposed a model for circular–circular regression by using a Möbius transformation to obtain a regression curve. Kato et al. (2008) provided a circular–circular regression model, which also uses a Möbius transformation but the errors are assumed to follow a wrapped Cauchy distribution. Taylor (2009) proposed a regression model using a slightly

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

more general transformation and also extended to a polynomial regression model and a multiple regression model. Kato and Jones (2010) proposed a regression curve, which is an extension of the regression models of Downs and Mardia (2002) and Kato et al. (2008). In order to choose between models, it is helpful to know the similarities and differences between them as well as some properties. These are studied in this chapter, including parameter estimation. In addition, we consider some diagnostic tools for circular regression.

This chapter is organized as follows. In Section 5.2, we review similarities and differences of existing circular–circular regression models. A useful strategy for estimating parameters in circular–circular regression context is introduced in Section 5.3. In Section 5.4, we investigate diagnostic analysis. Jammalamadaka and Sarma (1993) provide methods for identification of outliers in circular–circular regression. Here, we firstly investigate an approach for checking the von Mises distribution assumption and introduce a method for detecting influential observations, which can be used more generally. A practical example and a simulated example are given in Section 5.5. We conclude with a discussion.

5.2 Review of models

In this section, we briefly survey existing circular–circular regression models and set these into a common framework; this will allow us to highlight similarities and differences. Let Y be a circular response variable and X be a circular explanatory variable, where X and Y take values in the circle S_1 conveniently represented as the interval $[-\pi, \pi)$ or as the real numbers mod 2π. A general regression model is expressed as follows:

$$\begin{aligned} y_i &= \mu(x_i;\psi) + \varepsilon_i, \quad i = 1,\ldots,n \\ &= \text{atan2}\{g_2(x_i;\psi), g_1(x_i;\psi)\} + \varepsilon_i \pmod{2\pi}, \end{aligned} \tag{5.1}$$

say, where $\mu(\cdot)$ represents the conditional mean direction of y given x, ψ is the vector of all parameters, ε is the angular error and the function atan2(v, u) returns the angle between the x-axis and the vector from the origin to (u, v). This is undefined when $v = u = 0$. It may be noted that $g_1(\cdot)$ and $g_2(\cdot)$ are not uniquely identifiable in (5.1) since atan2$(v, u) =$ atan2(cv, cu) for $c > 0$.

General expressions for $g_j(x)$ are trigonometric polynomial functions of degree m,

$$g_j(x;\psi) = a_{j0} + \sum_{k=1}^{m}(a_{jk}\cos kx + b_{jk}\sin kx), \quad j = 1, 2, \tag{5.2}$$

where $\psi^T = (a_{10},\ldots,a_{1m},a_{20},\ldots,a_{2m},b_{11},\ldots,b_{1m},b_{21},\ldots,b_{2m})$ are $4m + 2$ parameters of the model.

Jammalamadaka and Sarma (1993) suggested to consider transformations $y_1 = \cos y$ and $y_2 = \sin y$ and then regressing y_i on the parameters in $g_i(x)$. However, this approach will not lead to fitted values $\hat{y}_i, i = 1, 2$ which satisfy $\hat{y}_1^2 + \hat{y}_2^2 = 1$, or even $|\hat{y}_i| \leq 1$. Moreover, the use of least squares is problematic since there is a lack of independence between $\cos y$ and $\sin y$, and there will also be heteroscedasticity in the errors. However, even though it is not straightforward to use a bivariate regression model, we nevertheless consider this strategy as an *ad hoc* method, which we discuss further in Section 5.3.

Various models have been proposed in the past 20 years, which can be recast in the form of (5.1) with various forms for g_1 and g_2 and for the angular error distribution. In particular, Rivest (1997), Downs and Mardia (2002), and Taylor (2009) have proposed models that take the form as follows:

$$\begin{aligned}
&\text{Rivest:}\quad \mu(x;\alpha,\beta,r) = \beta + \text{atan2}(\sin(x-\alpha), r+\cos(x-\alpha)),\\
&\text{Downs and Mardia:}\quad \mu(x;\alpha,\beta,\omega) = \beta + 2\text{atan}\{\omega\tan[(x-\alpha)/2]\},\\
&\text{Taylor:}\quad \mu(x;\alpha,\beta,a,b) = \beta + \text{atan2}(a\sin(x-\alpha), b\cos(x-\alpha)+1),
\end{aligned} \tag{5.3}$$

where r, a, b are real numbers, $\omega \in [-1, 1]$ is a slope parameter, and α and β are angular location parameters. The mean function μ is centered on (α, β). We note that Rivest (1997) also introduced an apparent five-parameter model, though it was not considered further.

We briefly review these models and establish a framework that highlights the similarities and differences. It can be easily shown that all three models in Equation (5.3) can be re-written in the form of (5.1) in which

$$\begin{aligned}
g_1(x) &= \cos\beta + (B\cos\beta\cos\alpha + A\sin\beta\sin\alpha)\cos x + (B\cos\beta\sin\alpha\\
&\quad - A\sin\beta\cos\alpha)\sin x,\\
g_2(x) &= \sin\beta + (B\sin\beta\cos\alpha - A\cos\beta\sin\alpha)\cos x + (B\sin\beta\sin\alpha\\
&\quad + A\cos\beta\cos\alpha)\sin x.
\end{aligned}$$

These are both of the form given by Equation (5.2) and so the apparent six parameters $a_{10}, a_{11}, a_{20}, a_{21}, b_{11}, b_{21}$ are determined by only four parameters A, B, α, β. Moreover, the correspondence of the parameters A and B for each of the aforementioned models is given in Table 5.1.

The following properties are then easily seen:

(i) The Rivest model constrains $A = B$, obviously reducing it to a three parameter model.

(ii) For the Downs & Mardia model, both A and B are determined by ω. Also $B \geq 1$ since $|\omega| \leq 1$.

(iii) The regression curve is continuous, unless $|B| = 1$.

(iv) The curve passes through the locations $(\alpha, \beta - \pi(I[B > -1] - 1))$ and $(\alpha + \pi, \beta + \text{sign}(A)\pi I[B > 1])$.

Table 5.1 The parameters A, B and the gradient at $x = \alpha$ and $x = \alpha + \pi$ for each of the models.

Model	Parameters		Gradients	
	A	B	$x = \alpha$	$x = \alpha + \pi$
Rivest	$1/r$	$1/r$	$1/(1 + r)$	$1/(1 - r)$
Downs and Mardia	$2\omega/(1 - \omega^2)$	$(1 + \omega^2)/(1 - \omega^2)$	ω	$1/\omega$
Taylor	a	b	$a/(b + 1)$	$a/(b - 1)$

(v) The turning points are at $(\alpha + \text{atan2}(\pm\sqrt{1-B^2}, -B), \beta + \text{atan2}(\pm A\sqrt{1-B^2}, 1-B^2))$ which are real only if $|B| < 1$.

(vi) The behavior of the model with $|B| < 1$ and the model with $|B| > 1$ are very different. As angle x changes in clockwise direction, $\mu(\cdot)$ oscillates (clockwise and anticlockwise) in a range less than π for the case $|B| < 1$, whereas $\mu(\cdot)$ changes in a uniform direction (clockwise for $AB > 0$ and anticlockwise for $AB < 0$) for the case $|B| > 1$.

(vii) Note that $\mu(x; \alpha, \beta, A, B) = \mu(x; \alpha + \pi, \beta, -A, -B)$ for $B < 0$. This means we can restrict attention to the model with $B \geq 0$ without loss of generality.

Note that properties (v) and (vi) are not relevant to Downs and Mardia's model since

$$|B| = \left|\frac{1+\omega^2}{1-\omega^2}\right| \geq 1 \quad \text{for} \quad |\omega| \leq 1.$$

The second ingredient of the model is the distribution of errors. In the papers of Rivest (1997), Downs and Mardia (2002), and Taylor (2009), they focus on the model in which the angular error is distributed as a von Mises distribution. Before considering other distributions of errors, it is worth noting an alternative representation of the model in Equation (5.1), which uses complex numbers (Kato et al. 2008; Kato and Jones 2010). The conditional mean formulation given by Downs and Mardia (2002), Kato et al. (2008), and Kato and Jones (2010) are the same, but the error distributions are different. The angular error is distributed as a wrapped Cauchy distribution in Kato et al. (2008's) model, whereas the error distribution of Kato and Jones (2010's) model is in a family of four-parameter distributions on the circle.

Finally, we note the following points for a general model (5.2) :

(a) A parameterization that makes use of all a_{jk} and b_{jk} will lead to obvious redundancies since dependencies exist, for example, scale all parameters by constant.

(b) Two seemingly different sets of parameters can lead to very similar predicted values.

(c) Alternative representations, although equivalent, may have parameters that are not so easy to interpret.

5.3 Parameter estimation and inference

Given an error distribution, the maximum likelihood function is easily expressed as a function of parameters. However, in maximizing the likelihood function, the model must avoid obvious redundancies. If the number of parameters is large relative to the number of observations, unless there is high concentration, it will be numerically difficult to find the maximum likelihood estimators. In our experience using simulated data, the R optimization routines nlm and optim can sometimes reach local maxima rather globally optimal solutions. The success will depend on the parameterization used as well as the number of parameters.

One possibility is simply to try several starting values, but another useful heuristic strategy that seems to work well in practice is to *initially* follow the approach of Jammalamadaka and Sarma (1993) and use least squares to fit

$$y_1 = g_1(x) + e_1,$$
$$y_2 = g_2(x) + e_2,$$

where $y_1 = \cos y$ and $y_2 = \sin y$ and $g_j(x)$ are in general form (5.2). Using this approach leads to a vector of least squares estimates, $\hat{\psi}$, and standard errors (computed in the standard way) for each estimate, $s_{\hat{\psi}_l}$; see (Jammalamadaka and SenGupta 2001, p. 191). We note that the fact that the fitted $\hat{y}_i$ do not lie on the circle is not so problematic, since, as previously noted, atan2(v, u) = atan2(cv, cu) for $c > 0$. This strategy was found to be useful in selecting an appropriate value of m, but also for removing redundancies in the final model through standard variable selection procedures.

Our proposed method is to find the coefficient that has the largest absolute value of a t-ratio, say $\hat{\psi}_p$, such that

$$p = \text{argmax}_{l=1,\ldots,4m+2} \left\{ \left| \frac{\hat{\psi}_l}{s_{\hat{\psi}_l}} \right| \right\}.$$

Then, in view of point (a) discussed earlier, we compute standardized parameters, $\hat{\psi}'_l = \hat{\psi}_l / \hat{\psi}_p$ such that all parameters are multiplied by a constant. We then treat $\hat{\psi}'_p \equiv 1$ as a fixed value. In the second stage, we assume a von Mises distribution for the errors and maximize the log-likelihood function

$$L(\psi'_{(-p)}) = \kappa \sum_{i=1}^{n} \cos\left\{y_i - \text{atan2}\left(g_2(x_i), g_1(x_i)\right)\right\} + \text{constant}$$

over $4m + 1$ parameters, $\psi'_{(-p)}$, where $\psi'_{(-p)}$ is the vector of parameters left for consideration after fixing $\psi'_p = 1$. The maximum likelihood estimators must be computed numerically and suitable starting values can be defined by

$$\psi^{(0)} = \hat{\psi}'_{(-p)}.$$

The validation of the Hessian matrix for obtaining the standard errors of parameter estimates after maximizing the likelihood (in a circular context) was investigated via a small simulation. It was found that the Hessian matrix was invariant to changes in angular location (x or y) of the data, and it can also be used to perform variable reduction by stepwise method. Alternatively, stepwise variable selection can be carried out using AIC or BIC.

In simulations, this approach worked well for large κ, but in other cases the issue of dependencies can cause trouble. However, the reliability of the solution was greatly improved when the parameters were normalized as described earlier. It could be that use of some MCMC scheme could help to overcome the problem of multiple solutions, but this would need to be investigated further.

5.4 Diagnostic analysis

The aim of this section is twofold. First, we introduce two statistical tests for checking the distribution assumption of circular residuals. The second goal is to investigate ways for detecting *influential* observations in circular–circular regression.

5.4.1 Goodness-of-fit test for the von Mises distribution

In this section, our main aim is to compare the power of two tests for the von Mises distribution. The first test is Watson's U^2 test, proposed by Watson (1961), and it is an analogue, for circular data, of the Craḿer-von Mises test. The test statistic given by

$$U^2 = \sum_{i=1}^{n} \left\{ u_{(i)} - \frac{(2i-1)}{2n} \right\}^2 - n\left(\bar{u} - \frac{1}{2}\right)^2 + \frac{1}{12n},$$

where $u_i = F(\theta_i; \hat{\mu}, \hat{\kappa})$; $\hat{\mu}, \hat{\kappa}$ are the MLE of μ and κ, respectively, $u_{(i)}$ denotes the ordered u_i and $\bar{u} = \sum u_i/n$. The table of critical values for this test is given by Lockhart and Stephens (1985).

The second test is a score test proposed by Cox (1975) and then documented in Barndorff-Nielsen and Cox (1989) and Mardia and Jupp (2000, pp. 142–143). However, there are a few discrepancies in these three publications; all have (different) misprints. The von Mises distribution can be extended to a density function, which is proportional to

$$\exp(\alpha_1 \cos\theta + \alpha_2 \sin\theta + \beta_1 \cos 2\theta + \beta_2 \sin 2\theta).$$

To test the von Mises distribution, the hypothesis $\beta_1 = \beta_2 = 0$ is examined. Considering the conditional distribution of $V = (\sum \cos 2\theta_i, \sum \sin 2\theta_i)$ given $U = (\sum \cos\theta_i, \sum \sin\theta_i)$ leads to the test statistic

$$S = \frac{s_c^2}{nv_c(\hat{\kappa})} + \frac{s_s^2}{nv_s(\hat{\kappa})},$$

where s_c and s_s are defined by

$$s_c = \sum_{i=1}^{n} \cos\{2(\theta_i - \hat{\mu})\} - nA_2(\hat{\kappa}) \quad \text{and} \quad s_s = \sum_{i=1}^{n} \sin\{2(\theta_i - \hat{\mu})\}.$$

Here

$$A_2(\hat{\kappa}) = \frac{I_2(\hat{\kappa})}{I_0(\hat{\kappa})},$$

where $I_p(\kappa)$ is the modified Bessel function of the first kind and order p, and s_c and s_s are independently normally distributed with zero mean and variance $nv_c(\hat{\kappa})$ and $nv_s(\hat{\kappa})$ respectively, where

$$v_c(\hat{\kappa}) = \frac{I_0^2 + I_0 I_4 - 2I_2^2}{2I_0^2} - \frac{(I_0 I_3 + I_0 I_1 - 2I_1 I_2)^2}{2I_0^2(I_0^2 + I_0 I_2 - 2I_1^2)},$$

$$v_s(\hat{\kappa}) = \frac{(I_0 - I_4)(I_0 - I_2) - (I_1 - I_3)^2}{2I_0(I_0 - I_2)},$$

and $I_p = I_p(\hat{\kappa})$. Therefore, the asymptotic distribution of S is chi-squared with two degrees of freedom.

A simulation study was conducted to investigate type I errors of the aforementioned two tests and then compare the power of the tests under some alternatives. We first generated data that have a von Mises distribution with $\mu = 0$ and $\kappa = 0.5, 2, 3, 4$. We used sample sizes $n = 20, 30, 50, 100, 200, 500$ and the number of replications in the simulations was

Table 5.2 Type I errors for various values of κ and n. Nominal level of test $\alpha = 0.05$.

	$\kappa = 0.5$		$\kappa = 2$		$\kappa = 3$		$\kappa = 4$	
n	U^2	S	U^2	S	U^2	S	U^2	S
20	0.050	0.053	0.055	0.045	0.053	0.036	0.049	0.039
30	0.043	0.049	0.050	0.043	0.050	0.040	0.051	0.040
50	0.055	0.057	0.049	0.048	0.053	0.041	0.053	0.045
100	0.046	0.052	0.049	0.052	0.059	0.045	0.047	0.049
200	0.052	0.051	0.049	0.047	0.063	0.044	0.049	0.045
500	0.051	0.055	0.049	0.053	0.057	0.046	0.049	0.054

5000. The results of the simulation are shown in Table 5.2. As can be seen both tests can control type I errors (the nominal level is 0.05) in most settings.

We then examined the power of these tests via a further simulation study. In this case, we generate data from an alternative hypothesis, which is not von Mises. Here, we chose to use data that has a mixture of von Mises distributions with two components with $\mu_1 = 0, \mu_2 = 3\pi/4$, $\kappa_1 = \kappa_2 = 0.5, 2, 3, 4$ and a mixing proportion of 0.5. We considered sample sizes $n = 20, 30, 50, 100, 200, 500$ with the number of replications equal to 5000. The results of the simulation are shown in Table 5.3. It can be seen that the power of the score test is greater than the power of Watson's U^2 test when κ is small but that their powers get closer as n and κ increase.

5.4.2 Influential observations

In this section, we investigate a possible way to detect an influential observation in circular–circular regression. Cook (1977) investigated the detection of influential observations in linear regression and proposed a distance measure for judging the influential observations on the basis of difference between the parameter estimates with and without the ith data point. However, we cannot easily apply this to circular–circular regression, since some of the parameters in a circular model are not on a line. So, we would need to obtain a "mixed distance" with some measured on the circle, and others on a linear scale. More importantly, in circular regression, some models that have a fairly large difference in magnitude of parameters can look similar in shape of regression curve. Therefore, we will consider an approach based on the likelihood function to identify influential observations.

Table 5.3 The power of the test for various values of κ and n.

	$\kappa_1 = \kappa_2 = 0.5$		$\kappa_1 = \kappa_2 = 2$		$\kappa_1 = \kappa_2 = 3$		$\kappa_1 = \kappa_2 = 4$	
n	U^2	S	U^2	S	U^2	S	U^2	S
20	0.049	0.053	0.231	0.278	0.529	0.617	0.799	0.868
30	0.051	0.055	0.353	0.417	0.759	0.837	0.950	0.975
50	0.052	0.060	0.574	0.649	0.949	0.973	0.998	0.999
100	0.058	0.061	0.892	0.932	0.999	1	1	1
200	0.057	0.067	0.998	0.999	1	1	1	1
500	0.088	0.101	1	1	1	1	1	1

As with Cook's distance, values will be computed with and without the ith observation in order to measure influence.

Let $L(\hat{\psi})$ and $L(\hat{\psi}_{(i)})$ denote the log-likelihoods of all observations. The former is based on the maximum likelihood estimates obtained by using all observations, while the latter is based on the maximum likelihood estimates obtained by omitting the ith observation. We consider a test statistic given by

$$D_i = L(\hat{\psi}) - L(\hat{\psi}_{(i)}).$$

It is straightforward to show that the distribution of D_i for a random sample from a *normal* population is a scaled chi-square, that is $D_i \sim a\chi^2_\nu$, where $a = 1/(2(n-1))$ and $\nu = 1$ are the parameters of a scaled chi-square.

A simulation study was used to investigate the distribution of D_i in various circular–circular regression models, in which all the data were simulated from the model. We carry out a goodness of fit test for a scaled chi-square distribution of statistic D_i by using the Kolmogorov–Smirnov test where the estimates of a and ν are computed to match the first two moments, i.e.

$$\hat{a} = \frac{s^2}{2\bar{d}} \quad \text{and} \quad \hat{\nu} = \frac{\bar{d}}{\hat{a}},$$

where $\bar{d}$ and s^2 are sample mean and sample variance of D, respectively. Simulation results suggest that the distribution of D_i in circular–circular regression models is also a scaled chi-square. However, this statement will only be true if the null hypothesis (H_0) – that there are no influential observations – is correct. If outliers are present, this will affect both s^2 and $\bar{d}$, so estimates of the parameters of this distribution will result in a loss of sensitivity for the detection of outliers among the D_i and, of course, the outlying observation itself will not follow this distribution. To solve this, we consider the concept of censored data, which is akin to that of the trimmed mean, which is often used as a robust location measure. Let

$$D_{(1)} \leq D_{(2)} \leq \cdots \leq D_{(j)} \leq D_{(j+1)} \leq \cdots \leq D_{(n-1)} \leq D_{(n)}$$

be the order statistics.

To allow for the fact that some – say the largest $n-j$ – of these D_i may be influential, we consider a log-likelihood function of the parameters to be given by

$$L(a,\nu) = \log\left\{\prod_{i=1}^{j} f(D_{(i)} \mid a,\nu)[1 - F(t \mid a,\nu)]^{n-j}\right\}, \tag{5.4}$$

where t is a value chosen (independently of a and ν) to threshold the (largest j) observations, which – in this case – will be outliers, and $1 - F(t \mid a,\nu) = \int_t^\infty f(x \mid a,\nu)dx$, with f given by the scaled chi-square, $a\chi^2_\nu$. As can be seen, it is necessary to define the value of t. In the case where there are no outliers, then a larger value of t will lead to less precise (though still unbiased) estimates of a and ν, whereas if there are outliers, then a smaller value of t will lead to a "corrupt" sample and biased estimates. We first investigate the selection of t by using sample quantiles (based on the data) corresponding to values from 0.90 to 0.99. The results are shown in Table 5.4. From this simulation study in which there were no outliers, it was found that the estimates of a and ν were similar over this range of sample quantiles. Based on these results, we fixed t equal to the sample quantile corresponding to 0.90 for the remainder of this chapter.

Table 5.4 The estimates of parameters a and ν.

	Quantile									
	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
$\hat{a}$	0.054	0.052	0.050	0.049	0.048	0.047	0.045	0.044	0.043	0.042
$\hat{\nu}$	0.870	0.879	0.888	0.895	0.904	0.910	0.918	0.925	0.934	0.941

Taking $f(\cdot)$ to be a scaled chi-square in Equation (5.4), the log-likelihood is

$$\begin{aligned}L(a,\nu) = &-\frac{1}{2a}\sum_{i=1}^{j} D_{(i)} + \frac{(\nu-2)}{2}\sum_{i=1}^{j}\log D_{(i)}\\ &+ (n-j)\log\left[\int_t^{\infty}\frac{(x/a)^{(\nu-2)/2}\exp(-x/2a)}{a2^{\nu/2}\Gamma(\nu/2)}dx\right]\\ &- j\log[(2a)^{\nu/2}\Gamma(\nu/2)]\end{aligned}$$

and then the estimates of parameters a and ν can be obtained by maximizing this function.

Finally, to obtain a critical value, we consider the distribution of the nth order statistic because if we use the aforementioned scaled chi-square, even when H_0 is true, the largest observation will generally be detected as an influential observation. Therefore, the distribution of the nth order statistic needs to be evaluated. Let $Y = \max_i D_i$, then

$$P(Y > y) = 1 - [P(D_i \le y)]^n.$$

Hence, a critical value can be computed using the following Expression:

$$P(Y > y) = 1 - \left[\int_0^{y}\frac{(x/a)^{(\nu-2)/2}\exp(-x/2a)}{a2^{\nu/2}\Gamma(\nu/2)}dx\right]^n.$$

It should be noted that this is a two-stage procedure. The computation of the D_i does not involve identification of outliers. It is only the estimation of the scaled χ^2 parameters, which is influenced by the presence of larger values in the resulting set of D_i. If the threshold (t) is chosen incorrectly (so that outlying D_i are deemed to satisfy H_0), then this will clearly have an impact of the final p-values. However, a conservative choice of t will only lead to a small loss of efficiency. This is a very similar position to the amount of exclusion applied in a trimmed mean estimate.

In the case that there are several influential observations, then a sequential procedure could be attempted, though this may not work in the case that there is a cluster of similar-valued observations which have undue influence.

5.5 Examples

In this section, two examples are used to compare some existing models and illustrate the influential diagnostic in circular–circular regression, respectively.

Example 1. The wind directions at 6.00 and 12.00 a.m. are considered in order to consider how the wind direction at 12.00 a.m. is related to the wind direction at 6 a.m. These wind direction data were used in work of Kato and Jones (2010) and used for illustration of a nonparametric regression model in DiMarzio et al. (2013). They are part of full dataset that was measured at a weather station in Texas. The dataset contains hourly resolution surface meteorological data from the Texas National Resources Conservation Commission (TNRCC) Air Quality Monitoring Network. This data covers the period from 20 May to 31 July 2003 and is provided by NCAR/EOL under sponsorship of the National Science Foundation, available at data.eol.ucar.edu/codiac/dss/id=85.034.

We use the model proposed by Taylor (2009) for regressing the wind direction at 12.00 a.m. on the wind direction at 6.00 a.m. Figure 5.1(a) shows that the estimated regression model of Taylor's model represents the relationship between the wind direction at 6.00 a.m. and 12.00 a.m. reasonably well. A residual analysis was conducted and the

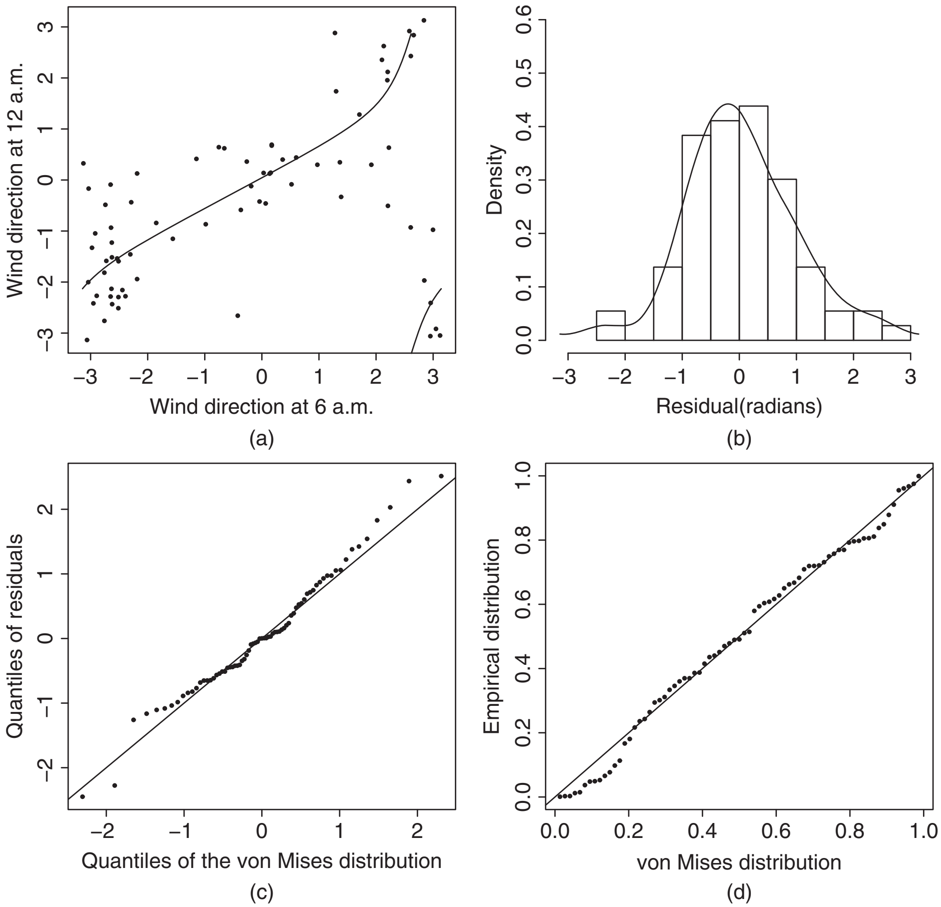


Figure 5.1 Top: plot of the wind direction at 6.00 a.m. and 12.00 a.m. with fitted regression line (a), and histogram of the residuals with the kernel density estimate (b), Bottom: Q-Q plot (c) and P-P plot (d).

results are also shown in the other panels of Figure 5.1. In addition, the goodness-of-fit test for von Mises was computed using the test statistic S. Here $S = 3.13$, which is less than the critical value of χ^2_2, confirming that the angular errors are compatible with the von Mises distribution. Finally, we calculated D_i and the critical value. All values of D_i are less than the critical value $(d_c = 0.509)$, so there is no influential observation in this dataset.

The maximum likelihood estimates (with standard errors), the maximum log-likelihood, AIC and BIC values for Taylor's model are given in Table 5.5. It can be seen that the estimate of β is less than two standard errors. Refitting the model with β constrained to be zero leads to a slightly improved AIC and BIC.

Table 5.6 shows the maximum likelihood estimates, the maximum log-likelihood, AIC and BIC values for model proposed by Jammalamadaka and SenGupta (2001). In this case, we simply apply the procedure as a bivariate regression, taking no account of over-parameterization. Even so, it can be seen that the likelihood is somewhat less than that in Table 5.5.

We then compare these models to the ones in Kato and Jones (2010's) paper and these are reproduced in Table 5.7 (the entries in parentheses are the constrained value of the parameters). According to the AIC criterion, AIC for the Kato and Jones's model is 201.4, which is lower than the others, while the Downs and Mardia's model (a von Mises case in Table 5.7) has lower BIC for this data set.

Example 2. In this example, a simulated data set is used to illustrate our method for detecting influential observations. We generated data from the following model:

$$y_i = \text{atan2}(2\sin x_i, 0.1\cos x_i + 1) + \varepsilon_i,$$

where $\varepsilon_i \sim \text{vM}(0, 10)$ and $i = 1, \ldots, 100$. Then, we manually added two more observations at $(-0.5, 2)$ and $(-3, -2.5)$.

Table 5.5 Maximum likelihood estimates (and SEs), the maximized log-likelihood, AIC and BIC for Taylor's full model (1) and constrained ($\beta = 0$) model (2).

	$\hat{a}$	$\hat{b}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\kappa}$	L	AIC	BIC
1	1.4434 (0.496)	1.405 (0.209)	–0.527 (0.150)	–0.273 (0.192)	1.811 (0.267)	–96.91	203.8	215.3
2	1.374 (0.605)	1.499 (0.262)	–0.356 (0.111)		1.784 (0.264)	–97.79	203.6	212.7

Table 5.6 Maximum likelihood estimates (and SEs), the maximized log-likelihood, AIC and BIC for model proposed by Jammalamadaka and SenGupta (2001).

		$\hat{\psi}$		L	AIC	BIC
a_0	0.297 (0.078)	c_0	–0.114 (0.054)	–111.2	238.4	256.7
a_1	0.570 (0.097)	c_1	0.277 (0.067)			
b_1	–0.128 (0.128)	d_1	0.593 (0.088)			

Table 5.7 Maximum likelihood estimates of parameters, the maximized log-likelihood, AIC and BIC for model proposed by Kato and Jones (2010) and two of its sub-models (*Source:* Kato and Jones 2010).

Model	$\arg(\hat{\beta}_0)$	$\lvert\hat{\beta}_1\rvert$	$\arg(\hat{\beta}_1)$	$\hat{\kappa}$	$\hat{r}$	$\hat{\nu}$	$\hat{\mu}$	L	AIC	BIC
K-J model	0.317	0.255	–0.558	1.40	0.399	–1.57	(0.772)	–94.7	201.4	215.1
von Mises	0.280	0.384	–0.498	1.79	(0)	(-)	(0)	–97.5	203.1	212.2
wrapped Cauchy	0.216	0.299	–0.470	(0)	0.609	(0)	(0)	–97.7	203.5	212.6

Table 5.8 The estimates, log-likelihood and value of D_i for selected observations.

i	x	y	$\hat{a}$	$\hat{b}$	$\hat{\alpha}$	$\hat{\beta}$	$L(\hat{\psi}_{(i)})$	D_i	Rank
102	–3.000	–2.500	1.885	0.195	–0.078	0.020	–51.384	4.770	1
99	1.868	0.887	2.770	1.045	0.133	0.204	–46.758	0.144	2
35	–0.243	0.247	2.749	1.016	0.139	0.198	–46.751	0.136	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
101	–0.500	2.000	3.063	1.251	–0.033	0.045	–46.654	0.039	30
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Firstly, we fit a model for all $n = 102$ observations and calculate the log-likelihood, $L(\hat{\psi})$. Then, we compute $L(\hat{\psi}_{(i)})$ and D_i; selected values are shown in Table 5.8.

The critical value d_c at 0.05 significance level was computed as $d_c = 0.625$. It shows that the observation 102 is an influential observation, while the observation 101 and the other observations do not significantly influence the model. The fitted regression curves when the observation 101 is omitted and 102 is omitted are shown in Figure 5.2.

5.6 Discussion

All models considered here can be expressed as a general form of the tangent link function of two trigonometric polynomial functions. In addition, the models proposed by Downs and Mardia (2002), Kato et al. (2008), and Kato and Jones (2010) are the same. However, the distributions of the angular errors are different. Using the form of the models given in Equation (5.3), it is hard to see how to generalize Downs and Mardia's model. As usual, increasing the number of parameters gives more flexibility in the fitted values, but this comes with a cost in estimation and identifiability.

In our experience with simulated data, the proposed test statistic for detecting influential observations in circular–circular regression gives a satisfactory performance for detecting an influential observation. Of course, this test statistic requires computational effort; for example, full analysis for about 100 observations takes about a minute on a 3 GHz desktop. However, this will become less of a disadvantage in time.

For further research, it could be interesting to investigate whether or not the proposed starting value approach will work also for alternative error distributions. Similarly, the

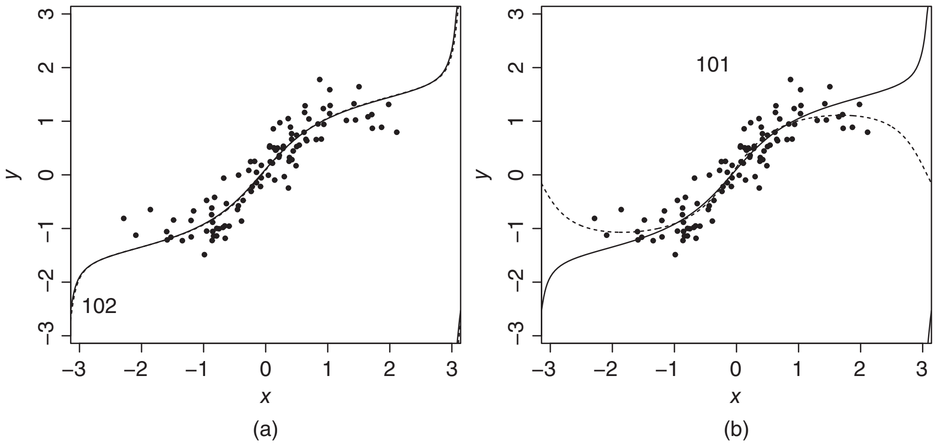


Figure 5.2 The continuous curve is the fitted regression model for all observations, $\hat{y} = 0.034 + \text{atan2}(3.003\sin(x + 0.053), 1.248\cos(x + 0.053) + 1)$ where $L(\hat{\psi}) = -46.615$. The dashed fitted regression curves; (a) when omitting the observation 101, (b) when omitting the observation 102.

extension to higher order models and multiple explanatory variables requires further investigation.

References

Barndorff-Nielsen OE and Cox DR 1989 *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London.

Cook RD 1977 Detection of influential observation in linear regression. *Technometrics* **19**, 15–18.

Cox DR 1975 Contribution to discussion of Mardia (1975a). *Journal of the Royal Statistical Society Series B (Methodological)* **37**, 380–381.

Di Marzio, M and Panzera, A and Taylor CC 2013 Nonparametric regression for circular responses. *Scandinavian Journal of Statistics* **40**, 238–255.

Downs T 1974 Rotation angular correlation In Biorhythms and Human Reproduction M. Ferin, F. Halberg and L. van der Wiele. John Wiley & Sons, Inc., New York.

Downs TD and Mardia KV 2002 Circular regression. *Biometrika* **89**, 683–698.

Jammalamadaka SR and Sarma YR 1993 Circular regression *Proceedings of the 3rd Pacific Area Statistical Conference*, pp. 109–128.

Jammalamadaka SR and SenGupta A 2001 *Topics in circular statistics*. World Scientific, Singapore.

Kato S and Jones MC 2010 A family of distributions on the circle with links to, and applications arising from, Möbius transformation. http://stats-www.open.ac.uk/TechnicalReports/KJJASA.pdf (last checked: 17.02.15).

Kato S, Shimizu K and Shieh GS 2008 A circular-circular regression model. *Statistica Sinica* **18**, 633–645.

Lockhart RA and Stephens MA 1985 Tests of fit for the von Mises distribution. *Biometrika* **72**, 647–652.

Lund U 1999 Least circular distance regression for directional data. *Journal of Applied Statistics* **26**, 723–733.

Mardia KV and Jupp PE 2000 *Directional Statistics*. John Wiley & Sons, Ltd, Chichester, UK.

Rivest LP 1997 A decentred predictor for circular-circular regression. *Biometrika* **84**, 717–726.

Taylor CC 2009 Directional data on the torus, with applications to protein structure *Proceedings of the SIS 2009 Statistical Conference on Statistical Methods for the Analysis of Large Data-Sets*, pp. 105–108.

Watson GS 1961 Goodness-of-fit tests on a circle.. *Biometrika* **48**, 109–114.

6

On two-sample tests for circular data based on spacing-frequencies

Riccardo Gatto[1] **and S. Rao Jammalamadaka**[2]

[1]*Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland*

[2]*Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA*

6.1 Introduction

In many scientific disciplines, observations are directions and are referred to as "directional data". A two-dimensional direction can be represented by (i) a vector in $\mathbb{R}^2$ of length one since magnitude has no relevance, (ii) by a complex number of unit modulus, (iii) by a point of S^1, the circumference of the unit circle centered at the origin, or (iv) by an angle measured in radians or degrees. In this chapter, we adopt this last representation using radians. Data representing two-dimensional directions is referred as "circular data." Circular data arise in many natural sciences, including geology, seismology, meteorology, animal behavior, and so on just to name a few. Moreover, any periodic phenomenon with a known period can be represented in terms of two-dimensional directions, such as the circadian rhythms.

The analysis of circular data relies on specific statistical procedures, which differ from usual statistical methodology for the real line. Since there is no prescribed null direction or sense of rotation (either clockwise or anticlockwise), it is important that procedures for circular data remain independent of the arbitrary choices of the zero direction and of the sense of rotation. The von Mises distribution provides one of the basic models for circular

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

data. It is often considered as central as the normal distribution is, for linear data. However, since there is no systematic mathematical rationale for invoking the von Mises distribution as much as there is for using a normal distribution on the line, distribution-free or nonparametric techniques assume a more important role in the context of circular data. This chapter focuses on nonparametric tests for circular data and in particular on nonparametric two-sample tests based on the so-called "spacing-frequencies". In this chapter, the importance of this type of tests is stressed in terms of invariance properties. Moreover, tests based on "circular ranks" on the circle can be reexpressed in terms of these spacing-frequencies.

Two seminal publications on circular distributions are Langevin (1905) and Lévy (1939) and one pioneering statistical analysis of directional data is due to Fisher (1953). Two general references are Mardia and Jupp (2000) and Jammalamadaka and SenGupta (2001). There is considerable literature on modeling and analysis of circular data including, for example, Rao (1969) and Gatto and Jammalamadaka (2007).

The remaining part of this chapter is organized as follows. Section 6.2 presents an overview of spacing-frequencies tests for circular data. In particular, it presents some careful analysis of the invariance, the maximality, and the symmetry properties. It then reviews three well-known two-sample tests for circular data, which are the Dixon, the Wheeler–Watson, and the Wald–Wolfowitz tests. A slight generalization based on high-order spacing-frequencies, called multispacing-frequencies, is then reviewed. The end of Section 6.2 mentions a conditional representation for the distribution of the multispacing-frequencies, which allows one to derive the asymptotic normality and a saddlepoint approximation. Section 6.3 provides an extension of Rao's one-sample spacings test (see Rao 1969, 1976) to the two-sample setting using the spacing-frequencies. A geometrical interpretation of the proposed test statistic is provided. Its exact distribution and a saddlepoint approximation are then discussed. Section 6.4 provides a Monte Carlo comparison of the powers of Wheeler–Watson's, Dixon's and Rao's two-sample spacing-frequencies tests. In this study, it is demonstrated that if one of the two samples is suspected of coming from a certain bimodal distribution, Rao's and Dixons's spacing-frequencies tests have comparable power, whereas Wheeler–Watson test, which is commonly used in this context, has substantially lower power. It may be remarked that this deficiency is comparable to that suffered by Rayleigh's test for uniformity in a single sample, when the data is suspected of not being unimodal.

6.2 Spacing-frequencies tests for circular data

Suppose we have two independent samples of circular data, the first sample consisting of m independent and identically distributed (iid) circular random variables $X_1, \ldots, X_m$, with probability distribution P_X and a second sample of n iid circular random variables $Y_1, \ldots, Y_n$, with probability distribution P_Y. As mentioned, these samples represent angles in radians, with respect to some arbitrary origin and sense of rotation. P_X and P_Y are circular distributions in the sense that they assign total measure one to $[c, c + 2\pi)$, $\forall c \in \mathbb{R}$. The general two-sample problem is to test the null hypothesis that both these samples come from the same parent population, viz.

$$\mathrm{H}_0 \colon P_X = P_Y. \tag{6.1}$$

Stating H_0 in terms of the probability distributions or measures P_X and P_Y, instead of the usual formulation in terms of cumulative distribution functions (cdf), is more appropriate because the cdf depends on the choice of the null direction and the sense of rotation. For convenience, we denote $X = (X_1, \dots, X_m)$ and $Y = (Y_1, \dots, Y_n)$.

6.2.1 Invariance, maximality and symmetries

Let $X_{(1)} \leq \cdots \leq X_{(m)}$ denote the circularly ordered values $X_1, \dots, X_m$, for a given origin and sense of rotation. With $\mathrm{I}\{A\}$ denoting the indicator of statement A, the random counts

$$S_j = \sum_{i=1}^{n} \mathrm{I}\{Y_i \in [X_{(j)}, X_{(j+1)})\}, \text{ for } j = 1, \dots, m-1, \text{ and } S_m = n - \sum_{j=1}^{m-1} S_j,$$

are commonly called (circular) spacing-frequencies, as they provide the number of observations $Y_1, \dots, Y_n$ which lie in-between successive gaps made by $X_{(1)}, \dots, X_{(m)}$. A substantial amount of nonparametric theory for the real line is based on the "ranks," for example, refer to Sidak et al. (1999). If one were to define "ranks" on the circle with respect to the same origin and sense of rotation (on which they depend), then the spacing-frequencies $S_1, \dots, S_m$ could be related to such ranks. Specifically, if R_k denotes the circular rank of the k^{th} largest $X_1, \dots, X_m$ in the combined sample, with origin given by $X_{(1)}$ and same sense of rotation as before, then

$$R_k = k + \sum_{j=1}^{k-1} S_j, \quad \text{for } k = 1, \dots, m, \tag{6.2}$$

(where $\sum_{j=1}^{0} \overset{\text{def}}{=} 0$). Conversely,

$$R_{k+1} = R_k + S_k + 1, \text{ for } k = 1, \dots, m-1, \text{ and } R_m = m + n - S_m$$

yield

$$S_k = R_{k+1} - R_k - 1, \text{ for } k = 1, \dots, m-1, \text{ and } S_m = m + n - R_m, \tag{6.3}$$

so that, $S_1, \dots, S_m$ may be thought of "rank-differences" when such ranks are well defined, as they are on the line. Moreover, note that in this context, the spacing-frequencies are well defined even in the presence of ties, that is, repeated values in the combined sample. Indeed, there is no reason to assume absolute continuity (with respect to the Lebesgue measure) of either P_X or P_Y, whereas ranks have to be adapted whenever ties have positive probability of occurring, for example, by defining "midranks."

A natural question that arises in this context is the symmetry with respect to roles of the two samples X and Y in the construction of the spacing-frequencies tests. Precisely, let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ denote the circularly ordered values $Y_1, \dots, Y_n$, for the same origin and sense of rotation used with $S_1, \dots, S_m$. The random counts

$$S'_j = \sum_{i=1}^{m} \mathrm{I}\{X_i \in [Y_{(j)}, Y_{(j+1)})\}, \text{ for } j = 1, \dots, n-1, \text{ and } S'_n = m - \sum_{j=1}^{n-1} S'_j,$$

are called the "dual spacing-frequencies." The next proposition addresses this question of sample symmetry.

Proposition 1
The dual spacing-frequencies $S'_1, \ldots, S'_n$ can be obtained as a one-to-one function of the original spacing-frequencies $S_1, \ldots, S_m$ and conversely, so that tests may be based on either set of spacing-frequencies.

We show this result in case where P_X and P_Y are absolutely continuous.

Proof
Assume P_X and P_Y absolutely continuous. Let R'_k denote the circular rank of the k^{th} largest $Y_1, \ldots, Y_n$ in the combined sample, for $k = 1, \ldots, n$, with origin given by $X_{(1)}$, which is the origin used for the original ranks, and same rotation sense as for the original ranks. Then, we can compute the dual circular ranks as follows:

$$R'_1 = 1 + \sum_{k=1}^{m} \mathrm{I}\{X_k \in [X_{(1)}, Y_{(1)})\} \text{ and } R'_k = k + R'_1 - 1 + \sum_{j=1}^{k-1} S'_j, \text{ for } k = 2, \ldots, n. \tag{6.4}$$

Given absolute continuity, we have

$$\{R'_1, \ldots, R'_n\} = \{1, \ldots, m+n\} \backslash \{R_1, \ldots, R_m\},$$

where the elements of the aforementioned sets are ordered from the smallest to the largest, when going from left to right. We then obtain

$$S'_k = R'_{k+1} - R'_k - 1, \text{ for } k = 1, \ldots, n-1, \text{ and } S'_n = m + n - R'_n + R'_1.$$

Conversely, absolute continuity yields

$$\{R_1, \ldots, R_m\} = \{1, \ldots, m+n\} \backslash \{R'_1, \ldots, R'_n\}$$

and $S_1, \ldots, S_m$ can be obtained through (6.3). □

We can thus arbitrarily decide which sample is used for constructing the spacings and which sample is used for obtaining the frequencies. Constructing tests based on either set of spacing-frequencies would make sense.

It turns out that the spacing-frequencies play a central role in comparing two circular distributions. This is because in many applied problems with circular data, the null direction and the sense of rotation are arbitrarily chosen. Assume that all circular random variables take values on $[0, 2\pi)$ and denote by $\mathcal{G}$ the transformation group consisting of all changes of origin (zero direction) and of the two changes of sense of rotation $[0, 2\pi)^{m+n} \to [0, 2\pi)^{m+n}$, that is, for the two samples. We recall that a (two-sample test) statistic $T : [0, 2\pi)^{m+n} \to \mathbb{R}$ is called invariant with respect to the transformation group $\mathcal{G}$ if, for any (X, Y) and $(\tilde{X}, \tilde{Y})$ $[0, 2\pi)^{(m+n)}$,

$$\exists\, g \in \mathcal{G} \text{ such that } (\tilde{X}, \tilde{Y}) = g(X, Y) \Longrightarrow T(\tilde{X}, \tilde{Y}) = T(X, Y).$$

If, in addition to this, for any (X, Y) and $(\tilde{X}, \tilde{Y})$ $[0, 2\pi)^{(m+n)}$,

$$T(\tilde{X}, \tilde{Y}) = T(X, Y) \Longrightarrow \exists\, g \in \mathcal{G} \text{ such that } (\tilde{X}, \tilde{Y}) = g(X, Y),$$

then the statistic T is a "maximal invariant". It can then be shown that the statistic T is $\mathcal{G}$-invariant iff T is a function of maximal $\mathcal{G}$-invariant. This leads us to ask whether

$(S_1 \dots, S_m)$ is invariant or maximal invariant with respect to the transformation group $\mathcal{G}$ and for the testing problem (6.1).

Consider first the equivalence classes generated by any maximal invariant for $\mathcal{G}$, cf. Schach (1969).

Proposition 2
The circular $[0, 2\pi)^{m+n}$-valued samples (X, Y) and $(\tilde{X}, \tilde{Y})$ belong to the same equivalence class generated by $\mathcal{G}$, iff

$$(S_1, \dots, S_m) = (\tilde{S}_{1+k}, \dots, \tilde{S}_{m+k}), \text{ for some } k \in \{0, \dots, m-1\},$$

with $\tilde{S}_j = \tilde{S}_{j-m}$, whenever $j > m$, or

$$(S_1, \dots, S_m) = (\tilde{S}_m, \dots, \tilde{S}_1),$$

where $(S_1, \dots, S_m)$ are the spacing-frequencies of (X, Y) and $(\tilde{S}_1, \dots, \tilde{S}_m)$ are the spacing-frequencies of $(\tilde{X}, \tilde{Y})$.

We often use the terminology that (X, Y) and $(\tilde{X}, \tilde{Y})$ are equal modulo $\mathcal{G}$.

Proof
The transformation group of all changes of origin is made of the set of functions $\mathcal{F}_1$ $[0, 2\pi)^{m+n} \to [0, 2\pi)^{m+n}$, which transform the spacing-frequencies of (X, Y) as

$$(S_1, \dots, S_m) \mapsto (S_2, \dots, S_m, S_1).$$

The transformation group of sense reversions is made of the set of functions $\mathcal{F}_2$ $[0, 2\pi)^{m+n} \to [0, 2\pi)^{m+n}$ yielding

$$(S_1, \dots, S_m) \mapsto (S_m, \dots, S_1),$$

when clockwise changes to anticlockwise. The transformation group $\mathcal{G}$ is made of $\mathcal{F}_1 \cup \mathcal{F}_2$. So we clearly obtain the equivalence classes mentioned in Proposition 2. □

Theoretically, one can obtain the desired $\mathcal{G}$-invariance by taking, for example, the supremum or the average of any function of $S_1, \dots, S_m$ over the given equivalence classes, but this approach seems clumsy and should not lead to any practical or useful statistic. Therefore, as a viable alternative, we consider functions of "ordered" $S_1, \dots, S_m$, which serve almost the same purpose and lead to $\mathcal{G}$-invariance. Obviously, the vector $(S_1, \dots, S_m)$ is not by itself $\mathcal{G}$-invariant: if we change for example the zero direction, then the new vector of spacing-frequencies is a permutation of the original one. So let $S_{(1)} \leq \cdots \leq S_{(m)}$ denote the ordered spacing-frequencies $S_1 \dots, S_m$. They constitute an invariant statistic for $\mathcal{G}$ and so is any statistic based on these ordered values. The complete description is given by the next proposition.

Proposition 3

1. T is a symmetric function of $S_1, \dots, S_m \iff T$ is a function of $(S_{(1)}, \dots, S_{(m)})$.
2. T is a symmetric function of $S_1, \dots, S_m \Longrightarrow T$ is $\mathcal{G}$-invariant.

Proof
1. ($\Rightarrow$) T is a function of any permutation of $S_1, \ldots, S_m$ and in particular of $(S_{(1)}, \ldots, S_{(n)})$. ($\Leftarrow$) T is invariant under permutations of $S_1, \ldots, S_m$, that is, T is a symmetric function of these values.
2. By part 1, T is a function of $(S_{(1)}, \ldots, S_{(m)})$. With Proposition 2, it is directly seen that any $\mathcal{G}$-transformation is without effect on these ordered values. □

We should remark that maximal invariance is, however, not obtained by $(S_{(1)}, \ldots, S_{(m)})$.

Proposition 4
The vector of ordered spacing-frequencies $(S_{(1)}, \ldots, S_{(m)})$ is not a maximal invariant statistic under the group $\mathcal{G}$.

Proof
Denote by $\tilde{S}_1, \ldots, \tilde{S}_m$ the spacing-frequencies obtained by the new samples $\tilde{X}$ and $\tilde{Y}$. Denote also $\tilde{S}_{(1)} \leq \cdots \leq \tilde{S}_{(m)}$ the corresponding ordered spacing-frequencies. "Maximality" means that

$$\tilde{S}_{(k)} = S_{(k)}, \text{ for } k = 1, \ldots, m \implies (\tilde{X}, \tilde{Y}) = g(X, Y), \text{ for some } g \in \mathcal{G}.$$

However, $S_{(k)} = \tilde{S}_{(k)}$, for $k = 1, \ldots, m$, means exactly that $(\tilde{S}_1, \ldots, \tilde{S}_m)$ is obtained through a permutation of the elements of $(S_1, \ldots, S_m)$. This last situation can be obtained in many different ways: for example, with $\tilde{X} = X$ and with $\tilde{Y}$ obtained from different individual transforms of the elements of Y, in such a way that $(\tilde{S}_1, \ldots, \tilde{S}_m)$ becomes the desired permutation. It is then not necessary that $\tilde{X}$ and $\tilde{Y}$ derive from a change of origin or sense of rotation, applied to X and Y simultaneously. Thus, we do not have maximality. □

As a concrete counter-example, the $\mathcal{G}$-invariant Wheeler–Watson statistic can be reexpressed as a function of $(S_1, \ldots, S_m)$, but not as a function of $(S_{(1)}, \ldots, S_{(m)})$: it is not a symmetric function of the spacing-frequencies. See Example 6 for details.

We may note the following observations about the unordered spacing-frequencies. First, $S_1, \ldots, S_m$ are exchangeable random variables under H_0 (i.e., any permutation of these random variables is equiprobable and follows the Bose–Einstein distribution in statistical mechanics). Second, consider any class of circular models parameterized by the null direction and by the sense of rotation. Then, $(S_1, \ldots, S_m)$ is an ancillary statistic for this class of models under H_0, that is, its distribution is invariant within this class.

6.2.2 An invariant class of spacing-frequencies tests

From the previous results, because the popular nonparametric Wilcoxon test statistic takes the nonsymmetric form $\sum_{k=1}^{m} kS_k$, it should not be used with circular data. Define $\mathbb{N} = \{0, 1, \ldots\}$. Assume $h : \mathbb{N} \to \mathbb{R}$ and $h_j : \mathbb{N} \to \mathbb{R}$, for $j = 1, \ldots, m$, satisfy certain mild regularity conditions. Holst and Rao (1980) consider nonparametric test statistics of the form

$$T_{m,n} = \sum_{j=1}^{m} h(S_j) \text{ and } T_{m,n}^{*} = \sum_{j=1}^{m} h_j(S_j), \tag{6.5}$$

which are called the symmetric and the nonsymmetric test statistics based on spacing-frequencies. As mentioned in Proposition 2, only the symmetric statistic $T_{m,n}$ is relevant with circular data, when considering $\mathcal{G}$-invariance. However, the asymptotic efficiencies of the nonsymmetric tests $T^*_{m,n}$ are shown to be superior by Holst and Rao (1980), when considering data on the real line.

The limiting null distribution of the most general nonsymmetric statistic $T^*_{m_\nu,n_\nu}$, when $\{m_\nu\}_{\nu\geq 0}$ and $\{n_\nu\}_{\nu\geq 0}$ are nondecreasing sequences in $\mathbb{N}^\infty$ such that, as $\nu \to \infty$,

$$m_\nu \to \infty,\, n_\nu \to \infty \quad \text{and} \quad \rho_\nu \overset{\text{def}}{=} \frac{m_\nu}{n_\nu} \to \rho, \quad \text{for} \quad \text{some} \quad \rho \in (0,\infty), \tag{6.6}$$

is given by

$$\frac{\sum_{j=1}^{m_\nu} h_j(S_j) - \mu_{m_\nu}}{\sigma_{m_\nu}} \xrightarrow{\text{d}} \mathcal{N}(0,1),$$

where μ_m and σ^2_m are defined as follows. If $V_1,\ldots,V_m$ are i.i.d. geometric random variables with

$$\mathsf{P}[V_1 = k] = \left(\frac{1}{1+\rho}\right)^k \frac{\rho}{1+\rho}, \text{ for } \ k = 0,1,\ldots, \tag{6.7}$$

then $\mu_m = \mathsf{E}[\sum_{j=1}^m h_j(V_j)]$ and $\sigma^2_m = \text{var}(\sum_{j=1}^m h_j(V_j) - \beta_m \sum_{i=j}^m V_j)$, in which $\beta_m = \text{cov}(\sum_{j=1}^m h_j(V_j), \sum_{j=1}^m V_j)/\text{var}(\sum_{j=1}^m V_j)$; refer to Corollary 3.1 on p. 41 of Holst and Rao (1980).

One can see that the circular Wald and Wolfowitz (1940) run test (see Example 6) and the circular Dixon (1940) test (see Example 5) have the symmetric form $T_{m,n}$, whereas the Wheeler and Watson (1964) test (see Example 7) is nonsymmetric with respect to the spacing-frequencies. One can also note that any linear function of the ranks $R_1,\ldots,R_m$ in the combined sample can be expressed in terms of the nonsymmetric statistic $T_{m,n}$. Further discussion on this type of tests can be found in Rao and Mardia (1980).

We now give two examples of symmetric statistics of the form given in (6.5). A third example will be suggested later in Section 6.3.1. Then we present Wheeler–Watson test, which will be analyzed numerically in Section 6.4.

Example 5 *Dixon's test* Theorem 4.2 at p. 48 of Holst and Rao (1980) states that the locally most powerful test among all symmetric tests in the spacing-frequencies given in (6.5) is

$$T_{m,n} = \sum_{j=1}^m S_j^2. \tag{6.8}$$

Note that this local optimality is under a sequence of alternative cdf's for Y_1 that converge to the cdf of X_1, both depending on the choices of zero direction and the sense of rotation, see Equation (4.2) in Holst and Rao (1980).

Example 6 *Wald–Wolfowitz run test* Another example in the class of symmetric two-sample test statistics is given by the circular version of Wald–Wolfowitz run test statistic; see also David and Barton (1962). The Wald–Wolfowitz run test statistic is $T_{m,n}$

as given by (6.5) with $h(x) = \mathrm{I}\{x > 0\}$. We define a "$Y$-run" in the combined sample as the largest nonempty group of adjacent Y-values. Since any positive value of $S_1, \dots, S_m$ constitute a Y-run, $T_{m,n}$ gives the number of Y-runs in the combined sample, and it takes values in $\{1, \dots, m\}$. But in the circle, the number of X- and Y-runs must be same and so $2T_{m,n}$ gives the total number of runs made by the combined sample. Large values of $T_{m,n}$ show evidence for equal spread, that is, for H_0. Note that Section 2.3 of Gatto (2000) provides a saddlepoint approximation to the distribution of this statistic under H_0, in the linear setting.

Example 7 *Wheeler–Watson test* This test has also been called the Mardia–Watson–Wheeler test, see e.g. p. 101 of Batschelet (1981), and the uniform scores test. It assumes absolute continuity of P_X and P_Y (in order to almost surely exclude ties). The idea is the following. Adjust the values of X and Y by respecting their relative order, in such a way to obtain $m + n$ equidistant values. So the spacings between any two consecutive adjusted values are all equal and equal to $2\pi/(m + n)$. For a given choice of origin and rotation sense, X and Y are thus mapped onto $\{2\pi k/n\}_{k=1,\dots,m+n}$. The values of X become $2\pi R_1/(m + n), \dots, 2\pi R_m/(m + n)$, which are called "uniform scores," where $R_1, \dots, R_m$ are, as before, the ranks of X in the combined sample. Because of being uniformly spread, the overall resultant vector V of the uniform scores is null, that is, $V = 0$. However, since $V = V_X + V_Y$, where V_X and V_Y are the resultant vectors of the transformed samples X and Y, it follows that $V_X = -V_Y$. (So only one of the statistics V_X and V_Y is relevant.) Under H_0, the two samples should be evenly spread over the circumference and thus $||V_X|| \simeq ||V_Y|| \simeq 0$. So a relevant decision rule is given by: reject H_0 if $||V_X||$ is large. But V_X can be obtained from the spacing-frequencies through (6.4),

$$\begin{aligned} ||V_X||^2 &= \left\{\sum_{k=1}^{m} \cos\left(\frac{2\pi}{m+n} R_k\right)\right\}^2 + \left\{\sum_{k=1}^{m} \sin\left(\frac{2\pi}{m+n} R_k\right)\right\}^2 \\ &= \left\{\sum_{k=1}^{m} \cos\left(\frac{2\pi}{m+n}\left[k + \sum_{j=1}^{k-1} S_j\right]\right)\right\}^2 + \left\{\sum_{k=1}^{m} \sin\left(\frac{2\pi}{m+n}\left[k + \sum_{j=1}^{k-1} S_j\right]\right)\right\}^2, \end{aligned} \tag{6.9}$$

which cannot have the symmetric form $T_{m,n}$ given in (6.5). From Proposition 3, it is not a function of $(S_{(1)}, \dots, S_{(m)})$. However, $||V_X||$ is clearly $\mathcal{G}$-invariant. This illustrates the non-maximality of $(S_{(1)}, \dots, S_{(m)})$ claimed by Proposition 4.

Note, however, the following drawback inherent to this test in the presence of bimodal distributions. Assume that the sample Y presents two similar modes, the second mode being located approximately at the antimode. For various configurations of the sample X, these modes lead to the cancelation in the uniform scores so that $||V_Y||$ and $||V_X||$ tend to be small, even without H_0 being true. Low power is thus expected in these cases. Our extensive simulations in Section 6.4 provide a numerical confirmation. This weakness, as we remarked in the introduction, is similar to that suffered by Rayleigh's test for uniformity when used in bimodal or multimodal samples.

6.2.3 Multispacing-frequencies tests

It turns out that the asymptotic power of the tests based on spacing-frequencies (6.5) can be improved by considering larger spacings or gaps in the following sense. Let $l \geq 1$ denote the order of the gap between the values of X and define the nonoverlapping or disjoint multispacing-frequencies as

$$S_j^{(l)} = \sum_{i=1}^{n} \mathrm{I}\{Y_i \in [X_{(jl)}, X_{((j+1)l)})\}, \text{ for } j = 1, \ldots, r, \quad \text{with } r \overset{\text{def}}{=} \left\lfloor \frac{m}{l} - 1 \right\rfloor.$$

So if $l = 1$, then $r = m - 1$ and $S_j^{(1)} = S_j$, for $j = 1, \ldots, m - 1$. In this case, S_m can be defined as before.

Assume $h : \mathbb{N} \to \mathbb{R}$ and $h_j : \mathbb{N} \to \mathbb{R}$, for $j = 1, \ldots, r$, satisfy certain regularity conditions (given under Assumption A in Jammalamadaka and Schweitzer 1985) and define the general classes of test statistics

$$T_{m,n}^{(l)} = \sum_{j=1}^{r} h(S_j^{(l)}) \quad \text{and} \quad T_{m,n}^{(l)*} = \sum_{j=1}^{r} h_j(S_j^{(l)}), \tag{6.10}$$

which represent, respectively, the symmetric and the nonsymmetric test statistics based on multispacing-frequencies. When $l = 1$, both sums in (6.10) go up to $m = r + 1$ (instead of r). Jammalamadaka and Schweitzer (1985) establish the asymptotic normality of these statistics (and of similar statistics based on overlapping multispacing-frequencies), under the null hypothesis and under asymptotically close alternatives as well. The locally most powerful test, for a given smooth sequence of alternative c.d.f. of Y_1 converging toward the cdf of X_1, is provided by Theorem 3.2 at pp. 41–42 of Jammalamadaka and Schweitzer (1985). We reject H_0 if

$$\sum_{j=1}^{r} g\left(\frac{j}{r+1}\right) S_j^{(l)} > c,$$

for some $c \in \mathbb{R}$, where the real-valued function g depends on the sequence of alternative cdf of Y_1 and on the cdf of X_1. So the optimal test statistic has the nonsymmetric form $T_{m,n}^{*}$ given in (6.10). For the same reason that nonsymmetric statistics in spacing-frequencies are not $\mathcal{G}$-invariant and symmetric statistics are $\mathcal{G}$-invariant, the nonsymmetric statistic in the multispacing-frequencies $T_{m,n}^{(l)*}$ is not $\mathcal{G}$-invariant, whereas the symmetric statistic $T_{m,n}^{(l)}$ is $\mathcal{G}$-invariant. Jammalamadaka and Schweitzer (1985) show that the sum of squared multispacing-frequencies, leading to the statistic

$$T_{m,n}^{(l)} = \sum_{j=1}^{r} \left(S_j^{(r)}\right)^2, \tag{6.11}$$

is the optimal choice among all symmetric and nonoverlapping statistics. When $l = 1$, this is the Dixon (1940) statistic of Example 5. We may note that the multispacing-frequencies statistics (6.10) are clearly nonsymmetric with respect to the roles given to the samples X and Y: if the spacings would be defined by Y and the frequencies by X, then we would obtain a different test statistic.

6.2.4 Conditional representation and computation of the null distribution

For the most general statistics based on the multispacing-frequencies, consider the independent random variables $W_1, \ldots, W_r$ with the negative binomial distribution with parameters

l and $p = \rho/(1+\rho)$, namely

$$\mathsf{P}[W_1 = k] = \begin{pmatrix} l+k-1 \\ k \end{pmatrix} \left(\frac{\rho}{1+\rho}\right)^l \left(\frac{1}{1+\rho}\right)^k, \quad \text{for } k = 0, 1, \ldots \tag{6.12}$$

The next proposition tells that under H_0, the r multispacing-frequencies have the same distribution as these negative binomial random variables, when conditioned to sum up to n.

Proposition 8
If $W_1, \ldots, W_r$ are independent random variables with probability function (6.12), then $\forall \rho \in (0, \infty)$,

$$(S_1^{(l)}, \ldots, S_r^{(l)}) \sim (W_1, \ldots, W_r) \mid Z_r = n, \tag{6.13}$$

where $Z_r = \sum_{j=1}^r W_j$.

This conditional representation is the central argument for the determination of the null asymptotic distribution of symmetric statistics, based on (nonoverlapping) multispacing-frequencies. The next proposition is a direct consequence of Theorem 4.2 on pp. 613–614 of Jammalamadaka and Schweitzer (1985).

Proposition 9
The following asymptotic distribution holds under H_0 and under the asymptotics (6.6),

$$r^{-\frac{1}{2}} \sum_{j=1}^{r} \{h(S_j^{(l)}) - \mathsf{E}[h(W_1)]\} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \zeta_l^2), \tag{6.14}$$

where

$$\zeta_l^2 = \mathrm{var}(h(W_1)) - \frac{\rho^2}{1+\rho} \cdot \mathrm{cov}^2(h(W_1), W_1). \tag{6.15}$$

We also note that the distributions of the most general test statistic $T_{m,n}^{(l)*}$ can be obtained with saddlepoint approximation suggested by Gatto and Jammalamadaka (2006), which also exploits the conditional representation (6.13); see also Section 6.3.3.

6.3 Rao's spacing-frequencies test for circular data

In this section, we provide an extension of the idea of Rao's one-sample spacings test (cf. Rao 1976) to the two-sample setting, making use of the spacing-frequencies. Although the Wheeler–Watson test is a popular two-sample nonparametric test, Rao's spacing-frequencies test has a simple intuitive interpretation and has efficiencies comparable to that of the locally most powerful Dixon's test. It also admits a nice geometrical interpretation, which is provided in Section 6.3.1. However, as mentioned in Example 6, Wheeler–Watson test has the drawback of not distinguishing the case where the P_X is bimodal at its antimode from H_0, a situation that often occurs when measuring wind directions, see for example, Section 3 of Gatto and Jammalamadaka (2007). The Wheeler–Watson test may have low power in this circumstance. A small sample power comparison in this situation and with these three tests, namely, the Wheeler–Watson test, the Dixon test, and the Rao spacing-frequencies test, is presented in Section 6.4.

6.3.1 Rao's test statistic and a geometric interpretation

Motivated by Rao's one-sample spacings test which takes the form $\sum_{j=1}^{m} |D_j - 1/m|$, where $D_1, \ldots, D_m$ denote the (one-sample) spacings (i.e., the gaps between successive points or the first-order differences) and which is widely used for testing isotropy of a single sample, we will define what we will call "Rao's two-sample spacing-frequencies test," by

$$T_{m,n} = \frac{1}{2} \sum_{j=1}^{m} \left| S_j - \frac{n}{m} \right|. \tag{6.16}$$

This is symmetric in the spacing-frequencies and has been briefly mentioned in the study by Rao and Mardia (1980).

An interesting geometrical interpretation can be given for this statistic similar to that available for the Rao's spacings test. We first note that

$$\sum_{j=1}^{m} \left(S_j - \frac{n}{m} \right) = 0 \implies T_{m,n} = \sum_{j=1}^{m} \max \left\{ S_j - \frac{n}{m}, 0 \right\}. \tag{6.17}$$

Consider for the moment a circle with circumference n (i.e.$n\mathrm{S}^1/(2\pi)$) and consider the spacing-frequencies $S_1, \ldots, S_m$ as spacings of a conceptual sample $Z = (Z_1, \ldots, Z_m)$ on this circle, that is, $S_j = Z_{(j+1)} - Z_{(j)}$, for $j = 1, \ldots, m-1$, and $S_m = Z_{(1)} - Z_{(m)}$. With this interpretation, we can consider the spacing-frequencies as $\{0, \ldots, n\}$-valued random variables. On this circle, we then place m arcs of equal length n/m, starting at each one of the m values of Z. In this situation, $T_{m,n}$ as given by (6.16) becomes the total "uncovered part of the circumference" of this circle. The case $T_{m,n} = 0$ means that all spacing-frequencies are exactly equal, that is,

$$S_1 = \cdots = S_m = \frac{n}{m},$$

which is clearly the strong evidence for $\mathrm{H}_0 : P_X = P_Y$. On the other extreme, the case $T_{m,n} = n(1 - 1/m)$ means that

$$\exists j \in \{1, \ldots, n\}, \text{ such that } S_j = n \text{ and } S_k = 0, \ \forall k \neq j \in \{1, \ldots, n\},$$

which is the strong evidence against H_0, that is, for dissimilarity between P_X and P_Y.

6.3.2 Exact distribution

It is difficult to obtain an analytical expression for the exact distribution of the circular Rao's spacing-frequencies test statistic given in (6.16). One can, however, obtain a formula for its characteristic function, along the lines of Bartlett (1938); see also Mirakhmedov et al. (2014). More generally, we consider the symmetric test statistic $T_{m,n}$ given in (6.5).

Consider the negative binomial random variables given in (6.12) with $l = 1$ and $\varphi : (\mathbb{R}^m, \mathsf{B}(\mathbb{R}^m)) \to (\mathbb{R}, \mathsf{B}(\mathbb{R}))$. Let $v \in \mathbb{R}$ and $k \in \mathbb{N}$, then

$$\mathsf{E}\left[\varphi(W_1, \ldots, W_m) \mathrm{e}^{\mathrm{i}vZ_m}\right] = \sum_{k=0}^{\infty} \mathrm{e}^{\mathrm{i}vk} \mathsf{P}[Z_m = k] \mathsf{E}[\varphi(W_1, \ldots, W_m) \mid Z_m = k]$$

(where $Z_m = \sum_{j=1}^m W_j$). The right side of the aforementioned equation is a Fourier series and from Fourier inversion we obtain

$$\mathsf{E}[\varphi(W_1,\ldots,W_m) \mid Z_m = k] = \frac{1}{2\pi \mathsf{P}[Z_m = k]} \int_{-\pi}^{\pi} \mathsf{E}[\varphi(W_1,\ldots,W_m)\mathrm{e}^{\mathrm{i}v(Z_m-k)}]\mathrm{d}v. \tag{6.18}$$

The conditional representation (6.13) directly yields

$$\mathsf{E}[\varphi(S_1,\ldots,S_m)] = \mathsf{E}[\varphi(W_1,\ldots,W_m) \mid Z_m = n],$$

which together with (6.18) at $k = n$ yields

$$\mathsf{E}[\varphi(S_1,\ldots,S_m)] = \frac{1}{2\pi \mathsf{P}[Z_m = n]} \int_{-\pi}^{\pi} \mathsf{E}[\varphi(W_1,\ldots,W_m)\mathrm{e}^{\mathrm{i}v(Z_m-n)}]\mathrm{d}v.$$

Define $\nu = \mathsf{E}[W_1] = (1-p)/p = \rho^{-1}$ and $\tau^2 = \mathrm{var}(W_1) = (1-p)/p^2 = (1+\rho)/\rho$. Given the function h of the symmetric test statistic given in (6.5) and $v_1, v_2 \in \mathbb{R}$, we define

$$\psi(v_1,v_2) = \\ \mathsf{E}\left[\exp\left\{\mathrm{i}\frac{v_1}{\zeta_1}\left(h(W_1) - \mathsf{E}[h(W_1)] - \frac{\mathrm{cov}(h(W_1),W_1)}{\tau}(W_1-\nu)\right) + \mathrm{i}\frac{v_2}{\tau}(W_1-\nu)\right\}\right]$$

and

$$\hat{\psi}_m(v_1,x) = \frac{1}{\sqrt{2\pi}} \int_{-\pi\tau\sqrt{m}}^{\pi\tau\sqrt{m}} \mathrm{e}^{-\mathrm{i}v_2 x} \psi^m(v_1,v_2)\mathrm{d}v_2,$$

for $x \in \mathbb{R}$, where ζ_1 is defined by (6.15). This last result and the inversion formula for the probability $\mathsf{P}[Z_m = n]$ provide a Bartlett-type formula for the characteristic function of

$$U_{m,n} = \frac{1}{\sigma_1} \sum_{j=1}^m \left\{h(S_j) - \mathsf{E}[h(W_1)] - \frac{\mathrm{cov}(h(W_1),W_1)}{\tau}(S_j - \nu)\right\},$$

which is given by

$$\mathsf{E}\left[\mathrm{e}^{\mathrm{i}vU_{m,n}}\right] = \frac{\hat{\psi}_n(v,x)}{\hat{\psi}_m(0,x)}. \tag{6.19}$$

Getting an analytical form for this characteristic function and inverting it to the exact probability is a difficult task, although asymptotic distribution and Edgeworth expansion can be obtained along the lines of Mirakhmedov et al. (2014).

However, given that one can compute the list of all possible realizations of $(S_1,\ldots,S_m)$, for any given $n \geq 1$, one can actually compute the value of the statistic for each of these $\binom{n+m-1}{n}$ equiprobable configurations and in this way determine the exact probability distribution of Rao's spacing-frequencies statistic $T_{n,m}$ given in (6.16).

6.3.3 Saddlepoint approximation

An alternative to finding all possible realizations of the spacing-frequencies is to approximate the exact distribution of Rao's spacing-frequencies statistic by the saddlepoint approximation. The saddlepoint approximation is a large deviations technique, which provides approximations to the exact distributions with bounded relative error. It is thus a very accurate method for computing small tail probabilities. It was introduced in statistics by Daniels (1954). In this section, we provide the cumulant generating function required for computing the saddlepoint approximation of Gatto and Jammalamadaka (1999) to the distribution of Rao's spacing-frequencies test statistics, under H_0.

For this purpose, we reexpress Rao's spacing-frequencies statistic (6.16) in the general M-statistic form $\sum_{j=1}^{m} \psi_1(S_j, T_{m,n}) = 0$, where

$$\psi_1(x, t_1) = \frac{1}{2}\left|\frac{n}{m} - x\right| - \frac{t_1}{m} = \begin{cases} \frac{1}{2}(\frac{n}{m} - x) - \frac{t_1}{m}, & \text{if } x \leq \frac{n}{m}, \\ \frac{1}{2}(x - \frac{n}{m}) - \frac{t_1}{m}, & \text{if } x > \frac{n}{m}. \end{cases}$$

We also define $\psi_2(x, t_2) = x - t_2/m$. Next, we compute the following joint cumulant generating function of these scores,

$$K(v_1, v_2; t_1, t_2) = \log\Big\{\mathsf{E}[\exp\{v_1\psi_1(W_1, t_1) + v_2\psi_2(W_1, t_2)\}]\Big\},$$

where W_1 has the distribution (6.12) with $l = 1$, which is a geometric distribution. After algebraic simplifications, we find

$$K(v_1, v_2; t_1, t_2) =$$

$$\log p + \log\left(\exp\left\{\frac{1}{m}\left[v_1\left(\frac{n}{2} - t_1\right) - v_2 t_2\right]\right\} \frac{1 - \left\{(1-p)\mathrm{e}^{v_2 - \frac{v_1}{2}}\right\}^{\lfloor \frac{n}{m} \rfloor + 1}}{1 - (1-p)\mathrm{e}^{v_2 - \frac{v_1}{2}}}\right.$$

$$\left. + \exp\left\{-\frac{1}{m}\left[v_1\left(\frac{n}{2} + t_1\right) + v_2 t_2\right]\right\} \frac{\left\{(1-p)\mathrm{e}^{\frac{v_1}{2} + v_2}\right\}^{\lfloor \frac{n}{m} \rfloor + 1}}{1 - (1-p)\mathrm{e}^{\frac{v_1}{2} + v_2}}\right),$$

$\forall v_1, v_2 \in \mathbb{R}$ such that $v_1/2 + v_2 < -\log(1-p)$. The saddlepoint approximation to $\mathsf{P}[T_{m,n} \geq t_1]$ can now be obtained by a direct application of Step 1 and Step 2 provided at p. 534 of Gatto and Jammalamadaka (1999) to the function $K_n = nK$, where K is the cumulant generating function given by the aforementioned formula. We also set $t_2 = n$ and $p = \rho_\nu/(\rho_\nu + 1)$, where $\rho_\nu = m/n$, see (6.6). One may refer to Gatto (2000) for a continuity correction for the case where the statistic is discrete, as it is the case here, and also for an algorithm for computing the quantiles, that is, the critical values of the test.

Although this approximation represents only the leading term of an asymptotic series, its accuracy is very good, even for small values of m and n and for very small tail probabilities. For the "exponential score" spacing-frequencies statistic, Table 1 in Gatto (2000) shows that with sample sizes as small as $m = 4$ and $n = 12$, this approximation is good: it yields 12% as relative error when applied to the upper tail probability 1%.

6.4 Monte Carlo power comparisons

This section presents a comparison of the power of Wheeler–Watson, Dixon's and Rao's spacing-frequencies tests, under a specific deviation from the null hypothesis, which appear unfavorable for Wheeler–Watson test. Numerical evaluations are done by Monte Carlo simulation, because it does not seem possible to extend the saddlepoint approximation of Section 6.3.3 to distributions under the alternative hypothesis. The reason is that the conditional representation (6.13) is valid only under the null hypothesis.

As is done on the real line, it is possible through a probability integral transform to make the distribution of say X uniform. Thus, let us consider the null hypothesis (6.1) wherein P_X is the circular uniform distribution with density $f_X(\theta) = 1/(2\pi)$, $\forall\theta \in \mathbb{R}$, and alternatives where and P_Y is a generalized von Mises distribution (GvM) of order two, with density given by

$$f_Y(\theta \mid \mu_1, \mu_2, \kappa_1, \kappa_2) = \frac{1}{2\pi G_0(\delta, \kappa_1, \kappa_2)} \exp\{\kappa_1 \cos(\theta - \mu_1) + \kappa_2 \cos 2(\theta - \mu_2)\}, \tag{6.20}$$

$\forall\theta \in \mathbb{R}$, where $\mu_1 \in [0, 2\pi)$, $\mu_2 \in [0, \pi)$, $\kappa_1, \kappa_2 \geq 0$, $\delta = (\mu_1 - \mu_2)\text{mod}\pi$ and where the normalizing constant is given by

$$G_0(\delta, \kappa_1, \kappa_2) = \frac{1}{2\pi} \int_0^{2\pi} \exp\{\kappa_1 \cos\theta + \kappa_2 \cos 2(\theta + \delta)\}\mathrm{d}\theta.$$

A circular random variable with density (6.20) is denoted $\text{GvM}(\mu_1, \mu_2, \kappa_1, \kappa_2)$. We refer to Gatto and Jammalamadaka (2007) for various interesting theoretical properties and characterizations regarding this class of distributions. We note that the well-known von Mises distribution is obtained by setting $\kappa_2 = 0$ in (6.20) and that the uniform distribution (with density f_X) is obtained by setting $\kappa_1 = \kappa_2 = 0$ in (6.20).

We consider alternative hypotheses where P_Y is the GvM distribution with $\mu_1 = \mu_2 = 0, \kappa_1 = 0.1$ and $\kappa_2 \in \{0.5, 1, \ldots, 7\}$. The graphs of some of these densities, over the interval $[-\pi, \pi)$, are given in Figure 6.1. We can see that each density is symmetric around zero and possesses two clear and quite similar modes. Figure 6.1 shows also that these GvM

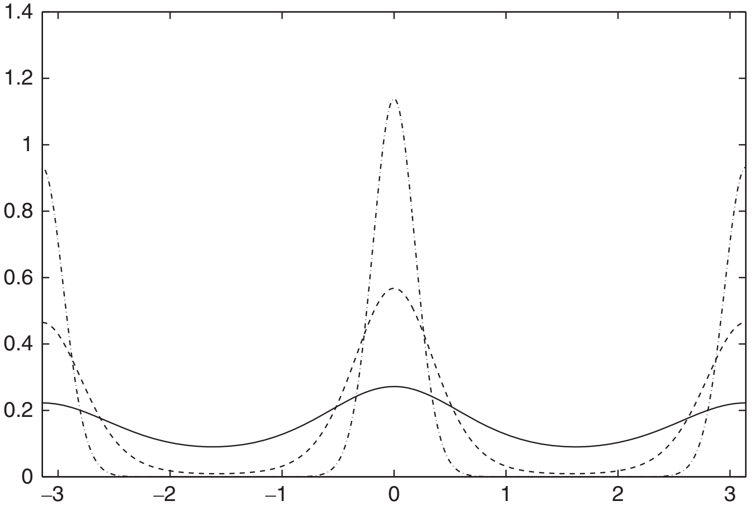


Figure 6.1 GvM densities over $[-\pi, \pi)$ with $\mu_1 = \mu_2 = 0$, $\kappa_1 = 0.1$ and $\kappa_2 = 0.5$ (solid line), $\kappa_2 = 2$ (dashed line), $\kappa_2 = 7$ (dashed-dotted line).

densities deviate increasingly from uniformity as the value of κ_2 increases. We compare the small sample power of the following tests: Wheeler–Watson test (see Example 7), Dixon's test (see Example 5) and Rao's spacing-frequencies test (see Section 6.3.1). All tests have (approximate) size 5% and the selected sample sizes are $m = 15$ and $n = 25$. Let us rewrite Wheeler–Watson test statistic $||V_X||$ given in (6.9) as $T^W_{m,n}$ and let us denote its α^{th} upper tail quantile as t^W_α. Let us also rewrite Dixon's spacing-frequencies test statistic $T_{m,n}$ given in (6.8) as $T^D_{m,n}$ and let us denote its α^{th} upper tail quantile as t^D_α. Let us also rewrite Rao's spacing-frequencies test statistic $T_{m,n}$ given in (6.16) as $T^R_{m,n}$ and let us denote its α^{th} upper tail quantile as t^R_α. Large values of $T^W_{m,n}$, $T^D_{m,n}$, and $T^R_{m,n}$ provide evidence against H_0. Based on $0.5 \cdot 10^6$ Monte Carlo simulations, we obtain $t^W_{0.05} = 5.3305$, $t^D_{0.05} = 149$, and $t^R_{0.05} = 14.6667$. In this setting, the powers of Wheeler–Watson and Rao's tests have been computed for various values of κ_2, each time based on 10^5 Monte Carlo generations.

The results are displayed in Table 6.1. We see that Wheeler–Watson test appears substantially less powerful for distinguishing the uniform distribution from the selected bimodal GvM distributions. This confirms the claim given at the end of Example 7, that the Wheeler–Watson test may not be appropriate when dealing with bimodal distributions displaying two similar well-separated modes with one at the antimode. Dixon's and Rao's spacing-frequencies test behave substantially better in this case. For other configurations with less accentuated bimodality, the power of Wheeler–Watson test is closer to the one of its competitors. Nevertheless, this important result and conclusion are in the same spirit as the well-known result that the Rayleigh test in one-sample case loses to tests such as the one-sample Rao's spacings test and is indeed inappropriate, when the data is not unimodal. We see also that Dixon's test shows slightly better power than Rao's test when κ_2 is small, that is, close to the null hypothesis and is known to be asymptotically locally most

Table 6.1 Power comparison between Wheeler–Watson, Dixon's and Rao's tests.

κ_2	$\mathsf{P}_{\kappa_2}[T^W_{m,n} > t^W_{0.05}]$	$\mathsf{P}_{\kappa_2}[T^D_{m,n} > t^D_{0.05}]$	$\mathsf{P}_{\kappa_2}[T^R_{m,n} > t^R_{0.05}]$
0.5	0.060	0.090	0.056
1.0	0.074	0.189	0.142
1.5	0.090	0.326	0.291
2.0	0.104	0.462	0.456
2.5	0.117	0.563	0.588
3.0	0.127	0.641	0.684
3.5	0.134	0.670	0.754
4.0	0.141	0.743	0.804
4.5	0.146	0.776	0.841
5.0	0.151	0.803	0.866
5.5	0.153	0.824	0.887
6.0	0.157	0.843	0.903
6.5	0.161	0.859	0.918
7.0	0.161	0.870	0.928

P_X: uniform distribution. P_Y: GvM distribution with $\mu_1 = \mu_2 = 0$, $\kappa_1 = 0.1$ and $\kappa_2 = 0.5, 1, \ldots, 7$. Each probability is obtained from 10^5 simulations. Size of tests: 5%. $m = 15$, $n = 25$.

powerful test among the symmetric tests in (6.5). However, this small advantage turns in favor of Rao's test as κ_2 increases.

Acknowledgments

The authors are grateful to an anonymous Referee for various suggestions and corrections and to the Editors for their encouraging support.

References

Bartlett MS 1938 The characteristic function of a conditional statistic. *Journal of the London Mathematical Society* **13**, 62–67.

Batschelet E 1981 *Circular Statistics in Biology*. Academic Press.

Daniels HE 1954 Saddlepoint approximations in statistics. *Annals of Mathematical Statistics* **25**, 631–650.

David FN and Barton DE 1962 *Combinatorial Chance*. Griffin and Company.

Dixon WJ 1940 A criterion for testing the hypothesis that two samples are from the same population. *Annals of Mathematical Statistics* **11**, 199–204.

Fisher RA 1953 Dispersion on a sphere. *Proceedings of the Royal Society of London Series A* **217**, 295-–305.

Gatto R 2000 Symbolic computation for approximating the distributions of some families of one and two-sample nonparametric test statistics. *Statistics and Computing* **11**, 449–455.

Gatto R, Jammalamadaka SR 1999 A conditional saddlepoint approximation for testing problems. *Journal of the American Statistical Association* **94**, 533–541.

Gatto R and Jammalamadaka SR 2006 Small sample asymptotics for higher order spacings In *Advances in Distribution Theory, Order Statistics and Inference, Part III: Order Statistics and Applications*, *Statistics for Industry and Technology* honor of BC Arnold, N Balakrishnan, E Castillo and J-M Sarabia (eds), Birkhäuser, pp. 239–252.

Gatto R and Jammalamadaka SR 2007 The generalized von Mises distribution. *Statistical Methodology* **4**, 341–353.

Holst L and Rao JS 1980 Asymptotic theory for some families of two-sample nonparametric statistics. *Sankhyā* Series A **42**, 19–52.

Jammalamadaka SR and Schweitzer RL 1985 On tests for the two-sample problem based on higher order spacing-frequencies In *Statistical Theory and Data Analysis* Matusita K (ed.), North-Holland, pp. 583–618.

Jammalamadaka SR and SenGupta A 2001 *Topics in Circular Statistics*. World Scientific, Singapore.

Langevin P 1905 Magnétisme et théorie des éléctrons. *Annales de Chimie et de Physique* **5**, 71–127.

Lévy P 1939 L'addition des variables aléatoires définies sur une circonférence. *Bulletin de la Société Mathématique de France* **67**, 1–41.

Mardia KV and Jupp PE 2000. *Directional Statistics*. John Wiley & Sons, Ltd, Chichester.

Mirakhmedov SM, Jammalamadaka SR and Ibrahim BM 2014 On Edgeworth expansions in generalized urn models. *Journal of Theoretical Probability* **27**, 725–753.

Rao JS 1969 *Some Contributions to the Analysis of Circular Data* Ph.D. Thesis, Indian Statistical Institute, Calcutta.

Rao JS 1976 Some tests based on arc-lengths for the circle. *Sankhyā* Series B **4**, 329–338.

Rao JS and Mardia KV 1980 Pitman efficiencies of some two-sample nonparametric tests In *Recent Developments in Statistical Inference* Matusita K (ed.), North-Holland, pp. 247–254.

Schach S 1969 On a class of nonparametric two-sample tests for circular distributions. *Annals of Mathematical Statistics* **40**, 1791–1800.

Sidak Z, Sen PK and Hajek J 1999 *Theory of Rank Tests*, 2nd ed., Academic Press.

Wald A and Wolfowitz J 1940 On a test whether two samples are from the same population. *Annals of Mathematical Statistics* **11**, 147–162.

Wheeler S and Watson GS 1964 A distribution-free two-sample test on a circle. *Biometrika* **51**, 256–257.

7

Barycentres and hurricane trajectories

Wilfrid S. Kendall
Department of Statistics, University of Warwick, Coventry, UK

7.1 Introduction

This paper is principally motivated by intellectual curiosity. After work by Huiling Le (Kendall and Le 2011) on laws of large numbers and central limit theorems for empirical Riemannian barycentres, it seemed natural to investigate the use of Riemannian barycentres in data analysis. This topic relates to contemporary interest in the statistical analysis of data comprised of intrinsically geometric objects, which can be viewed as part of the subject sometimes known as ‘object-oriented data analysis’. Indeed, the statistical use of barycentres has already been pioneered, for example, in Fournel et al. (2013) and Ginestet et al. (2012) (also see early work by Ziezold 1994); the purpose of this paper is to use a specific application to explore their use in analysing trajectories with strong geometric content. In the following, we use Riemannian barycentre theory to produce a simple non-parametric analysis of the extent to which consecutive North Atlantic hurricanes might have similar behaviour. Note that considerably more sophisticated methods of curve-fitting on manifolds could have been used here (see for example the use of smoothing splines described in Su et al. 2012): the barycentre approach is relatively simplistic, but nonetheless may be useful.

Writing this paper affords the opportunity of expressing sincere homage to Kanti Mardia for his seminal leadership in the application of geometry to statistics. I owe him thanks for kindness and encouragement stretching right back to 1978, when as a callow research

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

student I was invited by Kanti to make a most stimulating research visit to the University of Leeds Statistics group. Moreover, the present work originated in preparation for a talk I gave at one of the famous LASR workshops initiated and nurtured by Kanti at Leeds (specifically LASR 2011). I hope that Kanti finds pleasure in reading this brief account.

The paper commences (Section 7.2) with a speedy review of relevant aspects of Riemannian barycentres, with special attention paid to the simple but fundamental case of sphere-valued data. Section 7.3 then reviews `HURDAT2` a remarkable publicly available data set composed of hurricane trajectories (tropical cyclones in the North Atlantic) and some associated data concerning wind speeds and atmospheric pressures. This is the test data set: attention will be confined to the hurricanes viewed as trajectories on the terrestrial sphere. We then describe (Section 7.4) how barycentre theory interacts with non-parametric statistics *via* k-means clustering and discuss preliminary results for the test data set (Section 7.5). The concluding Section 7.6 reviews the results and considers some possible next steps.

7.2 Barycentres

Fréchet (1948) introduced barycentres in metric spaces as minimizers of 'energy functionals' $x \mapsto \mathbb{E}\left[\operatorname{dist}^2(X,x)\right]$ for random variables X taking values in metric spaces. Kendall (2013) presents a recent review of some subsequent theory; there is a strong link with convexity *via* 'convex geometry' (Kendall 1990, see also Afsari 2011). Our interest is focussed on the theory of Riemannian barycentres for random variables taking values in the 2-sphere S^2: indeed this can be viewed as normative for Riemannian barycentres (Kendall 1991). In particular, sphere-valued random variables can be guaranteed to possess unique barycentres when their distributions are concentrated in closed subsets of open hemispheres (which is to say, when the random variables are confined to 'small hemispheres'). Considerable work has been devoted to establishing laws of large numbers and central limit theorems for empirical barycentres (Bhattacharya and Patrangenaru 2003; 2005; Bhattacharya and Bhattacharya 2008); this has even opened up a new multivariate perspective on the classical Feller-Lindeberg central limit theory Kendall and Le (2011). An interesting non-Riemannian case is discussed in the pioneering work of Hotz et al. (2013); see also Barden et al. (2013). In this chapter, our interest centres on more data-analytic concerns, based on barycentres of measurable random maps $\Phi : [0,T] \to S^2$ from a time-interval to S^2. Convex geometry, hence uniqueness of barycentres, is maintained if for each time t the random variable $\Phi(t)$ is supported in a (possibly time-varying) small hemisphere.

There are a number of studies of iterative algorithms for computing Riemannian barycentres (e.g., Le 2004; Arnaudon et al. 2012). We shall finesse such considerations by approximating Riemannian barycentres on the sphere by *cosine-barycentres*, given by projecting the conventional expectation onto the sphere,

$$\operatorname*{argmin}_{x \in S^2} \mathbb{E}\left[1 - \cos\operatorname{dist}(X,x)\right] = \frac{\mathbb{E}[X]}{\|\,\mathbb{E}[X]\,\|}\,.$$

This is the 'mean direction' in the terminology of directional statistics. We choose to use the term 'cosine-barycentre', to emphasize that it minimizes what one might term the *cosine-energy* (related to chordal distance) $1 - \cos\operatorname{dist}(x,y) \approx \frac{1}{2}\operatorname{dist}(x,y)^2$. In particular, it provides a feasible and explicit approximation to the Riemannian barycentre if the dispersion of X on S^2 is not too large. Note, however, that its explicit and constructive definition provides no easy panacea for questions of uniqueness: evidently the construction only works when $\mathbb{E}[X] \neq 0$, and indeed if the support of the data cannot be contained in a closed subset of an open hemisphere, then it is possible for the barycentre to be ill-defined. (Consider, for example, the problem of finding the cosine-barycentre for a probability distribution spread out uniformly over a fixed great circle. Then all points on the sphere minimize the cosine-energy.)

Cosine-barycentres allow us to choose representative barycentre trajectories for a collection of hurricane trajectories, defined as solving the minimization problem

$$\underset{F:[0,T]\to S^2}{\operatorname{argmin}} \frac{1}{T}\int_0^T \mathbb{E}\left[1 - \cos \operatorname{dist}(F(t), \Phi(t))\right] \mathrm{d}t\,. \tag{7.1}$$

Here, the expectation is actually the empirical sample average over the collection of hurricane trajectories, so that Φ is viewed as drawn uniformly at random from this collection. Note that the minimization in (7.1) can be carried out separately for each time t, since there is no continuity requirement placed on F. We choose to avoid considerations of continuity or of smoothness of trajectories; close inspection of the hurricane trajectories in Figure 7.1 suggest that smoothness, at least, is perhaps not a paramount consideration. However, there are some practical issues that need to be faced. Our hurricane trajectories are not *a priori* registered to comparable starting and/or finishing times; in fact typically their durations do not overlap. We restrict attention to trajectories that make upcrossings on latitudes of 20°N and 35°N: we register times to agree at the first upcrossing of latitude 35°N. (The following analysis is sensitive to these choices: lower latitudes do not produce a clear statistical signal.)

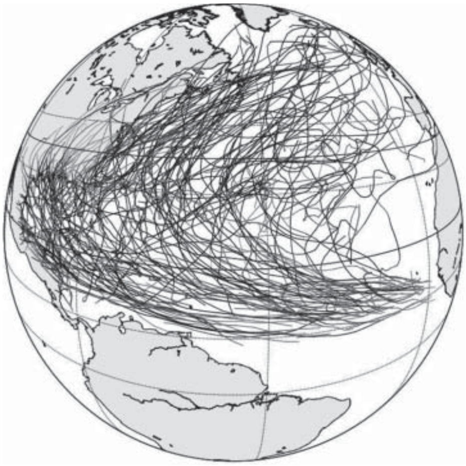


Figure 7.1 Trails of the 233 hurricanes recorded over the period 2000–2012 in the `HURDAT2` data set (darker trajectories indicate high maximum sustained wind speed). The viewpoint of this and similar following images is placed 5000 km above the centre of the image, which is therefore distorted at normal viewing distance for all but the most extremely short-sighted. We will consider the longer period 1950–2012 and will restrict attention to hurricanes crossing 20°N and 35°N (drawn as continuous lines in figure), and register hurricanes on their first crossing of 35°N.

A further issue is the need to average over hurricanes that, even when time-registered, start at different negative times and finish at different positive times. A possible solution is suggested by the observation that all hurricanes under consideration are confined to the Northern hemisphere. This suggests the following idea: if $\Phi_\omega(t)$ is not defined for some sample point $\omega \in \Omega$, then replace $1 - \cos \operatorname{dist}(F(t), \Phi_\omega(t))$ in (7.1) by the average when $\Phi(t)$ is uniformly distributed over the equator (this decision could be justified by arguments of maximum entropy type). However, this modification does not lead to good results in our current application. Instead, we solve the issue by restricting attention to the largest time interval over which our collection of time-registered hurricanes all has defined locations. This cropping procedure has the disadvantage of restricting the analysis to behaviour of the trajectories near the specified upcrossing used for registration.

Finally, we mention two more possible refinements, both of which would be computationally demanding and which we will not adopt. Firstly, one could attempt to solve the extended minimization problem allowing time-shifts of individual trajectories within the minimization problem (7.1). Secondly, one could replace the time integral in (7.1) by an integral over arc-length, or perhaps an integral over upcrossings of latitudes (though in this case one would be deliberately introducing discontinuity in cases when hurricane trajectories decrease as well as increase in latitude). Indeed one could envisage a whole variety of possible nonlinear warpings of time. We defer to another occasion the consideration of these refinements, as well as of assessment of the effect of approximating Riemannian barycentres by cosine-barycentres.

Our application needs to find a number of representative barycentre trajectories instead of just one. The natural remedy is to apply the k-means algorithm, adapted to use cosine-barycentres. Specifically, we aim to use barycentre k-means to cluster the chosen set of hurricane trajectories, so that we can study temporal association between cluster labels defined by the barycentre trajectory to which each hurricane trajectory is attached. We use Lloyd's algorithm (Lloyd 1982) for k-means: beginning with a random initial set of k trajectories serving as cluster centroid trajectories, the algorithm alternates between associating each trajectory to the closest cluster centroid trajectory (measured by cosine-distance), and then replacing each cluster centroid trajectory by the computed barycentre trajectory for the cluster. The algorithm has to be run repeatedly in order to find a good clustering; we choose to use 10 repetitions. Typically a single run of the algorithm will not produce an optimal set of cluster centroid trajectories (here, minimizing within-cluster sum of cosine-distances); indeed the task of producing such an optimal set is typically NP-complete.

Faster algorithms *do* exist for the one-dimensional problem (Wang and Song 2011), but it is an open problem to extend these to 'nearly one-dimensional' structure as exhibited by the set of hurricane trajectories. In this study, we use the k-means algorithm, setting $k = 20$, to group hurricanes into 20 groups linked to 20 barycentre trajectories. The groups are ordered from west to east according to where the barycentre trajectories first cross latitude 35°N.

7.3 Hurricanes

As an illustrative application, we consider the remarkable and freely available HURDAT2 data set concerning hurricanes (tropical cyclones) of the North Atlantic ocean, a collection of 1740 hurricane trajectories in the North Atlantic recorded by various means over

161 years from 1851 to 2012 (see Figure 7.1 for a display of recent hurricanes; the data set is discussed in Landsea and Franklin 2013; McAdie et al. 2009). The data set is available at

www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html.

Our interest centres on whether there is any evidence for temporal association; is there a tendency for successive hurricane trajectories to be close? MacManus (2011) used geometric methods to investigate similar issues (area of overlapping curvilinear strips based on paths and non-parametric measures of association). Here, we intend to use this question to illustrate application of the notion of barycentres based on hurricane paths considered as S^2-valued trajectories. Evidently the extent of these paths, ranging over wide expanses of the North Atlantic, means that their underlying geometric nature should be taken seriously.

We emphasize that this investigation is of purely methodological interest, aimed at illustrating the use of barycentres in data analysis. Addressing the question of temporal association properly would require serious attempts to relate HURDAT2 to other data sets and sources of information.

The following remarks are taken from McAdie et al. (2009) (further useful background is also given in a striking statistical survey of data acquired from three years of USAF flight missions flown into Northwest Pacific tropical cyclones which is reported by Weatherford and Gray 1988a; 1988b). Each hurricane trajectory in the HURDAT2 data set is a 'Best Track', defined using best estimates of the location of the hurricane centre, reconciling measurements obtained by various means and taken at six hourly intervals with small-scale smoothing applied. In essence, each hurricane trajectory is represented by a timed sequence of latitude/longitude pairs, measured every quarter of a day in degrees of latitude and longitude to an accuracy of 1 decimal place. Recall that a degree of latitude or longitude at the equator represents separation slightly in excess of 110 km, so stated location precision is of order ± 5 km. For the North Atlantic basin, McAdie et al. (2009) reports that diameters of hurricane eyes lie in the range of 15–50 km, so this is an entirely adequate level of precision. Typically hurricane eyes move at speeds of order of 5 m/s, so successive 6 hourly measurements are separated by order of 100 km (Figure 7.2). Measurements of maximum sustained surface wind speed and (subsequent to 1979) central surface barometric pressure are also recorded at various distances from hurricane centres; However, we focus on geographic location alone.

Surveillance methods for data capture have, of course, varied over the period of the data set. Early observations were acquired from sailing ship logs; in due course these were supplemented by radio reports, then by aircraft and radar observations, and finally by satellite observation and other systems such as dropsondes. Whether because of increased observational capacity or whether because of secular change (long-term non-periodic variation) in global weather conditions, the number of recorded hurricanes over the period 1851–2012 is clearly increasing with time (see the lowess curve for annual counts of hurricanes crossing $20°$N and $35°$N, given in Figure 7.3): for example, later years seem more likely to experience six or more such hurricanes. Further evidence of secular change is obtained by categorizing these hurricanes using the k-means algorithm, with $k = 4$, applied to those hurricanes crossing latitudes $20°$N and $35°$N. Figure 7.4 indicates that the East–West distribution of recorded hurricanes also varies over this period; it appears that more easterly hurricanes are recorded in later years.

It is evident from this discussion that, for the purposes of study of temporal association, the HURDAT2 data set is best considered as a temporally ordered sequence of short

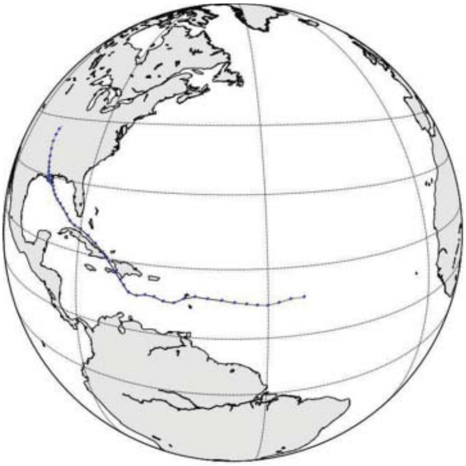


Figure 7.2 Sampling points (measured every 6 hours) of the Best Track of hurricane Isaac in 2012. Typical separation between measurement points is about 100 km.

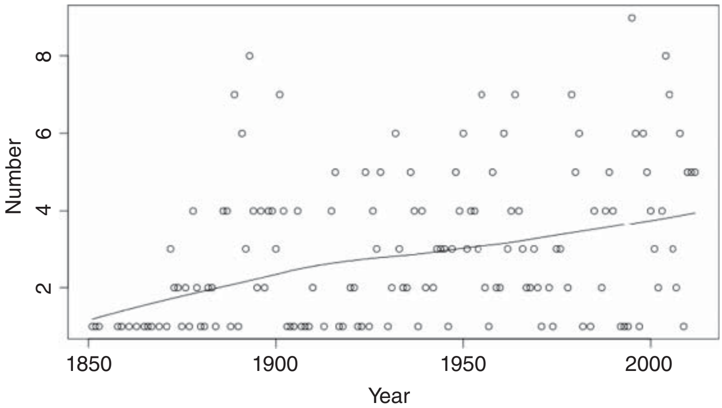


Figure 7.3 Numbers of hurricanes per year in the HURDAT2 data set over 1851–2012 which cross latitudes 20°N and 35°N, together with fitted lowess curve.

time-series, one short time-series per year over the period 1851–2012; moreover, the statistics of these short time-series should be expected to be different in later as opposed to earlier years. We shall focus subsequently on the period 1950-2012, but should bear in mind the trends illustrated in Figure 7.4.

7.4 Using k-means and non-parametric statistics

We seek to employ barycentre techniques to investigate temporal association between successive hurricanes.

Regardless of the *ad hoc* aspects of clustering using the k-means algorithm (specifically, a potential dependence of actual implementation outcomes on initial conditions), the crucial

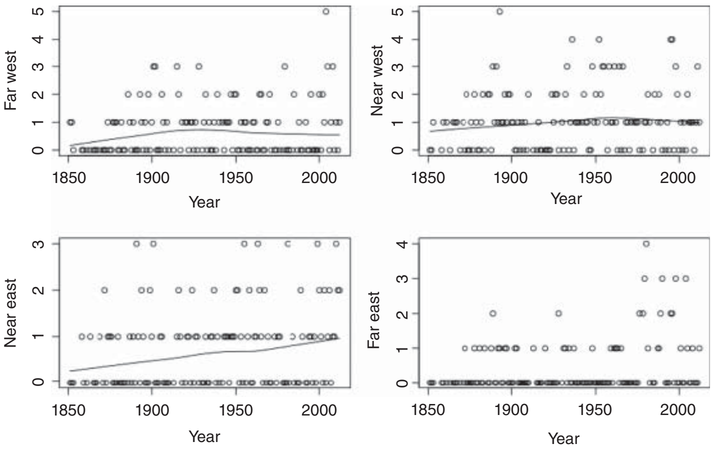


Figure 7.4 Numbers of hurricanes per year in the `HURDAT2` data set over 1851–2012 which cross latitudes 20°N and 35°N, grouped according to whether they first cross latitude 35°N at far-west, near-west, near-east, or far-east locations as indicated using k-means clustering (with $k = 4$) together with fitted `lowess` curves (except in the far-east case, for which the large majority of years record no hurricanes).

point is that the clustering takes no account of time order, whether by year or by time within year. Consequently, this clustering method can be used to detect temporal association using non-parametric statistical permutation tests.

Note that we do not consider the issue of estimation of the number of clusters k, as we are interested in the output of the k-means algorithm solely as an intermediate device to facilitate the detection of temporal association.

The simplest choice is to compute the statistic T counting the number (summed over all years) of pairs of hurricanes, one immediately succeeding the other in a given year, such that both hurricanes are categorized as belonging to the same k-mean cluster. Note that this is effectively a multiple category variant of the runs test for randomness, as discussed by Wald and Wolfowitz (1940), since each hurricane either initiates a run or completes a successive pair of hurricanes from the same cluster. (A discussion of the multiple category variant is given by Mood 1940, who also records an informative early history of the runs test.)

It is enlightening to impose a narrative based on an informal statistical model of temporal variation, as this allows a somewhat more statistically principled approach and suggests a useful generalization. Independently for each year y, let $X_{y,i}$ (for $i = 1, 2, \ldots, h_y$) record the k-mean cluster containing the ith hurricane (measured in time order) of the h_y hurricanes observed that year. Then, the year y in question contributes the following summand T_y to the total non-parametric statistic $T = \sum_y T_y$:

$$T_y(x_1, \ldots, x_{h_y}) = \sum_{i=2}^{h_y} \mathbb{I}\left[X_{y,i} = X_{y,i-1}\right].$$

This suggests that we model the sequence of k-mean labels in a given year by a one-dimensional Potts model (Grimmett 2006, S1.3): conditional on h_y the total number of hurricanes in that year, the probability of observing the sequence $x_1, x_2, \ldots, x_{h_y}$ is

$$\mathbb{P}\left[X_{y,1} = x_1, \ldots, X_{y,h_y} = x_{h_y} | h_y\right] = \frac{1}{Z(\theta, h_y)} \exp(\theta\, T_y) = \frac{1}{Z(\theta, h_y)} \exp\left(\theta \sum_{i=2}^{h_y} \mathbb{I}\left[x_i = x_{i-1}\right]\right). \tag{7.2}$$

Here, $\theta \geq 0$ is the parameter relating to strength of association, and the partition function can be computed explicitly in this simple one-dimensional case: $Z(\theta, h) = k(\mathrm{e}^\theta + k - 1)^{h-1}$. (The $k = 2$ case corresponds to a one-dimensional Ising model and can be traced as far back as Ernst Ising's 1924 thesis, published in part in Zeitschrift für Physik in 1925.)

Treating each year y as independent, we choose to condition not only on the total number h_y of hurricanes in that year but also on the total numbers $R_{y,j} = \#\{i : X_{y,i} = j\}$ of hurricanes in the year y categorized as belonging to k-mean cluster j. The action of conditioning on the $R_{y,j}$ discards some information about θ, since large positive values of θ would promote dominance by a single cluster in each year y. However, the empirical evidence of secular trends supplied by Figure 7.4 suggests the need for a more realistic model for the measurements $R_{y,j}$, allowing for their distributions not being exchangeable over the label j; conditioning on the $R_{y,j}$ allows us to evade this difficulty. The score statistic for the resulting conditioned model at $\theta = 0$ is then the year's contribution T_y to the non-parametric statistic T. Thus, consideration of T amounts to performing a conditional Neyman–Pearson hypothesis test of a null hypothesis $\mathcal{H}_0 : \theta = 0$, against a one-sided compound hypothesis $\mathcal{H}_1 : \theta > 0$.

It is of course possible to develop this theme further, using the evaluation of $Z(\theta, h_y)$ for the one-dimensional Potts model. Thus,

1. Inference could be improved to take explicit account of the information provided by the pattern of values of $R_{j,y}$, for example, by imposing an external field on the Potts model (7.2) to obtain

$$\mathbb{P}\left[X_{y,1} = x_1, \ldots, X_{y,h_y} = x_{h_y} | h_y\right] \propto \exp\left(\theta \sum_{i=2}^{h_y} \mathbb{I}\left[x_i = x_{i-1}\right] + \sum_{i=1}^{h_y} \sum_{j=1}^{k} \psi_j\, \mathbb{I}\left[x_i = j\right]\right).$$

2. Or one could attempt maximum likelihood estimation of θ, or even of different θ_j pertaining to different clusters j using a refined probability mass distribution

$$\mathbb{P}\left[X_{y,1} = x_1, \ldots, X_{y,h_y} = x_{h_y} \mid h_y, R_{y,1}, \ldots, R_{y,k}\right] \propto \exp\left(\sum_{i=2}^{h_y} \theta_{x_i}\, \mathbb{I}\left[x_i = x_{i-1}\right]\right).$$

However, we avoid pursuing either of these leads here, not only because of the resulting increase in model complexity but also because this would commit us in excessive detail to a parametric model that is motivated largely by heuristic considerations. Moreover, the model is dependent on the output of the k-means Lloyds algorithm, itself potentially a random phenomenon, insofar as the actual outcome of the algorithm can depend on essentially random selection of initial conditions (mitigated by using repeated independent runs and taking the best resulting clustering). Finally, the model, as expressed here, uses (naïve) free boundary conditions (it treats initial and final hurricanes of each year in much the same way as all the others), which is a further reason not to take it too seriously.

To evaluate the significance of the score statistic T, we compute the conditional mean and variance of each T_y, add up the T_y over the years y under consideration, and refer the sum T to a normal distribution with matching mean and variance (this relies implicitly on a central limit theorem approximation of Lyapunov type), or use a simulation test based on random permutations within each year. Means and variances can be computed by straightforward combinatorial methods: suppose that in year y there are $R_{y,j} = r_j$ hurricanes present belonging to cluster j, for $j = 1, \ldots, k$. For convenience, we set $m_2 = \sum_{j=1}^k r_j(r_j - 1)$ and $m_3 = \sum_{j=1}^k r_j(r_j - 1)(r_j - 2)$, and find

$$\mathbb{E}\left[T_y | R_{y,1} = r_1, \ldots, R_{y,k} = r_k\right] = \frac{m_2}{h_y}, \tag{7.3}$$

$$\mathrm{Var}\left[T_y | R_{y,1} = r_1, \ldots, R_{y,k} = r_k\right] = \frac{m_2^2}{h_y^2(h_y - 1)} + \frac{(h_y - 3)m_2}{h_y(h_y - 1)} - \frac{2m_3}{h_y(h_y - 1)}. \tag{7.4}$$

Note that differentiation of the partition function $Z(\theta, h_y)$ would produce means and variances *not* conditioned on the pattern of $R_{y,j}$ values, which would not suit our purpose.

It is useful to modify the one-dimensional Potts model (7.2) to allow for a geometric decay in strength of association, with decay rate $\beta \in (0, 1)$ and $\beta \ll 1$:

$$\mathbb{P}\left[X_{y,1} = x_1, \ldots, X_{y,h_y} = x_{h_y} | h_y\right] =$$

$$\frac{1}{Z(\theta, h_y, \beta)} \exp(\theta\, T_{y,\beta}) = \frac{1}{Z(\theta, h_y, \beta)} \exp\left(\theta \sum_{\ell=1}^{h_y-1} \beta^{\ell-1} \sum_{i=\ell+1}^{h_y} \mathbb{I}\left[x_i = x_{i-\ell}\right]\right). \tag{7.5}$$

The corresponding conditional score statistic (given the pattern of values $R_{y,j}$) is

$$T_{y,\beta} = \sum_{\ell=1}^{h_y-1} \beta^{\ell-1} \sum_{i=1+\ell}^{h_y} \mathbb{I}\left[X_{y,i} = X_{y,i-\ell}\right] = \sum_{\ell=1}^{h_y-1} \beta^{\ell-1} \# \left\{\text{agreements at lag } \ell\right\}, \tag{7.6}$$

and this provides some capacity to allow for detection of temporal association between trajectories, which are not immediately consecutive in time. The computations of mean and variance at (7.3) and (7.4) can be generalized to cover this case, though the formulae are too unwieldy to report here. In any case, in the sequel we shall evaluate test statistic scores by using a simulation test based on random permutations within each year.

7.5 Results

We consider a specific example, namely 179 North Atlantic hurricanes occurring over the time-range 1950–2012, and crossing latitudes 20°N and 35°N, and registered at first upcrossing of 35°N. It is candidly admitted that other crossing latitude choices lead to results which are not statistically significant, so assessment of the phenomena observed here needs to take judicious account of associated implicit selection effects. We have excluded years in which no more than two hurricanes occur, as they supply no information about clustering within a year. A k-means algorithm ($k = 20$, using Lloyd's algorithm based on 10 repetitions) classified the remaining 179 hurricanes from 37 years. The 20 groups provided by the k-means algorithm are loosely ordered on an East–West axis as illustrated in Figure 7.5 (see also Figure 7.4, which uses a k-means analysis with $k = 4$, based on the entire 1851–2012 data set). In the figure, the apparently anomalous barycentre trajectory running nearly horizontally near the more eastern end of the collection arises from a single rather long hurricane trajectory, whose initial behaviour (including its upcrossing of 20°N) is removed as part of the process of confining the barycentre trajectories to the strict intersection of registered time intervals for the component hurricane trajectories. Consequently the grouping by k-means relates to trajectory behaviour on quite a narrow band of latitudes, as can be seen from Figure 7.5. Labelling the groups in this order, so that the most westerly group at first crossing of 35°N is given index 0, we obtain Table 7.1. Reading from west to east, numbers in the 20 groups are given in Table 7.2. This procedure yields a test statistic (sum of numbers within each

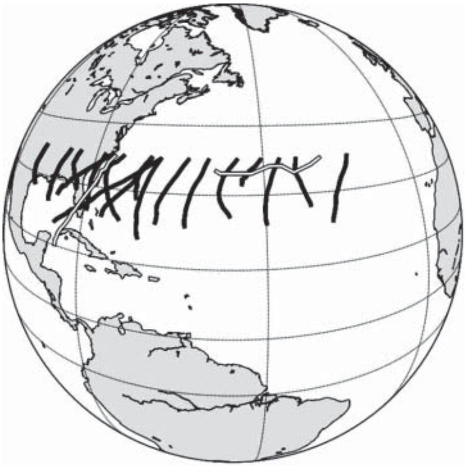


Figure 7.5 Plot of 20 barycentre trajectories arising from the k-means algorithm with $k = 20$, applied to the 1950–2012 data set of hurricanes crossing 20°N and 35°N. Barycentre trajectories are denoted by thick paths (the two outline paths correspond to clusters of just one or two hurricanes). The apparently anomalous trajectory running nearly horizontally near the more eastern end of the collection arises from a single rather long trajectory, whose initial behaviour (including its upcrossing of 20°N) is cropped as part of the process of cropping all hurricane trajectories to the same maximal time-interval.

Table 7.1 179 hurricanes over 37 years, classified by year and by 20 groups using the k-means algorithm with $k = 20$. Groups are ordered according to how westerly is the upcrossing by the corresponding barycentre trajectory of latitude (35°)N.

Year	Labels								
1950	5	1	13	8	2	13			
1951	12	14	7						
1952	6	3	8	9					
1953	11	12	11	4					
1954	11	5	5						
1955	7	5	13	0	14	5	11		
1958	5	14	7	13	8				
1961	14	0	17	8	8	9			
1962	19	14	11						
1963	12	13	16	12					
1964	7	13	2	4	12	8	3		
1965	1	1	17	14					
1966	7	9	9						
1969	5	13	1						
1975	1	13	11						
1976	14	19	15						
1979	0	11	3	1	19	16	16		
1980	19	19	8	17	16				
1981	3	18	8	13	14	13			
1985	1	1	5	7					
1988	3	0	16	10					
1989	12	17	17	13	5				
1990	14	16	15	17					
1995	4	8	8	16	13	11	11	19	1
1996	3	8	5	11	19	15			
1998	7	8	17	19	19	10			
1999	14	3	13	3	13				
2000	14	4	4	14					
2001	11	13	19						
2003	1	12	5	14					
2004	4	6	17	2	1	2	16	17	
2005	1	1	1	8	15	15	10		
2006	3	12	14						
2008	12	1	0	5	0	15			
2010	13	5	0	12	13				
2011	3	8	11	13	14				
2012	0	13	19	12	9				

year of consecutive pairs belonging to the same group) of $T = 15$. Computing mean and variance of T, conditional on the numbers of each cluster occurring in each year, and assuming the normal approximation of T to be valid, this can be referred to a conditional one-sided 5% level of $10.66 + 4.15 = 14.81$.

A quantile–quantile plot, based on 1000 randomized versions of the data displayed in Table 7.1, shows that the standardized distribution of T has somewhat lighter tails compared

Table 7.2 Numbers of hurricanes in each of the 20 groups determined by the k-means algorithm with $k = 20$. Groups are ordered according to how westerly is the upcrossing by the corresponding barycentre trajectory of latitude 35°N.

8	15	4	10	6	13	2	7	14	5	3	12	11	19	15	6	8	9	1	11

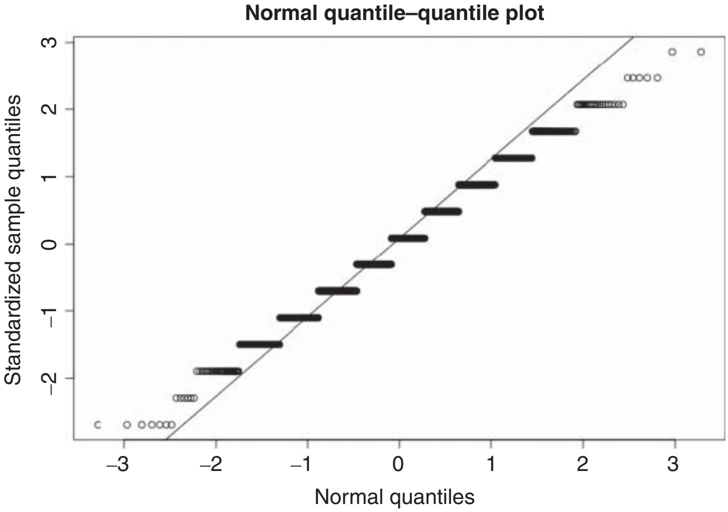


Figure 7.6 Quantile–quantile plot assessing approximate normality of the distribution of the test statistic T (with T constructed using k-means clustering with $k = 20$) based on 1950–2012 hurricanes crossing latitudes 20°N, 35°N.

to a normal distribution (see Figure 7.6). However, a simulation test based on 1000 simulations yields an unremarkable p-value of 7.5%, compared to a p-value of 4.3% using the normal approximation.

If we replace the test statistic T by T_β (obtained by summing the contributions from $T_{y,\beta}$ as defined in (7.6)), so as to make a weighted count of repetitions at longer lags, then there is a slightly stronger indication of clustering. We choose $\beta = 0.25$, so that longer lags are penalized quite heavily, (results do not appear to be particularly sensitive to the choice of small $\beta > 0$.) A quantile–quantile plot suggests normality of the distribution of $T_{\beta=0.25}$ under the hypothesis of no temporal association, though we omit this plot here, and in any case focus on assessment *via* a simulation-based permutation test. We obtain $T_{\beta=0.25} = 16.97$, and a simulation test based on 1000 simulations yields a p-value of 3.2%. (Examination of the data confirms that the dominant contribution to the modest improvement in p-value arises from pairs of similar hurricanes separated by lag 2.) This, therefore, suggests modest evidence of temporal association among this particular set of hurricanes. Given the hurricane trails have been classified into 20 clusters, over an East–West range of order 6000 km, the length scale of this association may be deemed to be of order of 300 km. This is supported by the boxplots of root-mean-square average distances of hurricanes from associated barycentres in Figure 7.7 (the calculation employs the approximation of Riemannian barycentres by cosine-barycentres). Mean distances between hurricanes and associated barycentres do indeed appear to be of order of 300 km. Visual inspection of the

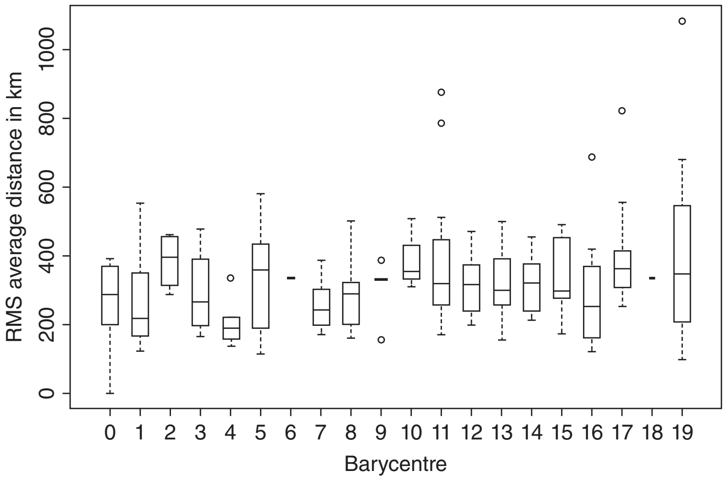


Figure 7.7 Boxplots of root-mean-square (RMS) average distances of hurricanes from associated barycentres (in kilometres). Boxplot widths are proportional to square roots of sample sizes.

individual clusters confirms that some clusters do group together rather different hurricane trajectories, underlining the need to be cautious in interpreting the formal statistical analysis given earlier. We note the two groups containing fewer than three hurricanes (groups 6 and 18); inspection shows that trajectories in both groups exhibit atypical behaviour. Various outliers and the larger dispersion of group 19 appear mostly to be linked to the more diverse behaviour of easterly hurricane trajectories.

7.6 Conclusion

This work illustrates how barycentres can be used in the analysis of trajectories with strong geometric content (here, hurricane trajectories lying on the surface of the terrestrial sphere). The conclusions drawn are modest, namely that there is rather limited evidence in favour of temporal interaction. We repeat that potential selection effects need to be borne in mind here. The nature of the data set (secular trends, structure of temporal sequence of short time-series) hamper further investigation. Were the purpose of this paper to develop such an applied theme, then the next step would be to pay greater attention to other features of the underlying data set (in particular, records of wind strength and atmospheric pressure), and also to combine the aforementioned analysis with inference drawn from other associated meteorological data sets. But the intention of this paper is more methodological: further development in such a direction could include the investigation of the more parametric inferential approaches mentioned in Section 7.4: including information derived from varying proportions of different groups of hurricanes, attempting maximum likelihood estimation of interaction parameters or even Bayesian inference exploiting the form of the likelihood of the heuristic parametric model.

Other avenues of investigation would require commitment of more computational resource: investigation of the effect of the cosine-barycentre approximation, factoring out time-shifts, or making a discontinuous time-change by referring trajectories to their times of upcrossing of successive latitudes, or working in terms of arc-length rather than time.

Finally, and more speculatively, the geometric context of these data is very simple. Useful insight might arise from consideration of more ambitious questions. For example, and following one of the applications in Su et al. (2012), this methodology could be extended to deal with the more complicated geometrical considerations that would arise when comparing three-dimensional trajectories arising in the study of chemotaxis, or more generally of motility of small organisms living at low Reynolds number (Purcell 1977). One speculates that it might be possible to use these three-dimensional trajectories to draw inferences concerning stochastic characteristics of trajectories in the rotational group, using the imputed orientation of the small organism in question; at small length scales (where Brownian effects cannot be neglected) this might lead to intriguing statistical applications of the techniques underlying the celebrated Eells–Elworthy stochastic development (Elworthy 1982, ch.VII.11).

Acknowledgment

This work was supported by EPSRC grant EP/K031066/1.

References

Afsari B 2011 Riemannian L^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society* **139**(2), 655–674.

Arnaudon M, Dombry C, Phan A and Yang L 2012 Stochastic algorithms for computing means of probability measures. *Stochastic Processes and their Applications* **122**, 1437–1455.

Barden D, Le H and Owen M 2013 Central limit theorems for Fréchet means in the space of phylogenetic trees. *Electronic Journal of Probability* **18**(25), 1–25.

Bhattacharya A and Bhattacharya RN 2008 Statistics on Riemannian manifolds: asymptotic distribution and curvature. *Proceedings of the American Mathematical Society* **136**(08), 2959–2967.

Bhattacharya RN and Patrangenaru V 2003 Large sample theory of intrinsic and extrinsic sample means on manifolds – I. *Annals of Statistics* **31**(1), 1–29.

Bhattacharya RN and Patrangenaru V 2005 Large sample theory of intrinsic and extrinsic sample means on manifolds – II. *Annals of Statistics* **33**(3), 1225–1259.

Elworthy KD 1982 *Stochastic Differential Equations on Manifolds*, *LMS Lecture Note Series*, CUP.

Fournel A, Reynaud E and Brammer M 2013 Group analysis of self-organizing maps based on functional MRI using restricted Frechet means. *Neuroimage* **1205.6158**, 1–23.

Fréchet M 1948 Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré* **10**(4), 215–310.

Ginestet C, Simmons A and Kolaczyk E 2012 Weighted Fréchet means as convex combinations in metric spaces: properties and generalized median inequalities. *Statistics and Probability Letters* **82**(10), 1–7.

Grimmett GR 2006 *The Random-cluster Model*, *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, Vol. 333, Springer-Verlag, Berlin.

Hotz T, Huckemann S, Le H, Marron JS, Mattingly JC, Miller E, Nolen J, Owen M, Patrangenaru V and Skwerer S 2013 Sticky central limit theorems on open books. *Annals of Applied Probability* **23**(6), 2238–2258.

Kendall WS 1990 Probability, convexity, and harmonic maps with small image I: uniqueness and fine existence. *Proceedings of the London Mathematical Society (Third Series)* **61**, 371–406.

Kendall WS 1991 Convexity and the hemisphere. *Journal of the London Mathematical Society (Second Series)* **43**, 567–576.

Kendall WS 2013 A survey of Riemannian centres of mass for data *Proceedings of the 59th ISI World Statistics Congress*, Hong Kong, pp. 1786–1791.

Kendall WS and Le H 2011 Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. *Brazilian Journal of Probability and Statistics* **25**(3), 323–352.

Landsea CW and Franklin JL 2013 Atlantic Hurricane Database uncertainty and presentation of a new database format. *Monthly Weather Review* **141**(10), 3576–3592.

Le H 2004 Estimation of Riemannian barycentres. *LMS Journal of Computation and Mathematics* **7**, 193–200.

Lloyd SP 1982 Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**(2), 129–137.

MacManus L 2011 Smith Institute: modelling hurricane track memory.

McAdie CJ, Landsea CW, Neumann CJ, David JE, Blake ES and Hammer GR 2009 *Tropical Cyclones of the North Atlantic Ocean, 1851–2006*, Historical Climatogoloy Series (6th edn), National Oceanic and Atmospheric Administration, Asheville, NC.

Mood AM 1940 The distribution theory of runs. *Annals of Mathematical Statistics* **37**, 688–697.

Purcell E 1977 Life at low Reynolds number. *American Journal of Physics* **45**(1), 3–11.

Su J, Dryden IL, Klassen E, Le H and Srivastava A 2012 Fitting smoothing splines to time-indexed, noisy points on nonlinear manifolds. *Image and Vision Computing* **30**(6-7), 428–442.

Wald A and Wolfowitz J 1940 On a test whether two samples are from the same population. *Annals of Mathematical Statistics* **11**, 147–162.

Wang H and Song M 2011 `Ckmeans.1d.dp`: optimal k-means clustering in one dimension by dynamic programming. *R Journal* **3**(2), 29–33.

Weatherford CL and Gray WM 1988a Typhoon structure as revealed by aircraft reconnaissance Part I: data analysis and climatology. *Monthly Weather Review* **116**, 1032–1043.

Weatherford CL and Gray WM 1988b Typhoon structure as revealed by aircraft reconnaissance Part II: structural variability. *Monthly Weather Review* **116**, 1044–1056.

Ziezold H 1994 Mean figures and mean shapes applied to biological figure and shape distributions in the plane. *Biometrical Journal* **36**(4), 491–510.

Part III

SHAPE ANALYSIS

8

Beyond Procrustes: a proposal to save morphometrics for biology

Fred L. Bookstein[1,2]

[1]*Department of Statistics, University of Washington, Seattle, WA, USA*

[2]*Department of Anthropology, University of Vienna, Vienna, Austria*

8.1 Introduction

In his long and productive career, Kanti Mardia has serially focused his attention on a very broad range of methodological innovations that together link the foundations of statistical science to its cutting-edge applications. Some of these bridge designs have excited the interest of many adopters or further innovators – his new formal models for directional data, his methods for Bayesian analysis of unlabelled point matching, the kriged Kalman filter – whereas others that would appear on formal grounds to be equally brilliant have been overlooked by the applied statistics community. This generalization is true as well of my own collaborations with him. The paper by Mardia and me introducing the rigorous Procrustes analysis of bilateral symmetry (Mardia et al. 2000), for instance, is among my greatest hits (132 citations over 15 years, according to Web of Science), whereas our paper on intrinsic random field models for deformations (Mardia et al. 2006) has generated only seven citations over its nine years of postnatal life. I think this imbalance is unrepresentative of the importance of the innovations. This chapter is an attempt at redress.

The focus of this return visit is a model that explicitly contradicts the assumption of spatially uncorrelated variability at the foundation of the Procrustes method. Figure 8.1 is a pedagogic aid showing a sample of shapes derived from a starting 5×5 square grid by

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

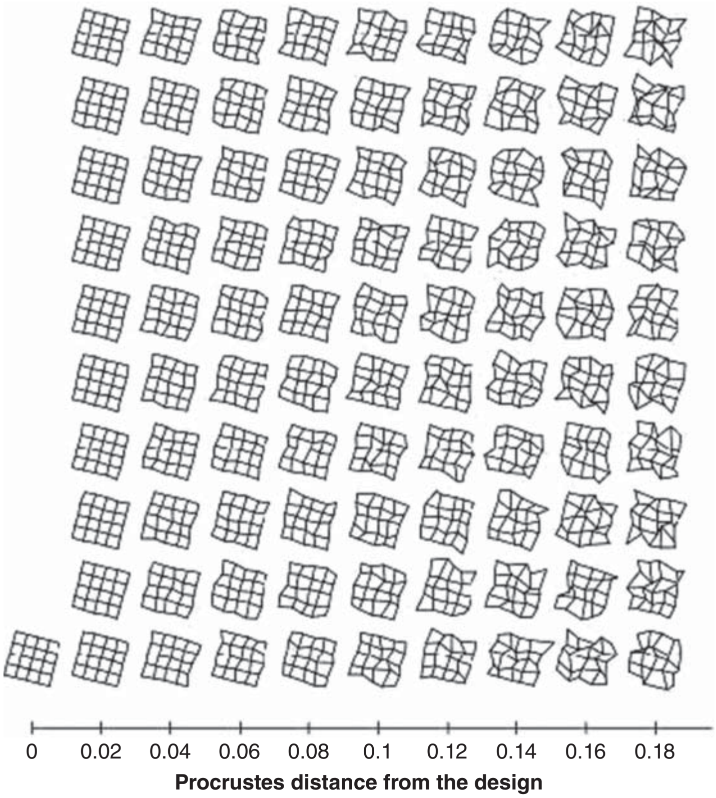


Figure 8.1 Deformations of a template (lower left) on an isotropic offset Gaussian model can be sorted into shells that are spherical in Procrustes distance.

independent and identically distributed (i.i.d.) perturbations at a range of amplitudes in the appropriate spherical shape space. By sorting into bins of Procrustes distance, we display these samples as a nest of spherical shells. While the grids of the successive shells certainly appear steadily more irregular, the specific irregularity proffered by the examples toward the right seems unconducive to most biologically useful pattern recognition approaches.

As I explained in Bookstein (2007), the reason these simulacra are unsatisfactory as hints of organismal variability is the total spatial noncorrelation of the underlying model. Figure 8.2 shows a far more appropriate generative model, whereby a 13-landmark form is parcellated into successively smaller compartments within each of which the variability is represented by one “new” landmark varying isotropically at a scale that shrinks with the size of its compartment. When the mean landmark positions involve such artificial symmetries, parcellations such as this can be extended indefinitely.

Regarding the prototype in Figure 8.2, for instance, the first four landmarks to be considered, as in the upper left panel, are the outer corners of the square, which, given their symmetry, are known to delimit a Procrustes shape space of four dimensions that takes the form of the tensor product of two planes, one for isotropic variation of the so-called uniform term (affine transformations) and the other for the *purely inhomogeneous transformations* (Bookstein 1991). In the upper central panel, the fifth landmark, at the center

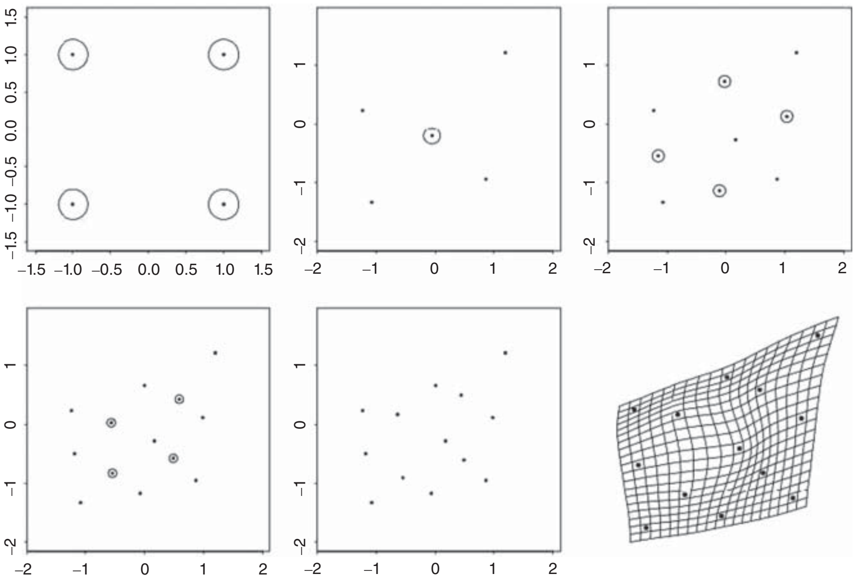


Figure 8.2 In an approach far more likely to be useful to the organismal biologist, deformations may be constructed serially from a parcellation of the form into cells involving one new landmark each with isotropic variance that is linearly scaled to the size of its compartment in some sense.

of the square, is perturbed with circular symmetry around the location imputed to it by the deformation of the square, with a variance that is half that of the corners of the square. Then (upper right) the midpoints of the edges of the square follow, independently in this simulation, each perturbed around *its* imputed location with variance reduced by a further factor of one-half, and so forth. If we stop at the 13-landmark stage, lower center, the resulting net deformation (graphed of course as the thin-plate spline at lower right) appears to have discrete features at a satisfying range of spatial scales – if this were a summary of some experimental or evolutionary phenomenon, we would be able to report it and speculate on its causes or effects.

A sampling of forms at varying amplitude for the cascade of isotropic generations here, Figure 8.3, shows this quite clearly. Now each deformation of the template seems to suggest a short list of one or two specific features *of* that deformation, a circumstance entirely contrary to that of the analogous offset isotropic shape distribution sampled in Figure 8.1.

Note. The examples in this paper are all two-dimensional, but preliminary simulations suggest that the protocol here extends to data in three dimensions in accordance with the rules set out in Section 3.1 of Mardia et al. (2006) for covariance kernel $|r|$ in place of $r^2 \log r$.

8.2 Analytic preliminaries

It proves helpful to approach this antinomy using a formalism that has proved useful in many other contexts of geometric morphometrics as well, the *bending energy of the thin-plate*

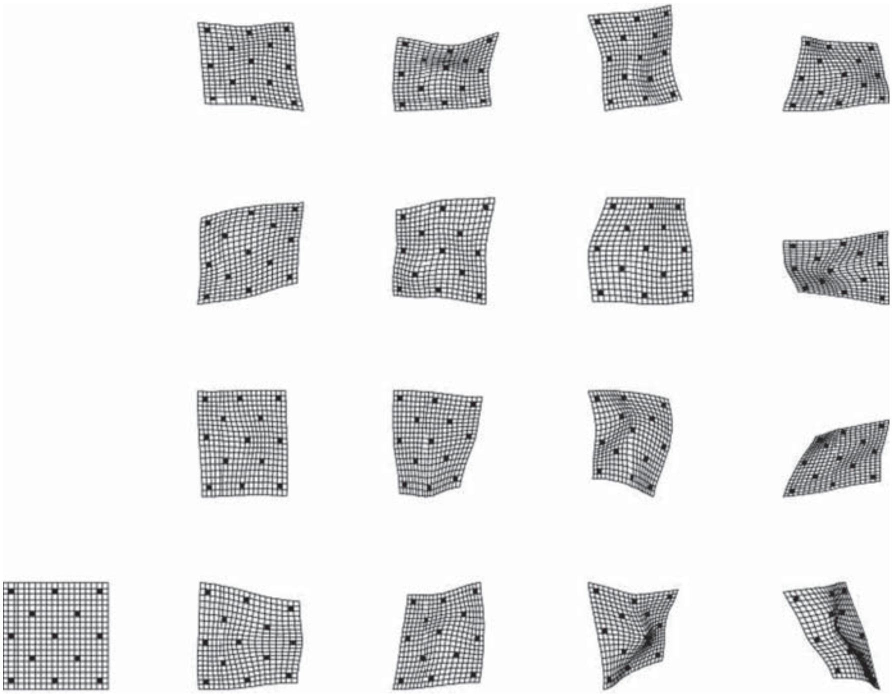


Figure 8.3 Further instances of multiscale models like that in Figure 8.2, over a range of parcellation-scaled variances that increase from left to right.

spline. We have already encountered this notion tacitly in Figure 8.2, where the mean location around which each "new" landmark was perturbed was the location assigned it by the thin-plate spline transformation on the landmarks already fixed – the target location of least bending from the template given the locations of the landmarks already assigned.

We have recourse to one standard notation. Let U be the function $U(r) = r^2 \log r$, and let $P_i = (x_i, y_i),\ i = 1, \ldots, k$, be k points in the plane. Writing $U_{ij} = U(P_i - P_j)$, build up matrices

$$K = \begin{pmatrix} 0 & U_{12} & \ldots & U_{1k} \\ U_{21} & 0 & \ldots & U_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ U_{k1} & U_{k2} & \ldots & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_k & y_k \end{pmatrix},$$

and

$$L = \begin{pmatrix} K & Q \\ Q^t & O \end{pmatrix}, \ (k+3) \times (k+3),$$

where O is a 3×3 matrix of zeros. Write $H = \left(h_1 \ldots h_k \; 0 \; 0 \; 0\right)^t$ and set $W = \left(w_1 \ldots w_k \; a_0 \; a_x \; a_y\right)^t = L^{-1}H$. Then the thin-plate spline $f(P)$ having heights (values) h_i at points

$P_i = (x_i, y_i)$, $i = 1, \ldots, k$, is the function

$$f(P) = \sum_{i=1}^{k} w_i U(P - P_i) + a_0 + a_x x + a_y y.$$

This function $f(P)$ has three crucial properties:

1. $f(P_i) = h_i$, all i: f interpolates the heights h_i at the landmarks P_i.

2. The function f has minimum **bending energy** of all functions that interpolate the heights h_i in that way: the minimum of

$$\iint_{\mathbf{R}^2} \left(\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \, \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right),$$

where the integral is taken over the entire picture plane.

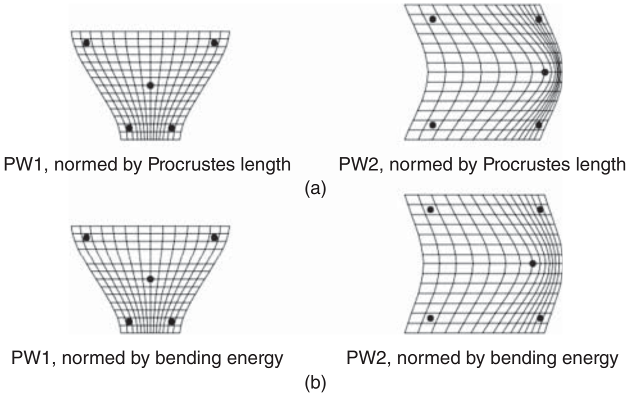
3. The value of this bending energy is

$$\frac{1}{8\pi} W^t K W = \frac{1}{8\pi} W^t \cdot H = \frac{1}{8\pi} H_k^t L_k^{-1} H_k,$$

where L_k^{-1}, the *bending energy matrix,* is the $k \times k$ upper left submatrix of L^{-1}, and H_k is the initial k-vector of H, the vector of k heights. The bending energy matrix has rank $k - 3$, corresponding to its three zero eigenvalues for the hyperplane of deformations that have no bending, the linear transformations $a_0 + a_x x + a_y y$.

In the application to two-dimensional landmark data, we compute two of these splined surfaces, one (f_x) in which the vector H of heights is loaded with the x-coordinate of the landmarks in a second form, another (f_y) for the y-coordinate. Then the first of these spline functions supplies the interpolated x-coordinate of the map we seek, and the second the interpolated y-coordinate. It is easy to show (see Bookstein 1989) that we get the same map regardless of how we place the (x, y) coordinate axes on the picture. For any such coordinate system, the resulting map $(f_x(P), f_y(P))$ is now a deformation of one picture plane onto the other that maps landmarks onto their homologues and has the minimum bending energy of any such interpolant. The bending energy of a grid is now the scalar sum of the bending energies in the x- and y-coordinates of the target configuration separately.

Bending energy scales as the inverse square of spatial scale; it will be our key to the link with organismal biology. To intuit one crucial geometric aspect of this link, the scale of features, it may be helpful to examine the fundamental diagram of dimensions of the shape space for a familiar prototype, the *quincunx* (pattern of the five-spot of a die). Corresponding to the five landmarks, there are two nonzero eigenvalues of L_k^{-1} corresponding to the two patterns in Figure 8.4. The eigenvalues of bending energy are in a ratio of 25:9. Whereas the less bent eigenmode of bending leaves the central landmark unmoved, displacing only the landmarks at the corners, the more bent has the opposite action, leaving the corners invariant but displacing only the centroid. The resulting gradients of squared second derivative, therefore, need to be quite a bit steeper, as assessed globally by that eigenvalue ratio. The intensification is also apparent visually: note, in Figure 8.4(a), the greatly increased compression of the grid lines at right center in the upper PW2 panel in comparison to those at lower center in the PW1 panel to its left. In Figure 8.4(b), this imbalance of spacing is effectively eradicated.



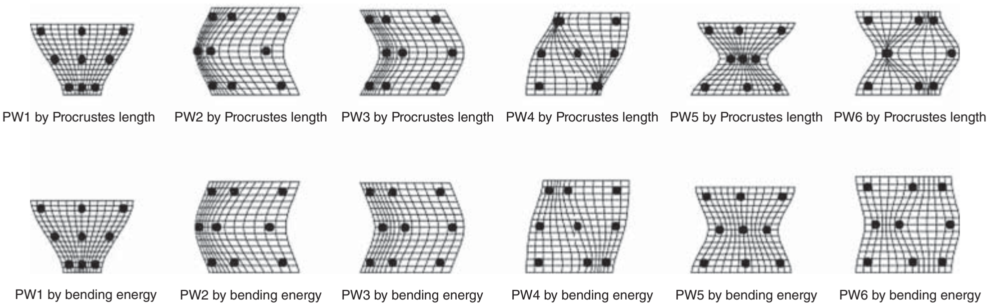
PW1, normed by Procrustes length PW2, normed by Procrustes length

(a)

PW1, normed by bending energy PW2, normed by bending energy

(b)

Figure 8.4 Nontrivial eigenvectors of a quincunx of landmarks, here drawn as parallel displacements in the horizontal direction. They can be drawn in either the Procrustes norm (a) or the bending energy norm (b).



PW1 by Procrustes length PW2 by Procrustes length PW3 by Procrustes length PW4 by Procrustes length PW5 by Procrustes length PW6 by Procrustes length

PW1 by bending energy PW2 by bending energy PW3 by bending energy PW4 by bending energy PW5 by bending energy PW6 by bending energy

Figure 8.5 The same for a 3×3 grid of landmarks. The attenuation by the square root of specific bending energy steadily increases from left to right.

We can pursue the analogous investigation for any other prototype scheme of landmark spacing. Figure 8.5, for instance, surveys the situation for a 3×3 grid of nine landmarks. The nonzero eigenvalues of the bending energy matrix are now in proportion to 6.75, 3.73 (twice), 2, 1.51, and 1.

Note that the uniform transformations (square to parallelograms), corresponding to the zero eigenvalues of L_k^{-1}, do not appear in diagrams of this style. These terms “have no scale,” or, rather, have infinite scale. They do not fit into this reformulation, but must be left outside as one additional two-dimensional aspect of any sample variation encountered.

8.3 The core maneuver

The key observation driving the claim that we can do better than the Procrustes approach for applications in organismal biology is the following observation (which follows from the analytic strategy set out in Mardia et al.

> When shapes are sampled in a covariance structure inverse to the bending energy matrix, as restricted to the subspace spanned by its eigenvectors of nonzero eigenvalue, then the distribution of component subshapes is self-similar as a function of scale.

Although it is relatively simple to demonstrate this proposition, its validity is startling. The corresponding proposition, after all, is false as it pertains to the isotropic Procrustes distribution itself: the smaller the square, the larger its own nonaffine shape variability when studied as a configuration of four landmarks only. As an accessible example, I return to the 13-landmark scheme of Figure 8.2. But this time, instead of producing those maps by an *ad hoc* parcellation, I rigorously deflate each dimension of the nonaffine shape space

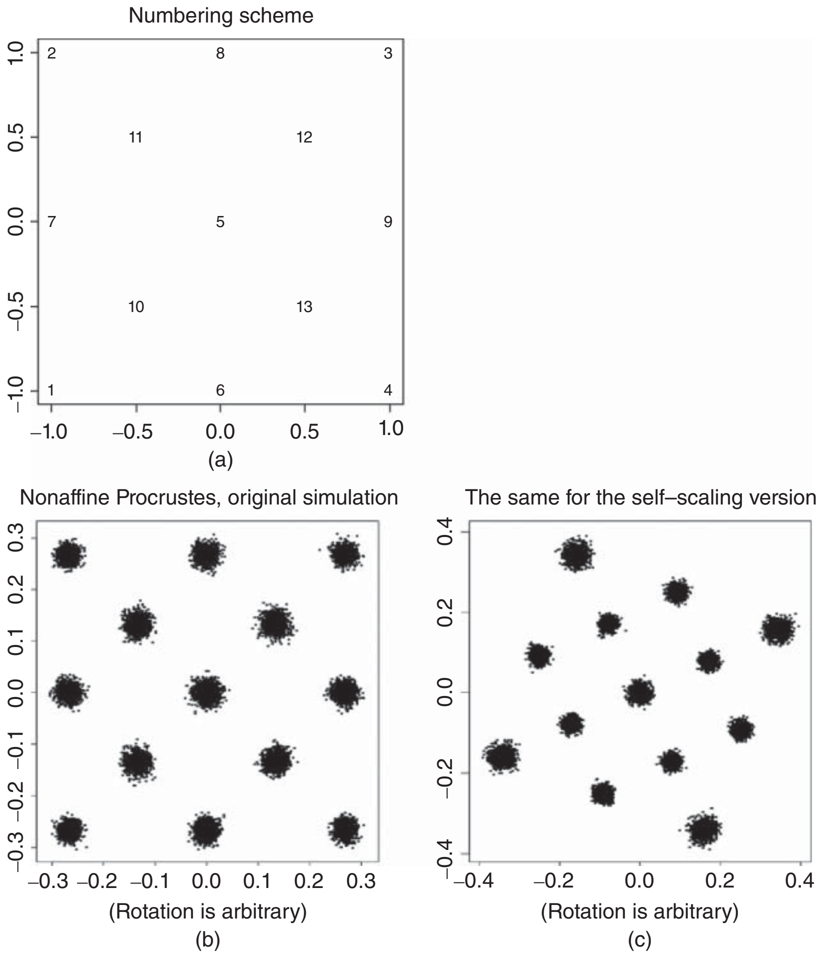


Figure 8.6 Ordinary Procrustes shape coordinate scatters for the isotropic offset Gaussian distribution on the indicated grid of 13 landmarks (b) versus the self-similar version (c). The inset (a) indicates the numbering scheme for the landmarks in the next two figures. See text.

according to the specific bending energies of its ten nontrivial dimensions. These bending energies are proportional to 15.22, 9.06, 8.41 (twice), 7.57, 3.57 (twice), 2.25, 1.26, and 1. The symmetries of this didactic configuration are irrelevant to the point being made here; they only afford the possibility of attending to a wide range of nominally square subconfigurations.

We get a hint of the new situation if we simply compare the usual Procrustes coordinate scatters for these two covariance structures. The two scenes are juxtaposed in Figure 8.6. The ordinary Procrustes shape coordinates, at left, show the familiar scaling of variance by $1 - r^2$ where r is the distance from the common centroid. The situation at the right is quite different. Here the corner points have the *most* variation, not the least, while the central landmark is actually more variable than the landmarks of the little square around it, as it contributes to larger scale features than they do.

Figures 8.7 and 8.8 combine the demonstration of imperfect scaling for the Procrustes shape distribution with evidence of perfect scaling for the new bending energy modification as applied to squares selected from the 13-landmark configuration in two orientations.

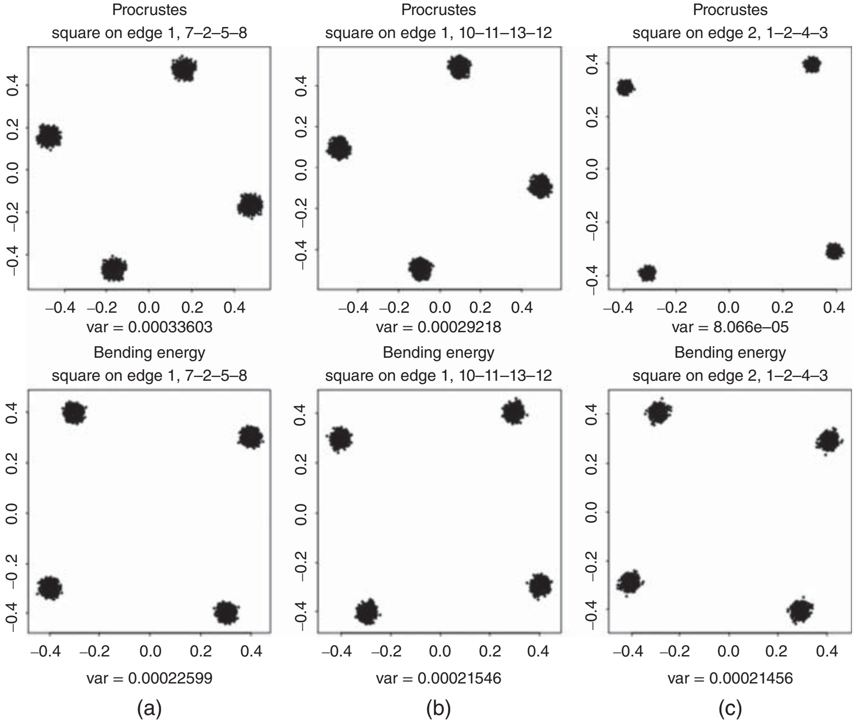


Figure 8.7 Ordinary Procrustes nonaffine shape scatters for square subconfigurations from the distributions in Figure 8.6. Upper row, to the isotropic Gaussian offset distribution; lower row, to the bending energy modification recommended in this paper. (a) For a small square in a corner of the full configuration. (b) For the square at the center. (c) For the four corners of the configuration as a whole. Under each panel is printed the variance of the x-coordinate of any one of the landmarks plotted.

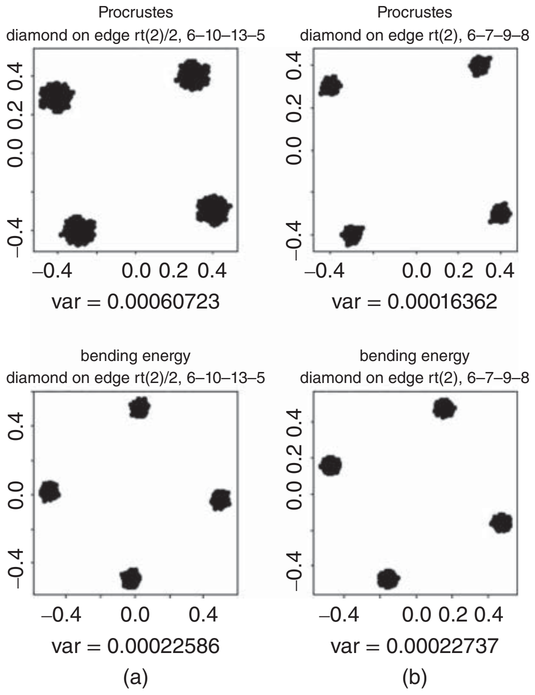


Figure 8.8 The same for squares rotated 45° to a diamond orientation, at small size (a) or large (b).

(Landmark numbering goes according to the guide shown in Figure 8.6(a).) In Figure 8.7, we show the distributions of the nonaffine component of the shape of three squares from our prototype: the square on landmarks 7, 2, 5, and 8, the upper left quadrant; the square on landmarks 10, 11, 13, and 12, the central four; and the square on the outside corners 1, 2, 4, 3. In Figure 8.8, by contrast, we rotate the square by 45°, comparing a selected small exemplar (6–10–13–5, on the lower edge) to the larger diamond on all four edge centers of the large square. In every panel of either picture, now that we are in the nonaffine subspace only two dimensions of shape variation remain to be displayed (i.e., the coordinates of each of the four landmarks plotted are equal or opposite), and all distributions are circular separately.

The results are unequivocal. The plots of the Procrustes squares show variances that differ by a factor of four for squares differing in edge length by a factor of two, and by a factor of two for the square versus the diamond on its diagonal. There is even a hint at a positional effect for identical squares in different positions within the configuration (Figure 8.7, upper central versus upper left panel – the variance-ratio here is significant with a p of about 0.01). By contrast, all five of the configurations in the lower row, corresponding to the bending energy norm, have effectively the same variance regardless of scale, position, or orientation.

> Hence, obviously, it is this shape space, not the Procrustes shape space, against which the biologist should be assessing shape covariances of empirical data sets

> whenever the purpose is to search for patterns of shape features that may or may not span the length and breadth of an organism over processes of growth, aging, or evolution.

We can put all this another way: comparisons of the same Procrustes length can have substantially different bending energies, and thereby substantially different feature languages. Figure 8.9 shows the scatter of Procrustes distance against bending energy for a simulation of 1000 isotropic perturbations of a 5×5 square grid at unit spacing. (The correlation of these two metrics is about 0.76.) The horizontal segments indicate a slab from near the middle of this distribution that spans a more than twofold range of bending energy for nearly constant Procrustes distance. Figure 8.10 shears this slab into a square plot and puts a little icon for the actual grid transformation at every point. Those toward the left are less bent and those toward the right more bent *at given Procrustes length.* Figure 8.11, finally, extracts three representative grids from the extreme left and likewise at the right, showing how the contrast between Procrustes length and bending energy is, precisely, the contrast between short-range and longer-range disorder in the shapes of the little grid polylines here. The simulations in the upper row, based in the 5×5 equivalent of the shape coordinates of the right-hand panel in Figure 8.6, would be a carefully structured oversampling of the perturbations tending to have the least bending energy for each slab in Figure 8.9.

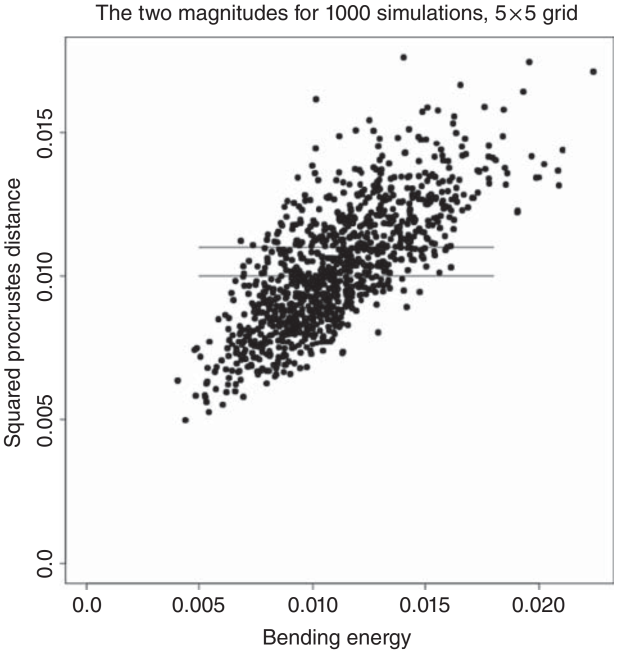


Figure 8.9 Squared Procrustes length versus bending energy for an isotropic sample of perturbations of a square 5×5 grid. The short segments indicate the slab extracted for closer examination in the next figure.

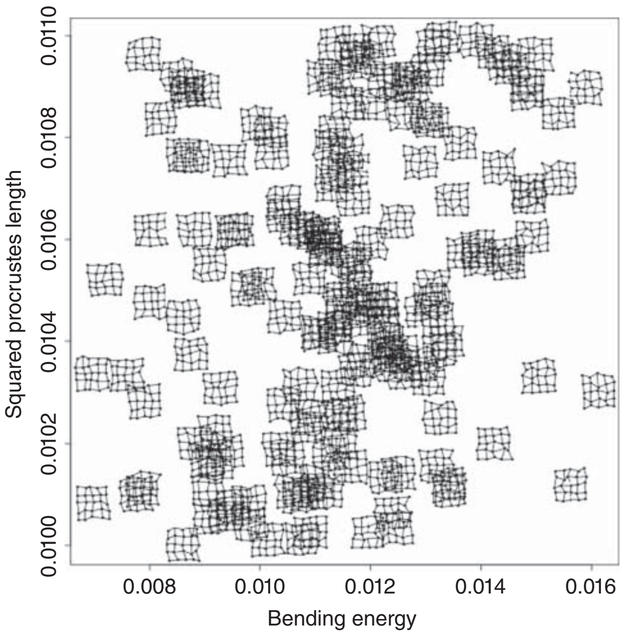


Figure 8.10 Vertical expansion of the slab from Figure 8.9, including an icon for each of the grid perturbations in this region.

8.4 Two examples

I would like to demonstrate the usefulness of this approach by way of two examples, each deriving from a data set that has been the object of study and manipulation not only by me but also by Mardia himself.

The first of these is my data set of midsagittal corpus callosum outlines gathered on a sample of 45 adult Seattle males, 30 of them with a pre-existing diagnosis in the fetal alcohol spectrum and the other 15 apparently normal. Mardia's description of these data can be found in Mardia et al. (2013). The Procrustes scatter we are examining is as in the guide figure in Figure 8.12(a): 40 points, of which only one is a proper landmark (the rest being sliding semilandmarks as explained in, e.g., Bookstein 2014), for the two-dimensional projection of the points along the curve of greatest local symmetry following around the waist of the corpus callosum, the neural structure that links the two halves of the cortex of the human brain. For the isotropic offset Gaussian distribution on this mean form, the rotation to eigendirections of our bending energy formalism is a function of the mean form alone; then it ought not to be associated with variances after the rotation. However, when we actually examine this pattern, Figure 8.12, we see that that naïve expectation of isotropy is not fulfilled. Instead, variances drop in nearly inverse proportion to specific bending energy, connoting exactly the pattern of self-similarity we just confirmed in Figures 8.7 and 8.8. In this interpretation, the emergence of ordinary Procrustes principal components (cf. Bookstein et al., 2001, 2002) as large-scale aspects of variation is not a property of the tissue per se, but only of the method of analysis.

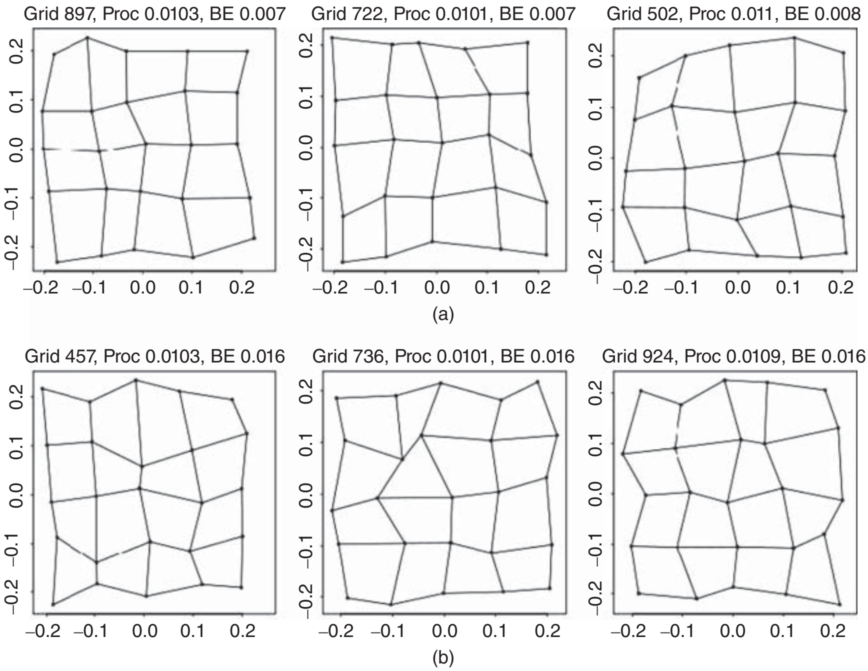


Figure 8.11 Grid transformations of extremely low (a) or high (b) bending for the given Procrustes distance. These are the three leftmost and rightmost grids, respectively, in the preceding figure.

A second example shows a falloff of variance with bending energy that is even more rapid than what we just saw in the brain data. The data undergoing reanalysis are the celebrated *Vilmann rodent skull octagons* originally analyzed by me in Bookstein (1984) and subsequently reanalyzed in (Bookstein 1991; 1994; 2014; 2015), Bookstein and Mitteroecker (2014), and Kent et al. (2001). There are eight landmarks involved, hence five nontrivial eigenvalues of the bending energy matrix. Figure 8.13 shows the five corresponding eigenvector scores as Cartesian (x, y) pairs (the so-called *partial warp scores*) along with the uniform term, which does something quite interesting over the course of growth from age 7 to 150 days in these 18 rats. At the lower right is a sketch showing the lie of this octagon *in situ* (the rat is facing to the right).

Following a hint from Bookstein (1991), it helps to approach trends such as these as dimensions of a composite *quadratic growth-gradient* that can be fitted by appending six more terms to the two-dimensional affine subspace that is already part of the standard decomposition (see the J-matrix in Bookstein 2014, Section 7.3 [which was originally work joint with Mardia]). In this sample of forms, there is only one dimension of such a quadratic component, corresponding to the left-hand grid in Figure 8.14. This feature does not, however, exhaust the correlated variation in the data. There is also an entirely local (i.e., not quadratic) term in the ordinary first Procrustes principal component of the nonaffine change (Figure 8.14(b)). This pattern modifies the growth gradient by a little twitch at the upper left (the top of the back of the head) that matches the corresponding grid feature at the upper left in the grid for the last (most bent) partial warp, Figure 8.13 (far right, second row).

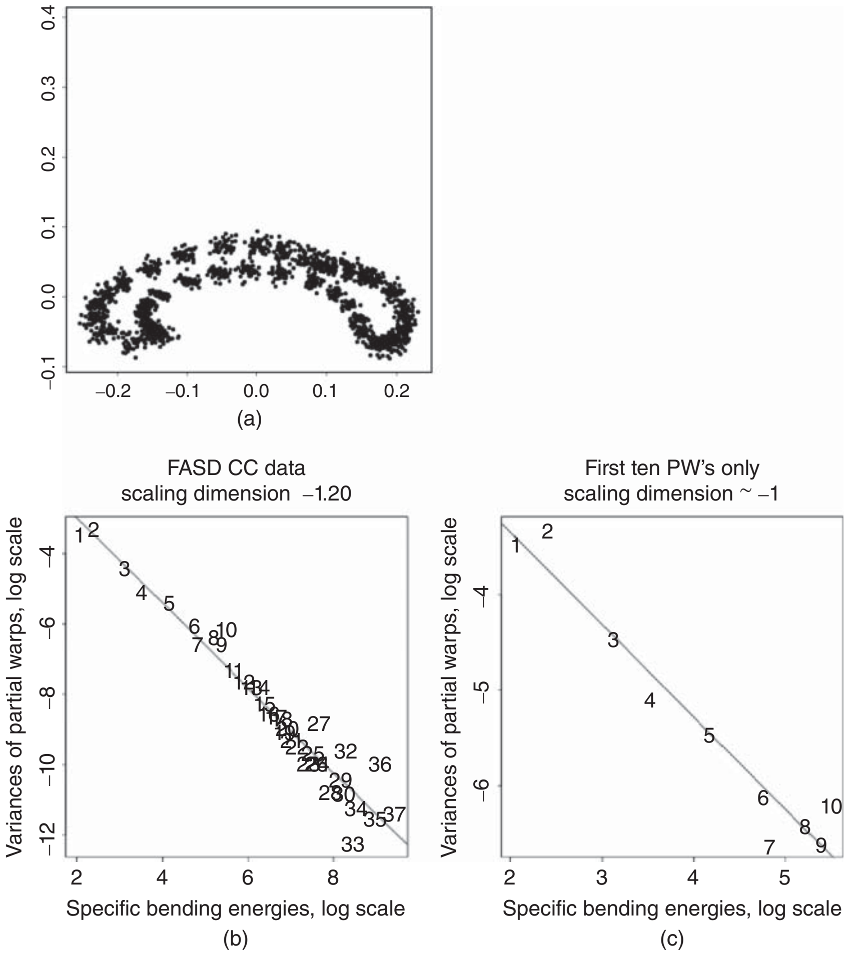


Figure 8.12 Toward the large-scale end of the callosal data set, the scaling of variation by eigenvectors of bending energy falls exactly inversely to bending energy, resulting in a spherical distribution after the standardization. Variation of these extended neural structures thus appears to be self-similar in the sense of the text. (a) The original Procrustes scatter, 40 points by 45 cases. Lower row: least-squares estimates of scaling dimension, all partial warps (b) or the 10 of the largest scale only (c).

We can move this into a more formal modeling context by actually plotting the variances of the partial warps (projections of the eigenvectors of bending energy) by their eigenvalue. As Figure 8.15 shows, these fall into two sets: a quartet with variance dropping with bending even faster than inversely, left panel, together with an orphan term, the smallest scale partial warp, showing specific local variation over time. In place of the loglinear fit to the scaling subset, at left, one might explore a modification that includes a nugget effect like the one previously suggested by Mardia et al. (2006) for the analogous application to variance components alone. In the present setting, a nugget would stand for an irreducible

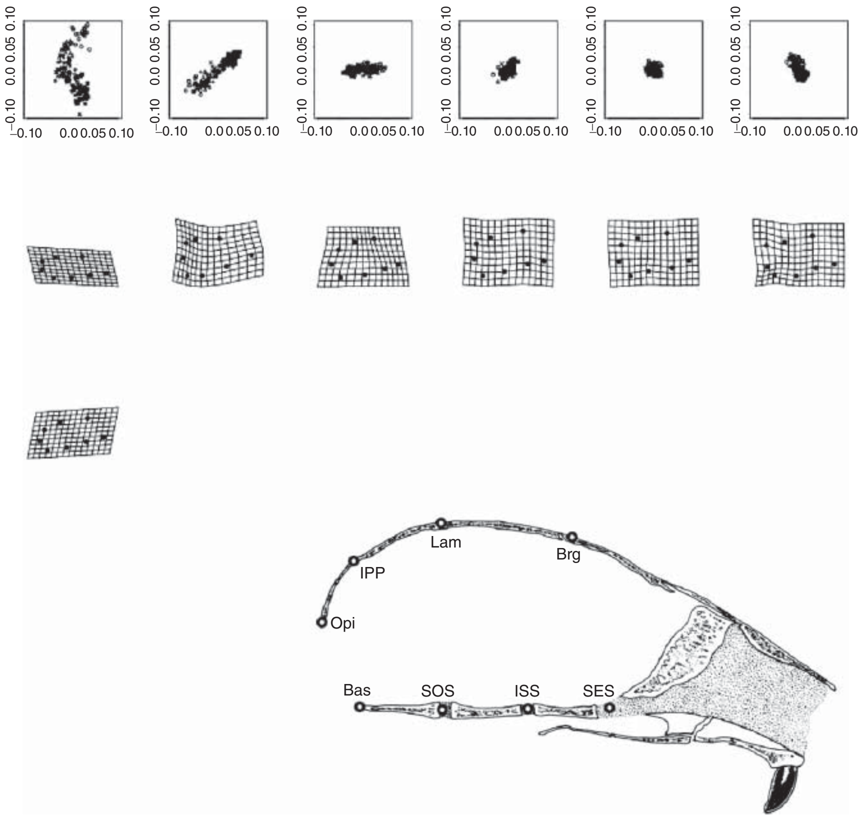


Figure 8.13 The five partial warp score two-vectors, plus the uniform component, for the 18 Vilmann rat skulls imaged at eight ages each. Each partial warp is shown along the principal component of its own partial warp scores, except that the uniform term (far left column) is displayed separately for its 7-to-30-day (upper) and 30-to-150-day (lower) orientations. Below right: the octagon of landmarks for a typical specimen cut midsagittally (up the middle of the skull, *Source:* from Bookstein 1991.

component of landmark perturbation that is uncorrelated with everything else, near or far, in its diagram. A least-squares estimate of this nugget term is 0.01142, which is most of the variance 0.016 of the second-last partial warp in Figure 8.13 but only a small fraction of the variance 0.21 of the largest-scale partial warp and an even smaller fraction of the variance 0.37 of the uniform term. Of course, in a sample of a mere 18 growing rats observed at 8 landmarks only, it is pointless to argue that the magnitude of this nugget effect has been identified with any precision.

This graphical representation with respect to the a priori basis of the partial warps, as exemplified in Figure 8.12 or Figure 8.15, is a limited view of a more formalized multivariate analysis, the *relative eigenanalysis* of Procrustes shape covariances with respect to the bending energy matrix. The algebra of this matrix maneuver, which is classical, was recently reviewed in Bookstein and Mitteroecker (2014) in a more general biometrical context. Its role in the modeling of scale-*specific* morphometric features was previously sketched in

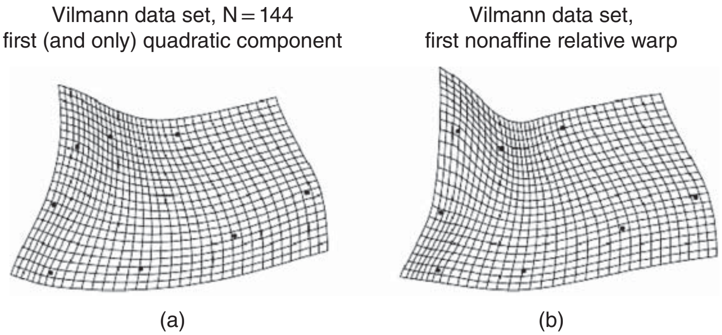

Vilmann data set, N = 144
first (and only) quadratic component
Vilmann data set,
first nonaffine relative warp
(a)
(b)

Figure 8.14 Summary of the single dimension dominating this covariance structure. (a) The large-scale (quadratic) growth gradient. (b) The first principal component of nonaffine growth, combining this gradient with a spatially focused additional feature at IPP (upper left margin of the configuration).

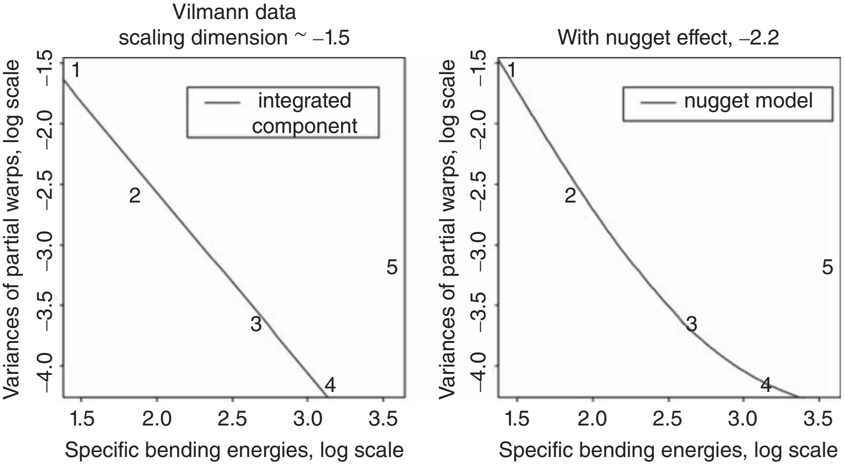


Figure 8.15 More detailed model of the Vilmann rodent skull. After the nugget correction, the diminution of variance with bending scales with dimension −2.2, much faster than self-scaling. The local term (fifth partial warp; Figure 8.13, rightmost column) is an entirely separate feature.

Bookstein (2007), under the name of "relative intrinsic warps," and well before that was introduced via the "relative warps" defined in Section 7.6.1 of Bookstein (1991). But please note that the meaning of the technical term "relative warps" has changed since then. The subject in 1991 was what Rohlf (1993) renamed "relative warps with $\alpha = 1$," whereas today's relative warps are, by convention (at least, within the community of people saying that they are using "geometric morphometrics"), those with $\alpha = 0$ – see, for example, Weber and Bookstein (2011), Chapter 4.

8.5 Some final thoughts

The point I am making is a methodological one, and it is far from trivial. For a statistical method to be of much use in any specific scientific application (here, in organismal biology), its models of variability and noise must be aligned to some extent with the actual patterns of variability or noise that characterize the real phenomena of that science or the explanations that customarily are found to account for those phenomena. In the application to covariance structures of biological shape, for the last twenty years or so our attention has been uncomfortably stretched between two extreme poles: the isotropic offset Gaussian model, which is totally disorganized and thus could not correspond to patterns from any actual living thing (if your features manifested no meaningful covariances across space or time, you would not have been liveborn), and the general model of unspecified positive-semidefinite covariances, which, in morphometric applications, has far too many parameters to be of much use in discriminating between models of quite different import that are equally reasonable a priori. This unsuitability of current statistical shape analysis for studies of actual living things has been masked by the simultaneous turn to the models of industrial biometrics and proteomics, which emphasize molecules over larger scale features of organisms and also privilege strategies of identification or classification in preference to the search for trends, equilibria, or other dynamical modes of explanation. In effect, we have been emphasizing control over understanding here in morphometrics, and we have been doing so long enough that our main toolkit has actually suffered some deformation of its own.

The model of self-scaling in landmark systems serves two functions, then. First, it has arisen as a point null model, a specific proposed covariance structure that is not remotely the sphere of the isotropic Procrustes shape model but that can nevertheless serve as a plausible null in some circumstances, for example, the regulation of brain form as we explored it in Figure 8.12. Its second function, though, might be even more important. If you ignore the nugget, which was just me showing off, the slope of the curve in Figure 8.15 is a *single* additional parameter for the organization of complex biological systems, a parameter that links the case of complete biomechanical homogeneity (which involves no gradients at all) to the case of isotropic Gaussian variation (which involves no organization) by a one-parameter family having a specific meaningful value of 1 (self-scaling) *along its dimension.* We can thus decompose biologically meaningful shape variations by a scheme that brings with it a system of the associated rhetorical tools: self-similarity or, by contrast, large-scale gradients (the curve in Figure 8.15) often paired with local features such as the outlying partial warp 5 here (which is too local to be self-scaling even while the rest of these rodent skull configurations are too *global* to be).

In the morphometrics of organisms, we do not want our null models to involve meaningless noise, the way they do in ordinary linear modeling. Instead, we want the noise terms to be *meaningful* expressions of the part of biological shape that is in fact being ignored by the organism, presumably for good reason, at the same time that it is actively managing the rest. (This has been a theme of theoretical biology for at least 50 years. Consider, for instance, the comment from the embryologist Paul Weiss during the 1956 meeting on “concepts of biology” (Weiss 1958): “Identical twins are much more similar than any microscopic sections from corresponding sites you can lay through either of them.” Or the equally nuanced insight of Walter Elsasser, in his *Chief Abstractions of Biology* (Elsasser 1977), that the crucial problem of representation in organismal biology is the selection of a finite set of constructs worth measuring out of the effectively infinite class of things that *might* be quantified, whether theoretically pertinent or not.) The issue for organismal biology is not mere hypothesis testing or the estimation of posterior probabilities. It is, rather, the far deeper issue of understanding the actual sources of variation encountered in samples of forms, and the developmental and evolutionary origins of that variation; likewise the sources of dimensions that do *not* vary, and the origins of that canalization; likewise the explanations of how features pass back and forth between these two complementary domains over the course of developmental and then evolutionary time. The models that appear in the molecular sciences are impoverished by comparison, as they offer so much less to explain.

The set of strategies that this organismal context suggests for spatiotemporal methodology, strategies that seem mostly to be missing from the work of others over in the related domain of geostatistics (see, e.g., Cressie and Wikle 2011), can be viewed as a generalization of the emphasis that was already present in Mardia's work on bilateral symmetry in the late 1990s. Bilateral symmetry is, after all, a version of integration, probably the most intuitively familiar we've got. Our reinterpretation of symmetry analysis in terms of complementary subspaces of symmetrical versus asymmetrical dimensions in shape space is a discrete analogue to the continuous rescaling of Procrustes variation according to bending energy that is being proposed here, following on Mardia's work of a decade ago, and the notion of the hyperplane of exactly symmetric structures against which all this geometry of asymmetry is calibrated is the analogue of the exactly self-scaling covariance structure that the corpus callosum data set here hints at, the structure that at last justifies the exactly self-scaling intrinsic random field models proposed in Mardia et al. (2006).

With Mardia it usually works like that. Typically he has his tools in hand years or decades before they are called out by the queries of others. All through the course of his long career, Kanti Mardia has made a habit of unearthing fundamental aspects of applied statistics along these lines – places where radical changes in the scope of uncertainty that is being modeled, and in the style of that modeling, bear huge implications for the understanding of the signal that remains – by seeing the corresponding analytic possibilities before anybody else has suspected their existence. Always, too, he has pursued the specific sort of concern I am concerned with here: the provenance of a single new parameter that enables the radical reorganization of one or another applied field. In this domain of descriptive features of landmark configurations, his fundamental intellectual strategies – the tie between splines and kriging, the role of Bayesian inferences in high dimensions, the willingness to dive into virtually any applied domain in search of new problems – usually provide the rest of us with the best hints we can muster about the directions in which to search for the most promising new discrete parameters.

Mardia's work on the spatial structure of deformation maps, greatly underappreciated in its initial incarnation, deserves far more attention than it has thus far received. Once the associated display conventions are more fully developed, the feature language it sketches against a background of self-similar processes will bear powerful implications for a whole host of spatiotemporal problems in organized systems from the scale of the Earth or even the solar system down to climate, biomes (ecosystems), single organisms, and their organs, tissues, and, ultimately, molecules. If the job of the applied mathematician is to find what is mathematizable in the world, then Mardia's self-assigned task has always been to find what is mathematizable in our *uncertainty* about the world, and then phrase the new language(s) necessary for reporting on that uncertainty. He has pursued this goal doggedly for more than half a century, and his innovations now make possible a thoroughgoing reformulation of the very field, geometric morphometrics, that he was responsible for bringing to the attention of the mathematical statisticians in the first place. I am honored to be part of this Festschrift.

8.6 Summary

The familiar Procrustes metric of contemporary morphometrics is fundamentally unsuited to organismal biology for a variety of reasons. One is the unrealistic nature of its symmetries, which involve uncorrelated errors at every digitized landmark separately *contra* the biologist's intuition of organisms as integrated systems of very high dimension. Another is the refractory nature of its covariance modeling: either a sphere in shape space, which is wholly unrealistic in these applications, or else some version of a general positive-semidefinite alternative that affords no practical possibility of meaningful biometrical pattern analysis. This essay reminds the reader of a different possibility: analysis with respect to the *bending energy metric* of the associated thin-plate splines, a metric closely associated with the intrinsic random field model of Mardia et al. (2006). To this approach corresponds a postulate of self-similar shape variability apparently aligned both with the cognitive psychology of the search for characteristics of groups in systematic biology, physical anthropology, and medicine and with the rhetoric used to convey such patterns once unearthed. The model appears relevant to understanding two of our standard data sets, the Vilmann rodent skull octagons and the midline corpus callosum semilandmark 40-gons from my study of the brain in the fetal alcohol spectrum disorders. A closing comment speculates on how this project exemplifies Kanti Mardia's approach to statistical science in general.

Acknowledgments

Some of the work reported here was supported by grant DEB–1019583 from the National Sciences Foundation of the United States to Joseph Felsenstein and F. L. Bookstein. For fruitful discussions on many of these points I am grateful to Philipp Mitteroecker (University of Vienna) and, of course, to Kanti Mardia as well.

References

Bookstein FL 1984 Tensor biometrics for changes in cranial shape. *Annals of Human Biology* **11**, 413–437.

Bookstein FL 1989 Principal warps: thin-plate splines and the decomposition of deformations. *I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence* **11**, 567–585.

Bookstein FL 1991 *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

Bookstein FL 1994 After landmarks In *Modern Morphometrics in Physical Anthropology* Slice DL (ed.), Kluwer Academic Publishers, New York, pp. 49–71.

Bookstein FL 2007 Morphometrics and computed homology: an old theme revisited In *Proceedings of a Symposium on Algorithmic Approaches to the Identification Problem in Systematics* MacLeod N (ed.), Museum of Natural History, pp. 69–81.

Bookstein FL 2014 *Measuring and Reasoning: Numerical Inference in the Sciences*. Cambridge University Press.

Bookstein FL 2015 *Biometrics and Morphometrics for Anthropologists*. Book manuscript under review, Cambridge University Press.

Bookstein FL and Mitteroecker PM 2014 Comparing covariance matrices by relative eigenanalysis, with applications to organismal biology. *Evolutionary Biology* **41**, 336–350.

Cressie N and Wikle CK 2011 *Statistics for Spatio-temporal Data*. John Wiley & Sons.

Elsasser W 1977 *The Chief Abstractions of Biology*. North-Holland.

Kent JT, Mardia KV, Morris RJ and Aykroyd RG 2001 Functional models of growth for landmark data In *Proceedings in Functional and Spatial Data Analysis* Mardia KV and Aykroyd RG (eds), Leeds University Press, pp. 109–115.

Mardia KV, Bookstein FL and Kent JT 2013 Alcohol, babies, and the death penalty: saving lives by analysing the shape of the brain. *Significance* **10**(2), 12–16.

Mardia KV, Bookstein FL, Kent JT and Meyer CR 2006 Intrinsic random fields and image deformations. *Journal of Mathematical Imaging and Vision* **26**, 59–71.

Mardia KV, Bookstein FL and Moreton IJ 2000 Statistical assessment of bilateral symmetry of shapes. *Biometrika* **87**, 285–300.

Rohlf FJ 1993 Relative warp analysis and an example of its application to mosquito wings In *Contributions to Morphometrics* Marcus LF, Bello E and Garcia-Valdecasas A (eds), Museo Nacional de Ciencias Naturales (CSIC) Madrid, Spain, pp. 131–159.

Weber GW and Bookstein FL 2011 *Virtual Anthropology: A Guide to a New Interdiscipinary Field*. Springer-Verlag.

Weiss PA 1958 [comments] In *Concepts of Biology* Gerard RW (ed.), National Academy of Sciences, Publication 560.

9

Nonparametric data analysis methods in medical imaging

Daniel E. Osborne[1], Vic Patrangenaru[2], Mingfei Qiu[2] and Hilary W. Thompson[3]

[1]*Department of Mathematics, Florida Agricultural and Mechanical University, Tallahassee, FL, USA*

[2]*Department of Statistics, Florida State University, Tallahassee, FL, USA*

[3]*School of Medicine, Division of Biostatistics, Louisiana State University, New Orleans, LA, USA*

9.1 Introduction

Shape-based statistical methods for medical imaging started around the early nineties (see Bookstein 1991; Dryden and Mardia 1998) and the first nonparametric methods started being used slightly later. A classical medical imaging library that was heavily used in developing nonparametric tests in medical imaging was the one resulting from the Louisiana Experimental Glaucoma Study (LEGS), consisting of two types of imaging outputs: Heidelberg Retina Tomograph (HRT) images and stereo pairs of images of the back of the eye (see Burgoyne et al. 2000). The LEGS images are from Rhesus monkeys retinae. Tragically, the animals survived all the experiments, only to fall victims of the hurricane Katrina in 2005. In each of the individuals in the LEG study, an increased internal ocular pressure (IOP) was induced in one eye, while the other eye was left as control. Both eyes were imaged, and for

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

each individual in the study, a complete set of observations, both HRT and stereo pairs were stored. The stereo pairs consisting of four optic nerve head (ONH) images were processed only in 2008 or later. They consisted of two images of the control eye (A) and two of the treated eye (B). Section 9.2 is dedicated to a review of results of nonparametric shape data analysis for HRT and stereo LEGS library data. HRT image data allows a recovery of the similarity shape information; therefore, for such data, the analysis is performed on the space Σ_3^k of direct similarity shapes of k-ads in 3D, known as *Kendall shape space*. Along these lines, we recall results from Derado et al. (2004) and from Bhattacharya and Patrangenaru (2005). For the stereo LEGS data, the camera parameters are unknown, thus only 3D projective shape data could be recovered. A 3D projective shape change analysis due to Crane and Patrangenaru (2011) is, therefore, pursued in this part of Section 9.2.

In Section 9.3, we focus on the important task of recovery of 3D data from CT scans of the human skull. This task includes preprocessing and postprocessing steps for CT images. The preprocessing step consists of the extraction the boundary of the bone structure from the CT slices, while the postprocessing step consists of 3D reconstruction of the virtual skull from these bone extractions. Given that the bilateral symmetry of the skull allows for a 3D size-and-reflection shape analysis on a manifold, therefore, in Section 9.4, we briefly introduce the general nonparametric bootstrap on manifolds methodology, based on extrinsic means and extrinsic sample covariance matrix computations. Next, in Section 9.5, we introduce in detail the 3D size-and-reflection shape space $SR\Sigma_{3,0}^k$, as orbifold (space of orbits of the action of the orthogonal group $O(3)$ on centered k-ads in general position in $\mathbb{R}^3$. The Schoenberg embedding, the Schoenberg extrinsic mean, and the asymptotic behavior of the Schoenberg sample mean are also given in this section, for which the main reference is Bandulasiri et al. (2009). In Section 9.6, we present preliminary results for skull shape analysis based on bootstrap distributions of the Schoenberg's sample mean size-and-reflection shape for a selected group of k anatomical landmarks, and report a confidence region for the Schoenberg mean configuration of the corresponding k-ads on the midface.

The third part of the chapter is dedicated to examples of nonparametric analysis on homogeneous spaces applied to MRI brain imaging. The first example, following results from Osborne et al. (2013), is given in Section 9.7. There a two-sample test for DTI intrinsic means, based on their nonparametric methodology, was applied to a concrete DTI small data set previously analyzed by Schwartzman et al. (2008), consisting of a voxelwise comparison of spatially registered DT images belonging to two groups of children, one with normal reading abilities and one with a diagnosis of dyslexia. The data provides strong evidence of differences between the intrinsic means of the two groups. The second example in Section 9.8 is on the infinitely dimensional homogeneous space of direct similarity shape of contours, in the context of neighborhood hypothesis testing on manifolds, as it was recently developed in Ellingson et al. (2013). As an illustrative application, a test is carried out to see how far is the average direct similarity shape of contours of the midsection of corpus callosum in elderly people from that of Albert Einstein, at the time when he just passed away.

9.2 Shape analysis of the optic nerve head

Since glaucoma is a disease affecting the 3D appearance of the ONH region due to high IOP, it leads to a change in the 3D shape of this region. Given the small sample size of the LEGS

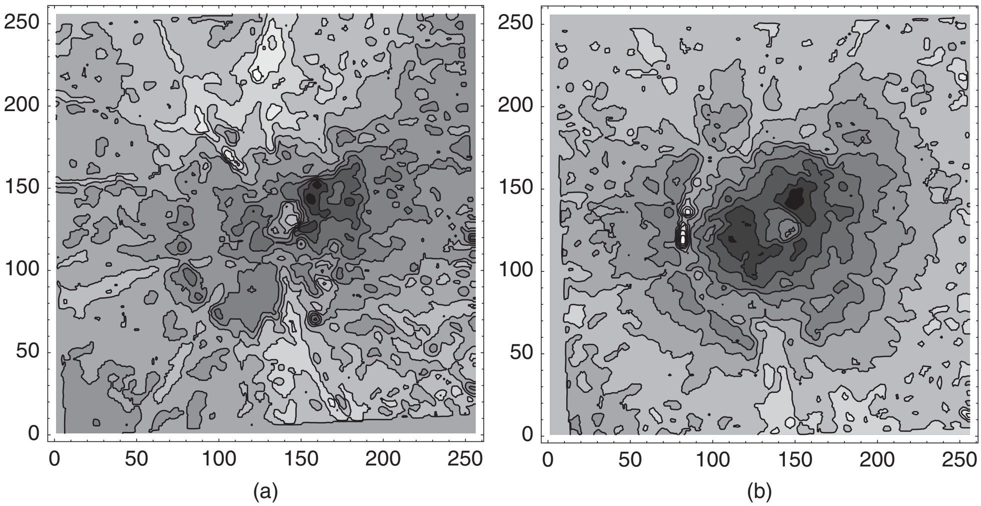


Figure 9.1 Change in the ONH topography from normal (a) to glaucomatous (b) (*Source:* Derado et al. 2004, Figure 3, p. 1243. Reproduced by permission of Taylor and Francis http://www.tandfonline.com).

data, any analysis has to undergo a drastic dimension reduction. For shape data, a first step in the dimension reduction consists of a selection of a few significant anatomical landmarks. In the case of HRT outputs, each "image" was presented in the form of a series of 256×256 2D arrays of ONH height values from a plane spanned by the ridge of the ONH cup. Due to the increased IOP, as the soft spot where the ONH enters the eye is pushed backward, eventually, the optic nerve fibers that spread out over the retina to connect to photoreceptors and other retinal neurons can be compressed and be damaged. Two processed images of the ONH cup surface before and after the IOP increment are shown in Figure 9.1.

Regarding landmark-based dimension reduction analysis, assume that the position vectors of these landmarks are $X_1, \dots, X_k, k \geq 4$. Two configurations of landmarks have the same *Kendall shape*, if they can be superimposed after a direct similarity. The *Kendall shape* of the configuration $x = (x_1, \dots, x_k)$ is labeled $o(x)$ and the space Σ_m^k, of shapes of configurations of k points in $\mathbb{R}^m$ at least two of which are distinct introduced in Kendall (1984) is the *Kendall shape space* of k-ads in m dimensions.

We now return to the shape of an ONH region, which resembles a "cup" in the shape of half an ellipsoid with an ellipse-shaped margin. Following Patrangenaru et al. (2000), in Bhattacharya and Patrangenaru (2005) four landmarks were used; the first three, denoted by S, T, and N were chosen to be the "top, left, and right" points on this ellipse, that is, (when referring to the left eye) Superior, Templar, and Nose papilla (see Derado et al. 2004). The fourth landmark, V, was called *vertex*, the deepest point inside the ellipse bordering the ONH cup; therefore, in Bhattacharya and Patrangenaru (2005) the data analysis was carried out on the shape space of tetrads, Σ_3^4, which is topologically a 5 dimensional nonstandard sphere, according to Kendall et al. (1999), p. 33. On the other hand, it is known that if a probability distribution on Σ_m^k has small support outside a set of singular points, any distance that is compatible with the orbifold topology considered is not relevant in data analysis

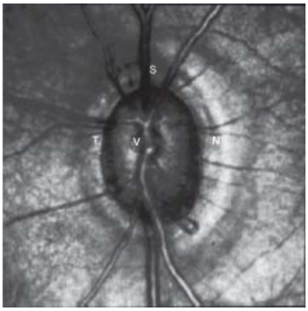


Figure 9.2 Landmarks on the ONH for HRT outputs (*Source*:Derado et al. 2004, Figure 1, p. 1242. Reproduced by permission of Taylor and Francis http://www.tandfonline.com).

(Dryden and Mardia 1998, p. 65) as the data can be linearized. Therefore, for practical considerations, distances that are derived using a partition of unity and Euclidean distances in various coordinate domains of Σ_3^4 are useful for such distributions. In Dryden and Mardia (1998), pp. 78–80, five coordinates, that were later called *DM coordinates* by Bhattacharya and Patrangenaru (2005), were defined on the generic subset of Kendall shapes of nondegenerate tetrads in Σ_3^4, and labeled $v^1, \dots, v^5$. The five DM coordinates proved useful in detecting a significant glaucomatous means shape "difference" due to the increased IOP, as shown in Bhattacharya and Patrangenaru (2005). Nevertheless, since it was preferable to have a single medical measurement to detect glaucoma from HRT outputs, Derado et al. (2004) defined a *glaucoma index* and showed that this index is useful in mean shape change detection (Figure 9.2). Due to its simplicity, the landmark-based glaucoma index method, is cited in the medical literature (see Hawker et al. 2007; Sanfilippo et al. 2009).

In the case of stereo data of the back of the eye, which is the most common imaging data for eye disease detection, Crane and Patrangenaru (2011) developed a landmark-based projective shape analysis approach. They analyzed data from LEGS consisting of fifteen independent complete paired observations of stereo pairs. Figure 9.3 displays the fifteen independent stereo pairs of observations. Unlike with HRT data, which is 3D from the outset, in the case of stereo imaging, one has to retrieve the 3D structure of the landmark configuration from its stereo pair images. The problem of reconstruction of a 3D configuration of points from a pair of its *ideal noncalibrated camera* images was solved by Faugeras (1992) and Hartley et al. (1992), who showed that:

Theorem 9.2.1 *A finite configuration $\mathcal{C}$ of eight or more points in general position in 3D can be reconstructed from the coordinates of the images of these points in two ideal noncalibrated digital camera views, and the reconstruction $\mathcal{R}$ is unique up to a projective transformation in 3D.*

This projective ambiguity in Theorem 9.2.1 was reinterpreted in Sughatadasa (2006), Patrangenaru et al. (2010), Crane and Patrangenaru (2011) as follows:

Corollary 9.2.2 *The projective shapes of the 3D configurations of points $\mathcal{R}$ and $\mathcal{C}$ in Theorem 9.2.1 are identical.*

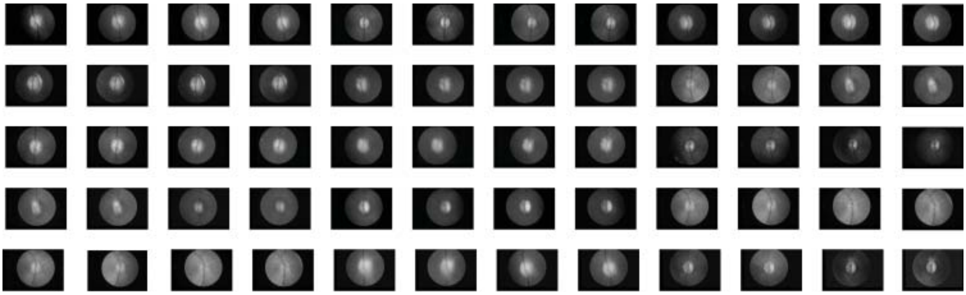


Figure 9.3 Optic nerve head region data (*Source*: Crane and Patrangenaru (2011), Figure 2, p. 232. Reproduced by permission of Elsevier).

Among reconstruction algorithms, Crane and Patrangenaru (2011) suggested using the eight-point algorithm in Ma et al. (2006, p. 121), for a conveniently selected camera internal parameters matrix, or the refined eight-point algorithms for the estimate of the fundamental matrix (Ma et al. 2006, p. 188, p. 395), (Hartley and Zisserman 2004, p. 282). For details on the reconstruction of the projective shape of a 3D configuration from the pixel coordinates of two of its digital images, see Patrangenaru et al. (2010) and the references therein. In a study by Crane and Patrangenaru (2011), coordinates of nine landmarks on the approximate elliptic contour of the ridge of the ONH were recorded, as well as those of certain blood vessels junctions and estimated location of the deepest points. These included the landmarks considered for HRT data. They were S(superior), I(inferior), N(nasal), T(templar), V(vertex-the deepest point of the ONH cup), SM(mid-superior), IM(mid-inferior), NM(mid-nasal), and TM(mid-templar), and their positions in the ONH cup are schematically displayed in Figure 9.4. Note that projective shape analysis can be performed using different approaches, by representing a projective shape on a certain projective shape space. The most recent approach, due to Kent and Mardia (2012), has the advantage of being independent of the landmark labels. On the other hand, the projective frame approach (Mardia and Patrangenaru 2005; Patrangenaru et al. 2010) has the advantage of being rooted in projective geometry and computationally faster; no 3D projective shape analysis was so far published using the approach in Kent and Mardia (2012). Moreover, in 3D, the projective shape space obtained via the projective frame approach has a Lie group structure, thus allowing a two-sample test for mean projective shape change in matched pairs to be reduced to a one-sample test. For such reasons, in their projective shape analysis of mean glaucomatous projective shape change, Crane and Patrangenaru (2011) used a projective frame approach, by selecting the projective frame $\pi = (\mathrm{N}, \mathrm{T}, \mathrm{S}, \mathrm{I}, \mathrm{V})$. For the analysis, the projective coordinates (defined in Mardia and Patrangenaru 2005) of the remaining four landmarks $[h_{1,ji}], [h_{2,ji}], [h_{3,ji}], [h_{4,ji}], j = 1, 2, i = 1, \ldots, 15$ were computed with respect to this frame. To test if there is a difference between the extrinsic mean projective shape change from the configuration in the control eye and the treated eye, given the small size of the

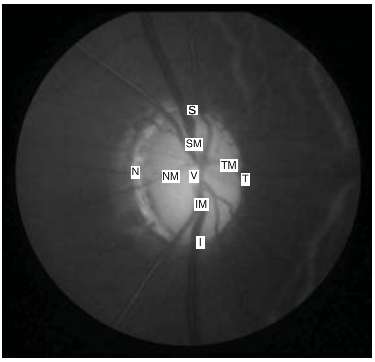


Figure 9.4 Nine anatomical landmarks of the ONH for one stereo image (*Source:* Crane and Patrangenaru (2011), Figure 3, p. 235. Reproduced by permission of Elsevier).

sample, Crane and Patrangenaru (2011) computed the bootstrap statistics $T_s^*, s = 1, 2, 3, 4,$ in Patrangenaru et al. (2010) for the four $\mathbb{R}P^3$ marginals for $20,000$ resamples. The histograms for the bootstrap distributions of $T_s^*, s = 1, 2, 3, 4$ corresponding to the marginal axes are displayed in Figure 9.5 (see also Crane and Patrangenaru 2011). The values of the statistics $T_s, s = 1, 2, 3, 4$ under the null hypothesis of *no projective shape change* are $T_1 = 1474.7, T_2 = 2619.9, T_3 = 860.2, T_4 = 1145.7$, and since the T_1, T_2, T_3, and T_4 are much larger than the corresponding cutoffs given earlier, there is a significant mean projective shape change due to the increased IOP in the treated eye.

It is worth noting that while test statistics for mean glaucomatous Kendall shape change based on HRT outputs, including tests for mean glaucoma index change in Derado et al. (2004) are easier to compute, most ophthalmologists cannot afford an HRT, while any ophthalmologist has access to stereo cameras designed for eye fundus imagery; thus, tests for mean projective shape change due to glaucoma onset might be more useful for the onset of glaucoma detection.

9.3 Extraction of 3D data from CT scans

In this section, our main focus is on preprocessing and postprocessing steps of CT images.

9.3.1 CT data acquisition

The CT images were taken using a computed tomography device (CT scanner). This was done for twenty-eight individuals. A computed tomography (CT) scan uses X-rays to make detailed pictures of structures inside the body. A CT scan is used to study all parts of the human body. In this study, one CT scan in our data set consists of about 100+ X-rays of the head above the mandible per individual. Figure 9.6 displays an example of one CT scan of an individual in our data set.

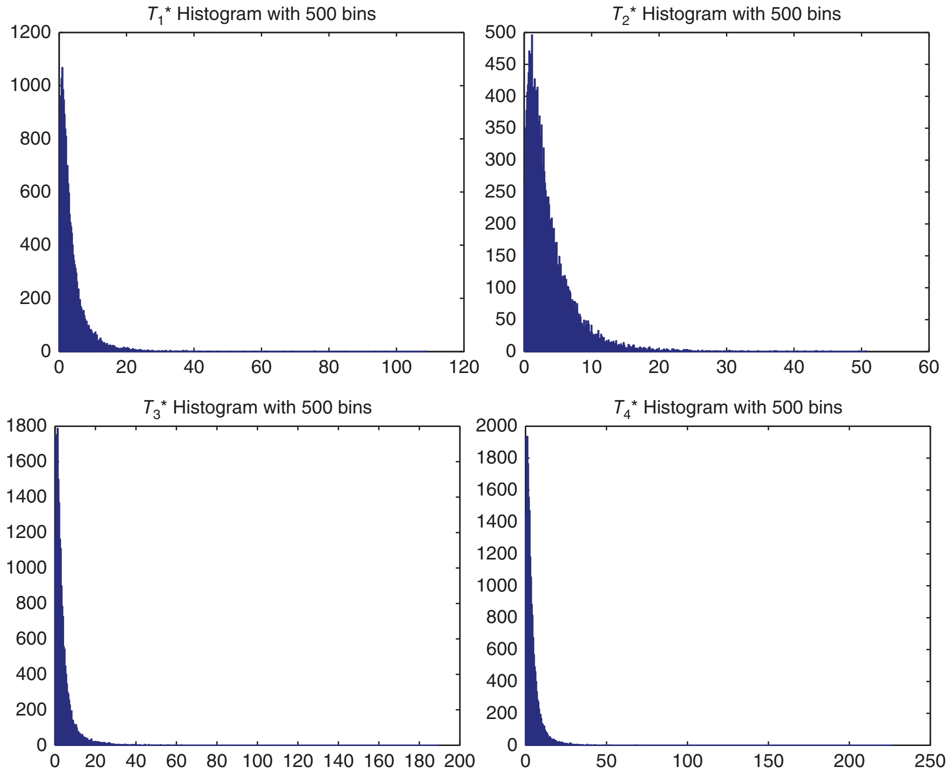


Figure 9.5 Histograms for the bootstrap distributions of $T_s^*, s = 1, 2, 3, 4$ for 20,000 resamples (*Source:* Crane and Patrangenaru (2011), Figure 4, p. 236. Reproduced by permission of Elsevier).

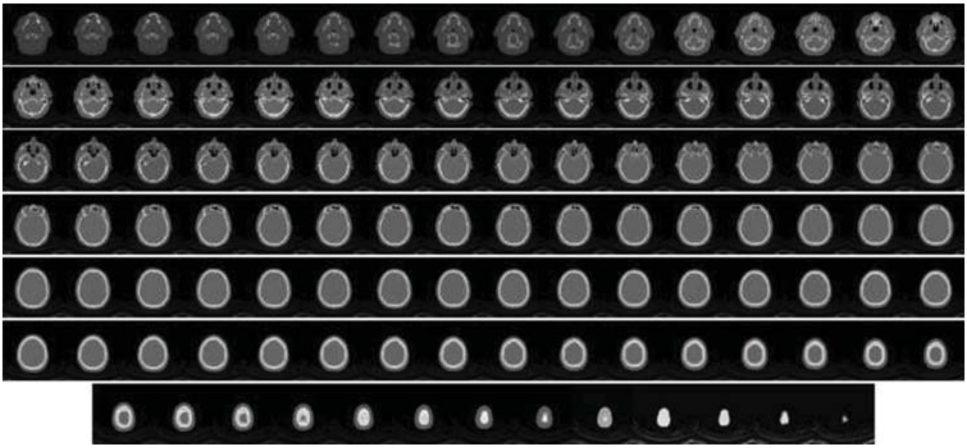


Figure 9.6 CT scan of an individual.

9.3.2 Object extraction

Numerous methods of thresholding and segmentation have been developed for object extraction from 2D for 3D display Harris and Camp (1984), Robb (1996), Serra et al. (1997), and Sher (1977). Surface rendering (Gibson et al. 1998; Heffernan and Robb 1985; Herman and Liu 1977; Lorensen and Cline 1987; and Robb 1994) and volume rendering (Cabra et al. 1994; Drebin et al. 1988; Kaufman et al. 1993; Levoy 1988; Pflesser et al. 1998; Robb 1994; and Robb and Barillot 1989) are two different techniques that have traditionally enabled the visualization of 3D biomedical volume data (images). Both techniques produce a visualization of selected structures in 3D volume data (images), but one should note that the methods involved in these techniques are quite different, and each has its advantages and disadvantages. Selection between these two approaches is often based on the particular nature of the biomedical image data, the application to which the visualization is being applied, the desired result of the visualization, and computational resources. Here, we focus on surface rendering techniques. Surface rendering, when based on a sequence of stacked 2D images, requires the extraction of contours (the edge of the intersection of the object with each slice, in our case of 2D slice level the skull surface). Then, a tiling algorithm is applied that places surface patches (or tiles) at each contour point and with hidden surface removal and shading, the surface is rendered visible. The advantage of this technique lies in the relatively small amount of contour data, resulting in a fast 3D rendering (reconstruction) speeds. The disadvantages may vary depending on object extraction (or segmentation) software or algorithms. Ideally, one would like to extract all objects of interest from 3D volume data (images) quickly and accurately. In other words, the extracted object should be a good representation of the original object inside the image. Here, we explored various segmentation methods in order to extract the bone structure from the CT slices and then perform 3D reconstruction of the virtual skull from these bone extractions.

9.3.2.1 Segmentation: minimizing the geodesic active contour

Segmentation is a well-studied area, and it is usually formulated as the minimization of a cost/energy function subjected to some constraints. Segmenting 3D image volumes slice by slice using image processing techniques is a lengthy process and requires a postprocessing step to connect the sequence of 2D contours into a continuous surface (3D reconstruction). Caselles et al. (1997) introduced the geodesic active contour (GAC) algorithm, as an enhanced version of the snake model of Kass et al. 1988. The GAC algorithm is defined as the following variation problem:

$$\min_{C} \{E_{\text{GAC}}[C]\}, \text{ where } E_{\text{GAC}}[C] = \int_0^{|C|} g(|\nabla I(C(s))|)dl. \tag{9.1}$$

In (9.1) $|C|$ is the Euclidean length of the curve C and dl the Euclidean element of arc. The edge detection function, $g \in (0, 1]$ in Equation (9.1) has the following meanings: values close to 0 are at strong edges in the image I, whereas values close to 1 are not at edges in the image I. $|\nabla I|$ acts as an edge detector. In particular, ∇I is the gradient of the gray level along the curve $C(s)$ Caselles et al. (1997). A (local) minimal distance path between given points is a geodesic curve. To show this Caselles et al. (1997) used the classical Maupertuis principle from dynamical systems (Caselles et al. 1997), which essentially explains when an energy minimization problem is equivalent to finding a geodesic curve in a Riemannian space (see also Milnor 1963). Typically, to find the global optimal solution of Equation (9.1), graph-based approaches are commonly used which rely on partitioning of a graph that is built based on the image I. Unfortunately, such approaches can lead to major systematic discretization error problems. Appleton and Talbot (2006) presented an approach that minimizes the GAC energy using continuous maximal flows. The amazing gain from their approach is that it does not suffer from any discretization errors. Bresson et al. (2005) produced a different approach, which uses the *weighted Total Variation*. The *weighted Total Variation* or simply *weighted TV* is defined as

$$TV_g(u) = \int_\Omega g(x)|\nabla u|dx. \tag{9.2}$$

$TV_g(u)$ is the weighted gradient of u. The active contour C is a level-set of a function $u : [0, a] \times [0, b] \to \mathbb{R}$. In other words, u is an implicit representation of the active curve C, since C coincides with the set of points $u = \text{c}onstant$. Bresson et al. showed that under certain conditions, namely if u is a characteristic function 1_C then Equation (9.2) is equivalent to E_{GAC} in (9.1). The details are provided in Bresson et al. (2005). In order to find the geodesic curve, the corresponding steepest descent flow of Equation (9.2) is computed. If we allowed u to vary continuously in $[0, 1]$, then Equation (9.2) becomes a convex function, meaning that one can compute the global minimizer of it. Unger et al. (2008) proposed the following variational image segmentation algorithm:

$$\min_{u\in[0,1]} \{E_{\text{Seg}}\} \text{ where } E_{\text{Seg}} = \int_\Omega g(x)|\nabla u|d\Omega + \int_\Omega \lambda(x)|u - f|d\Omega. \tag{9.3}$$

Here, the first term of the energy is the *weighted TV* of u as defined in Equation (9.2), which minimizes the GAC energy. The second term is used to incorporate constraints into the energy function. The variable $f \in [0, 1]$ is provided by the user, and it indicates foreground $(f = 1)$ and background $(f = 0)$ seed regions. The spatially varying parameter $\lambda(x)$ is responsible for the interpretation of the information contained in f. Figure 9.7 displays ten 3D reconstruction based on the method of Unger et al. (2008) summarized earlier.

9.4 Means on manifolds

9.4.1 Consistency of the Frećhet sample mean

Consider a separable metric space $(\mathcal{M}, \rho)$ and a random object

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathcal{M}, \mathcal{B}_\rho). \tag{9.4}$$

Given a probability measure Q associated with a random object X on a metric space $\mathcal{M}$ with the distance ρ, a natural index of location is the *Fréchet mean* (Fréchet 1948; Ziezold 1977) which is the minimizer of

$$F(p) = E(\rho^2(p, X)) = \int \rho^2(p, x)Q(dx), \tag{9.5}$$

if the minimizer is unique. The set of all such minimizers form the *Fréchet mean set*. Bhattacharya and Patrangenaru (2005) showed that if the Fréchet mean set has only one

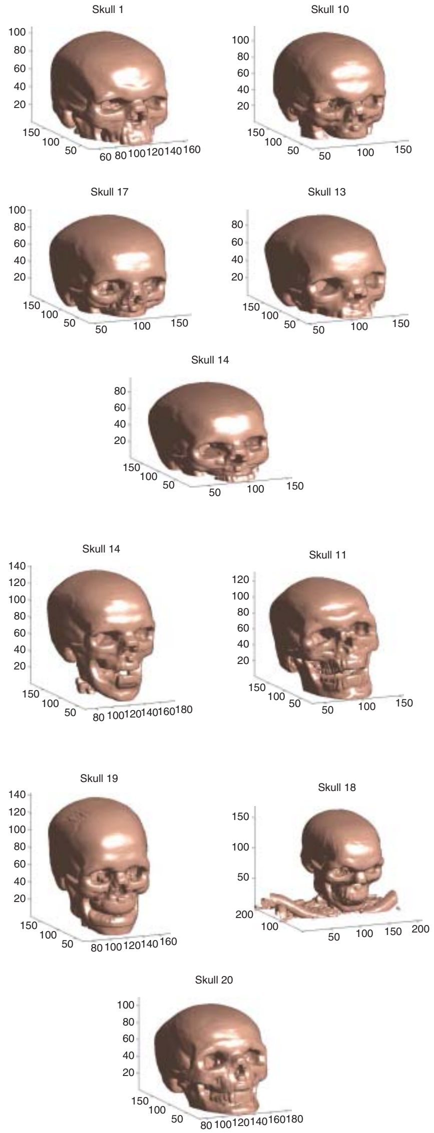


Figure 9.7 Select 3D reconstruction results via segmentation.

point (*Fréchet mean*), then the Fréchet sample mean (set) is a strongly consistent estimator of the Fréchet mean. In this paper, we define the definition of a distance between size-and-reflection shapes just like that presented in Bandulasiri and Patrangenaru (2005) and Bandulasiri et al. (2009). The Fréchet mean is called *extrinsic mean* if the distance ρ is induced by an embedding $j : M \to \mathbb{E}^N$, and *intrinsic mean* if the distance ρ is induced by a Riemannian structure on M. Furthermore, the extrinsic (intrinsic) sample mean is a consistent estimator of the extrinsic (intrinsic) mean.

The *projection map* $P_j : F^c \to j(M)$ is defined as

$$P_j(p) = j(x) \;\text{ if }\; d_0(p, j(M)) = d_0(p, j(x)), \tag{9.6}$$

where d_0 is the Euclidean distance and F^c is the set of nonfocal points of M in $\mathbb{E}^N$. In a study by Bhattacharya and Patrangenaru (2003), it was shown that if X is a j-nonfocal random object on $\mathcal{M}$, then the extrinsic mean is given by

$$\mu_j = j^{-1}(P_j(E(j(X)))), \tag{9.7}$$

where P_j is the projection on $j(M)$. Furthermore, if we let $X = (X_1, \ldots, X_n)$ be i.i.d. M-valued random variables with nonfocal measure Q on $(\mathcal{M}, j)$ and if the mean $\overline{j(X)}$ of the sample $j(X) = (j(X_1), \ldots, j(X_n))$ is a nonfocal point, then the *extrinsic sample mean* is given by

$$\overline{X}_j := j^{-1}\left(P_M(\overline{j(X)})\right). \tag{9.8}$$

The *extrinsic sample covariance matrix*, which shows in the *extrinsic* T^2 asymptotic statistics (see Bhattacharya and Patrangenaru 2005), is

$$\begin{aligned} S_{j,E,n} = &\left[\left[\sum d_{\overline{j(X)}} P_j(e_b) \cdot e_a(P_j(\overline{j(X)}))\right]_{a=1,\ldots,m}\right] \cdot S_{j,n} \\ &\left[\left[\sum d_{\overline{j(X)}} P_j(e_b) \cdot e_a(P_j(\overline{j(X)}))\right]_{a=1,\ldots,m}\right]^T, \end{aligned} \tag{9.9}$$

where $S_{j,n} = n^{-1} \sum_{r=1}^{N} (j(X_r) - \overline{j(X)})(j(X_r) - \overline{j(X)})^T$ is the sample covariance and $(e_a(y), a = 1, \ldots, N)$ is an orthoframe field around $P_j(\overline{j(X)})$, whose first m vectors are in $T_y j(M), y \in j(M)$, and $d_{\overline{j(X)}} P_j$ is the differential of P_j at the sample mean $\overline{j(X)}$.

9.4.2 Nonparametric bootstrap

Efron's nonparametric bootstrap methodology (Efron 1982) is extremely useful in data analysis on manifolds where the sample is small. If $\{X_r\}_{r=1,\ldots,n}$ is a random sample from the unknown distribution Q, and $\{X_r^*\}_{r=1,\ldots,n}$ is a bootstrap resample from $\{X_r\}_{r=1,..,.n}$, then $S^*_{j,E,n}$ is obtained from $S_{j,E,n}$ substituting $X_1^*, \ldots, X_n^*$ for $X_1, \ldots, X_n$. For example, if n is not large, from Bhattacharya and Patrangenaru (2005) it is known that a $100(1-\alpha)\%$ nonparametric bootstrap confidence region, for the extrinsic mean μ_j is given by $D^*_{n,\alpha} := j^{-1}(V^*_{n,\alpha})$, where

$$V^*_{n,\alpha} = \{\mu \in j(\mathcal{M}) : n\|S_{j,E,n}^{-\frac{1}{2}} \tan_{P_j(\overline{j(X)})}(P_j(\overline{j(X)}) - P_j(\mu))\|^2 \le d^*_{1-\alpha}\}. \tag{9.10}$$

Here $\tan_p(v)$ is the tangential component of v with respect to the splitting $T_v\mathbb{E}^N = T_vM \oplus (T_vM)^\perp$ and $d^*_{1-\alpha}$ is the upper $100(1-\alpha)\%$ point of the values

$$n\|{S^*_{j,E,n}}^{-\frac{1}{2}} \tan_{P_j(\overline{j(X^*)})}(P_j(\overline{j(X^*)}) - P_j(\overline{j(X)}))\|^2 \tag{9.11}$$

among the bootstrap resamples. This region has coverage error $O_p(n^{-2})$.

9.5 3D size-and-reflection shape manifold

9.5.1 Description of $\boldsymbol{SR\Sigma^k_{3,0}}$

We consider configurations $\mathbf{x} = (x^1, \ldots, x^k)$, which consist of $k > 3$ labeled points in $3D$, called k-*ads*. These k-ads are in *general position* (i.e., the minimal affine subspace containing the landmarks in $\mathbf{x}$ spans $\mathbb{R}^3$) and they represent k locations on an object. Translation is removed by centering the k-ad $\mathbf{x} = (x^1, \ldots, x^k)$ to

$$\xi = (\xi^1, \ldots, \xi^k), \xi^j = x^j - \overline{x}, \forall\, j = 1, \ldots, k. \tag{9.12}$$

The set of all centered k-ads form a vector subspace $L^3_k \subset (\mathbb{R}^3)^k = M(3, k; \mathbb{R})$ of dimension $3k - 3$, where

$$L^3_k = \{\xi \in M(3, k; \mathbb{R}), \xi \mathbf{1_k} = 0\}. \tag{9.13}$$

The orthogonal group $O(3)$ acts on L^3_k on the left, via the action α given by $\alpha(A, \xi) = A\xi$. The *3D size-and-reflection shape* $[\mathbf{x}]_{RS}$ of a k-ad $\mathbf{x}$ is the $O(3)$-orbit of the corresponding centered configuration ξ under the diagonal action $\alpha_k(A, \xi) = (\mathbf{A}\xi^{\mathbf{1}}, \ldots, \mathbf{A}\xi^{\mathbf{k}})$ of the orthogonal group $O(3)$ on the set of all centered k-ads:

$$[\mathbf{x}]_{RS} = \{A\xi : A \in O(3)\}. \tag{9.14}$$

A k-ad is in general position if and only if $\{\xi_1, \ldots, \xi_k\}$ spans $\mathbb{R}^3$. The *3D size-and-reflection shape space* $SR\Sigma^k_{3,0}$ is the set of all size-and-reflection shapes of k-ads in general position

$$SR\Sigma^k_{3,0} = \{[\mathbf{x}]_{RS},\ \text{rank}(\mathbf{x}) = 3\}. \tag{9.15}$$

This space is a manifold because the action of an orthogonal matrix on $\mathbb{R}^3$ is uniquely determined by its action on a basis of $\mathbb{R}^3$, and a centered k-ad in general position includes such a basis (Bandulasiri and Patrangenaru 2005). The dimension of $SR\Sigma^k_{3,0}$ is $3k - 6$. This space, $SR\Sigma^k_{3,0}$, can be represented as a quotient space $(L^3_{k,0} \backslash \{O_3\})/O(3)$, where $L^3_{k,0}$ is given by (9.13).

9.5.2 Schoenberg embeddings of $\boldsymbol{SR\Sigma^k_{3,0}}$

Bandulasiri and Patrangenaru (2005) introduced the Schoenberg embedding of reflection shapes in higher dimensions to perform an extrinsic analysis. The *Schoenberg embedding* of the size-and-reflection shape manifold is $J : SR\Sigma^k_{3,0} \to S(k, \mathbb{R})$, given by

$$J([\xi]_{RS}) = \xi^T \xi. \tag{9.16}$$

The range of the Schoenberg embedding of $SR\Sigma^k_{3,0}$ is the subset $SM_{k,3}$ of $k \times k$ positive semidefinite symmetric matrices A with $\ rank(A) = 3,\ A\mathbf{1}_k = 0$. Also M_k is the space of $k \times k$ symmetric matrices A with $A\mathbf{1}_k = 0$. Dryden et al. (2008) and Bandulasiri et al. (2009) showed that if the map ϕ from M_k to $S(k-1, \mathbb{R})$, given by $\phi(A) = HAH^T$, where $(1_k, H^T) \in O(k)$, is an isometry, then $\psi : S\Sigma^k_{3,0} \to S(k-1, \mathbb{R})$, given by

$$\psi([\mathbf{x}]_{RS}) = H\xi^{\mathbf{T}}\xi H^T, \tag{9.17}$$

is an embedding; the Schoenberg embedding and the embedding ψ induce the same distance on $SR\Sigma^k_{3,0}$.

9.5.3 Schoenberg extrinsic mean on $SR\Sigma_{3,0}^{k}$

Let $\mathbf{X}$ be a random k-ad in general position, which is centered as $\mathbf{X_0} = (X^1 - \overline{X}, \ldots, X^k - \overline{X}) \in (\mathbb{R}^3)^k \simeq M(3, k; \mathbb{R})$.

Theorem 9.5.1 *(Bandulasiri et al. 2009) Assume $C = \sum_{i=1}^{k} \lambda_i e_i e_i^T$ is the spectral decomposition of $C = E(\mathbf{X_0}\mathbf{X_0}^T)$, and $v_j = \sqrt{\lambda_j} e_j, j = 1, \ldots, k$. Obviously, $C1_k = 0, C \geq 0$. Let $\xi = V^T$, where*

$$V = (v_1 v_2 v_3). \tag{9.18}$$

Then, the extrinsic mean μ_J size-and-reflection shape exists if $\lambda_3 > \lambda_4$ and $\mu_J = [\xi]_{RS}$.

Furthermore, if $k = 4$, then the projection P_ψ is the identity map, and any distribution Q is ψ-nonfocal and $\psi(\mu_S)$ is the mean μ of $\psi(Q)$. The approach taken in Theorem 9.5.1 is the same as saying that, given C, ξ is a classical solution in $\mathbb{R}^3$ to the MDS problem, as given in Mardia et al. (1979) in terms of the three largest eigenvalues of C.

For estimation purposes, let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a sample of k-ads in general position in $\mathbb{R}^3$, where $\mathbf{x}_j = (x_j^1, \ldots, x_j^k)$, for $j = 1, \ldots, n$. The *extrinsic sample mean size-and-reflection shape* is $\overline{[\mathbf{x}]}_E = [\hat{\xi}]_{RS}$, where $\hat{\xi}$ is given by the eigenvectors corresponding to the three largest eigenvectors of

$$\hat{C} = \frac{1}{n} \sum_{j=1}^{n} \xi_j^T \, \xi_j \tag{9.19}$$

assuming that $\hat{\lambda}_3 > \hat{\lambda}_4$, where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_k$ are the eigenvalues of $\hat{C}$. $\hat{\xi}$ is the classical solution in $\mathbb{R}^3$ to the MDS problem (Mardia et al. 1979, p. 397) for the matrix $\hat{C}$. Note that ξ_j is the matrix obtained from $\mathbf{x}_j$ after centering (removing translation). If $\lambda_3 > \lambda_4$, then $\mu_J = [\mu]_{RS}$, and $[\hat{\mu}_{\mathrm{MDS}}]_{RS}$ (see Bandulasiri et al. 2009) is a consistent estimator of $[\mu]_{RS}$. The asymptotic distribution of the extrinsic sample mean size-and-reflection shape is given in a study by Bandulasiri et al. (2009).

Related results are given by Dryden et al. (2008) and Kent (1994).

9.6 3D size-and-reflection shape analysis of the human skull

Here, we give a comprehensive application of size-and-reflection shape space $SR\Sigma_{3,0}^{k}$ of k-ads in general position in 3D. One potential application is to surgery planning, where a natural approach is to take into account size in addition to shape when analyzing the CT scan data. In this context, one performs a nonparametric analysis on the 3D data retrieved from CT scans of adults, on the size-and-reflection shape space $SR\Sigma_{3,0}^{k}$ of k-ads in general position in 3D.

9.6.1 Confidence regions for 3D mean size-and-reflection shape landmark configurations

Once we obtained the 3D reconstruction of the virtual skull from the bone extractions, we proceed to perform landmark-based analysis based on the Schoenberg embedding. For the purpose of one analysis, we were interested in $k = 9$ and $k = 17$ matched landmarks around the eyes. The landmarks were registered on the reconstructed 3D virtual skulls.

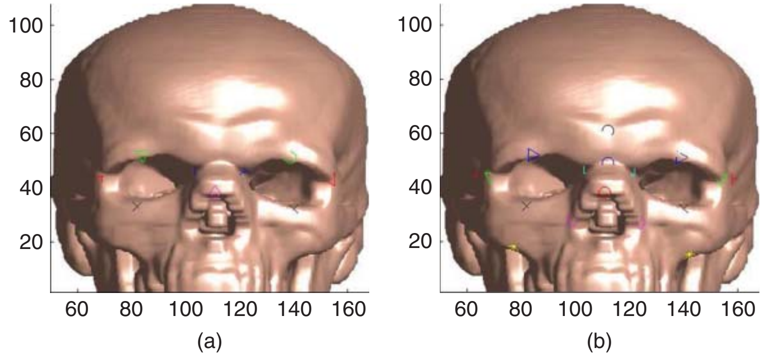


Figure 9.8 Two groups of landmarks around the eye: (a) $k = 17$ and (b) $k = 9$.

Here, we consider nonparametric statistical analysis size-and-reflection shape data using landmarks in which each observation $\mathbf{x} = (x^1, \ldots, x^9)$ and $\mathbf{x} = (x^1, \ldots, x^{17})$ consists of 9 points and 17 points in $\mathbb{R}^3$ (see Figure 9.8). The landmark coordinates can be found in Appendix A and Appendix B of Osborne (2012), pp. 74–77 and 78–84, respectively.

We remove translation by centering the k-ads $\mathbf{x} = (x^1, \ldots, x^9)$ and $\mathbf{x} = (x^1, \ldots, x^{17})$ to

$$\xi = (\xi^1, \ldots, \xi^9) \text{ and } \xi = (\xi^1, \ldots, \xi^{17})$$
$$\xi^j = x^j - \overline{x}, \forall\, j = 1, \ldots, 9 \text{ and } j = 1, \ldots, 17.$$

The set of these centered k-ads lies in the vector subspace $L^3_9 \in (\mathbb{R}^3)^9$ and $L^3_{17} \in (\mathbb{R}^3)^{17}$, respectively. The dimensions of the manifolds $SR\Sigma^9_{3,0}$ and $SR\Sigma^{17}_{3,0}$ are $3k - 6$, where $k = 9$, respectively $k = 17$.

Finally, in order to estimate the 3D size-and-reflection shape for the selected group of landmarks, we compute the Schoenberg sample means. That is, we used 500 bootstrap resamples based on the original 20 skull configurations ($k = 9$ and $k = 17$), represented by the 3 by k matrices (where k was the number of landmarks selected in the analysis). For the purpose of one analysis, we were interested in $k = 9$ and $k = 17$ landmarks around the eye region. Registered representations, for these mean size-and-reflection shapes yield the bootstrap mean size-and-reflection shape configurations given in Figures 9.9 and 9.10.

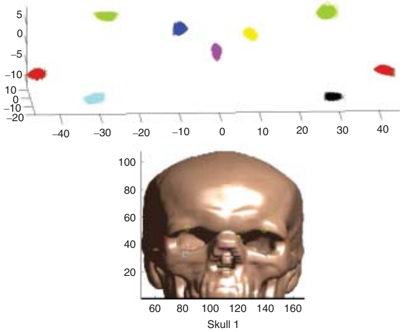


Figure 9.9 Bootstrap distribution for the Schoenberg sample mean configurations $k = 9$ based on 500 resamples.

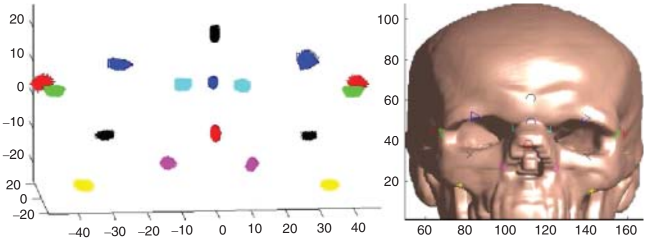


Figure 9.10 Bootstrap distribution for the Schoenberg sample mean configurations $k = 17$ based on 500 resamples.

Table 9.1 90% Lower confidence limit for the bootstrap distribution of the 3D sample mean size-and-reflection shape configuration.

Landmark	1	2	3	4	5	6	7	8	9
x	−45.76	−28.65	−9.75	−32.06	−0.90	7.84	27.15	40.79	26.28
y	10.10	−2.37	−5.91	0.27	−19.20	−3.84	0.86	10.04	−0.06
z	−0.19	9.73	4.24	−11.67	−8.06	3.00	−12.70	−1.70	9.36

Table 9.2 90% Upper confidence limit for the bootstrap distribution of the 3D sample mean size-and-reflection shape configuration.

Landmark	1	2	3	4	5	6	7	8	9
x	−42.03	−24.44	−7.27	−28.29	0.41	10.19	30.65	44.52	30.30
y	11.85	−0.93	−3.39	1.82	−15.92	−1.69	2.52	12.83	4.17
z	1.43	11.12	5.93	−10.02	−5.73	4.83	−11.25	−0.32	13.08

In addition, we provide a 90% simultaneous confidence limits for the 3D mean size-and-reflection shape configuration are given in Tables 9.1 and 9.2. Similar tables, with 90% simultaneous confidence bounds for the 3D mean size-and-reflection shape configuration of 17 landmarks given in Figure 9.8, based on the nonparametric bootstrap distribution displayed in Figure 9.10, are given in Osborne (2012). For practical purposes, these simultaneous confidence regions may be used, for example, to design helmets or other protection devices of the midface area region of an average individual.

9.7 DTI data analysis

In this section, we analyze the DTI data according to the new methodology presented in Osborne et al. (2013) using a concrete DTI example. The data was collected from two groups of children, a group of six children with normal reading abilities and a group of six children with a diagnosis of dyslexia. Twelve spatially registered diffusion MRIs (DT images) were obtained from the two groups of children, respectively. The prognosis is generally helpful for individuals whose dyslexia is identified early, who have supportive family and friends and a strong self-image, and who are involved in a proper remediation program.

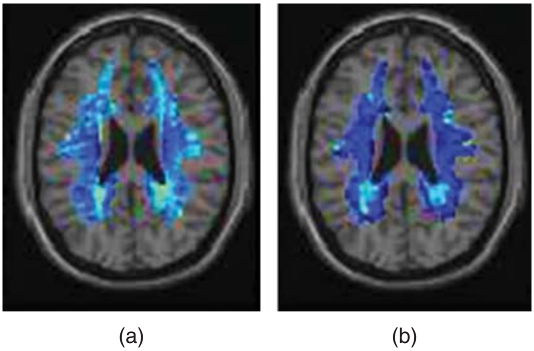


Figure 9.11 DTI slice images of a control subject (a) and of a dyslexia subject (b) (*Source:* Osborne et al. (2013), Figure 1, p. 171. Reproduced by permission of Elsevier).

In Figure 9.11, we display DTI slices including a given voxel recorded in a control subject and a dyslexia subject.

Commonly in DTI group studies, a typical statistical problem is to find regions of the brain whose anatomical characteristics differ between two groups of subjects. Typically, the analysis consists of registering the DT images to a common template so that each voxel corresponds to the same anatomical structure in all the images and then applying two-sample tests at each voxel.

Osborne et al. (2013) presented a nonparametric analysis of a single voxel at the intersection of the corpus callosum and corona radiata in the frontal left hemisphere that was found in Schwartzman et al. (2008) to exhibit the strongest difference between the two groups. Table 1 in Osborne et al. (2013) shows the data at this voxel for all 12 subjects. The d_{ij} in the table are the entries of the DT on and above the diagonal (the below-diagonal entries would be same since the DTs are symmetric).

For this analysis, the primary goal is to demonstrate that the nonparametric two-sample testing procedure in Section 3 of Osborne et al. (2013) is able to detect a significant difference between the generalized Frobenius means of the clinically normal and dyslexia groups without increasing the dimensionality in the process. For distances, other than Riemannian ones, on the set $Sym^+(3)$ of 3×3 positive definite matrices, see Dryden et al. (2009). Namely, we are interested in detecting, on average, from diffusion tensor images (DTI), dyslexia in young children compared to their clinically normal peers, *without making any distributional assumptions*.

Given two independent populations with i.i.d. samples of random SPD matrices $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1} \in Sym^+(3)$ from the clinically normal population and $X_{2,1}, X_{2,2}, \ldots, X_{2,n_2} \in Sym^+(3)$ from the dyslexia population with sample sizes of $n_1 = 6$ and $n_2 = 6$ and the total sample size $n = n_1 + n_2 = 12$, where, for $a = 1, 2, X_{a,1} \sim \mu_{F,a}$, the sample generalized Frobenius mean for the clinically normal population and dyslexia population is given by

$$\bar{x}_{1,F} = \begin{pmatrix} 0.6318 & 0.0046 & -0.0924 \\ 0.0046 & 0.9863 & -0.0873 \\ -0.0924 & -0.0873 & 0.7803 \end{pmatrix} \text{ and } \bar{x}_{2,F} = \begin{pmatrix} 0.6146 & -0.0261 & -0.1910 \\ -0.0261 & 0.8118 & -0.0901 \\ -0.1910 & -0.0901 & 0.9537 \end{pmatrix}.$$

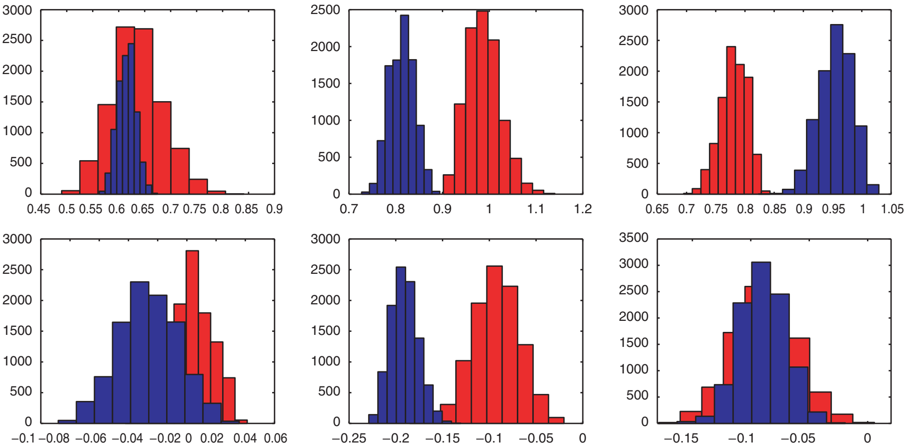


Figure 9.12 Marginals of the bootstrap distribution for the generalized Frobenius sample means for d_{11}, d_{22}, d_{33}, d_{12}, d_{13}, and d_{23}; clinically normal (light gray) versus dyslexia (dark gray).

The test statistics $\hat{T}$ and V, previously described in Osborne et al. (2013), are given by

$$\hat{T} = \begin{pmatrix} 0.9862 & 0.0000 & 0.0000 \\ -0.0485 & 0.9067 & 0.0000 \\ -0.1487 & -0.0152 & 1.0781 \end{pmatrix} \text{and} V = \begin{pmatrix} -0.0139 & 0.0000 & 0.0000 \\ -0.0513 & -0.0980 & 0.0000 \\ -0.1446 & -0.0153 & 0.0752 \end{pmatrix}.$$

In addition, let $\hat{t}_{ij}$ and v_{ij} correspond to the entries of the test statistics $\hat{T}$ and V on and below the diagonal (since the test statistics $\hat{T}$ and V are lower triangular matrices).

In order to test hypothesis 3.9 or hypothesis 3.10 from a study by Osborne et al. (2013), for $\delta = I_3$, we repeatedly resample observations from the original data and compute the generalized Frobenius sample mean for each respective group. The generalized Frobenius sample means are computed as described in Section 2 of a study by Osborne et al. (2013). Figure 9.12 displays a visualization of the bootstrap distributions of the Generalized Frobenius sample means. They used 10,000 bootstrap resamples and computed the bootstrap generalized Frobenius sample mean for each respective group.

In addition, for each bootstrap resample, we calculate the Cholesky decomposition of the bootstrap generalized Frobenius sample mean for each respective group and then proceed to calculate the bootstrap distribution of our test statistics $\hat{T}$ and V as described in Equation (3.16) of Osborne et al. (2013). Figures 9.13 and 9.14 display a visualization of our nonpivotal bootstrap distribution of our test statistics $\hat{T}$ and V.

Under the null hypothesis 3.10 of Osborne et al. (2013), $\delta = I_3$ on $T^+(3, \mathbb{R})$ or $\log(\delta^{-1}) = \mathbf{0}_3$ on the vector space $T(3, \mathbb{R})$ of lower triangular 3×3 matrices; however, after visually examining Figures 9.13 and 9.14, we informally conclude that there is a significant difference between the generalized Frobenius means of the clinically normal and dyslexia group, since the $\hat{T}^*_{22}$ and V^*_{22} values do not overlap with $\delta_{22} = 1$, respectively, and with $\mathbf{0}_{3,22} = 0$. Moreover, we also observed that the distributions of $\hat{T}^*_{33}$, V^*_{33} and $\hat{T}^*_{31}$, V^*_{31} barely touch $\delta_{33} = 1$, $\mathbf{0}_{3,33} = 0$ and $\delta_{31} = 0$, $\mathbf{0}_{3,31} = 0$.

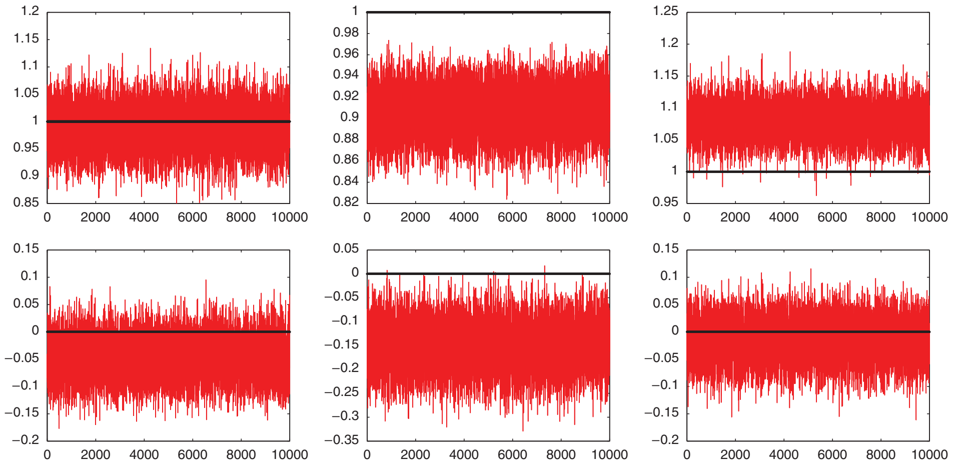


Figure 9.13 Bootstrap distribution of our test statistics $\hat{T}$: The images (1–3) in the first row correspond to the diagonal entries of the matrices $\hat{T}^*$: t_{11}, t_{22}, t_{33} and images (4–6) in the second row corresponds to the lower triangular off-diagonal entries of the matrices $\hat{T}^*$: t_{21}, t_{31}, t_{32} (*Source:* Osborne et al. (2013), Figure 2, p. 172. Reproduced by permission of Elsevier).

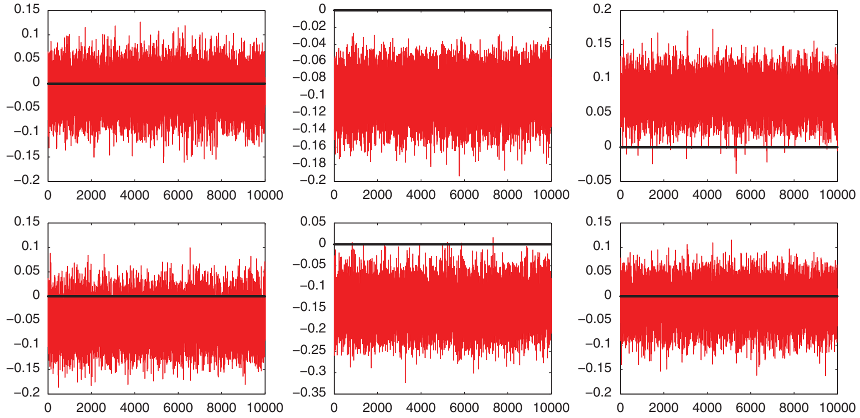


Figure 9.14 Bootstrap distribution of our test statistics V: The images (1–3) in the first row correspond to the diagonal entries of the matrices V^*: v_{11}, v_{22}, v_{33} and images (4–6) in the second row corresponds to the lower off-diagonal entries of the matrices V^*: v_{21}, v_{31}, v_{32}.

These results are formally confirm at level α, that there is significant evidence that the clinically normal and dyslexia children display on average different DTI responses. The results were obtained by constructing a $100(1-\alpha)\%$ – simultaneous bootstrap confidence intervals, as described in Remark 3.8 of Osborne et al. (2013), for $\hat{T}_{ij}$ and V_{ij}. Tables 2 and 3 in Osborne et al. (2013) display the results of the Bonferroni $100(1-\alpha)\%$ – simultaneous bootstrap confidence intervals for $\hat{T}_{ij}$ and V_{ij} at various significance levels: for example, the

94% simultaneous c.i. for T_{22} and T_{33} are $(0.8488, 0.9600)$ respectively, $(1.0085, 1.1465)$, and the simultaneous c.i. 94% for V_{22} and V_{33} are $(-0.1640 - 0.0409)$ respectively $(0.0084, 0.1367)$, both pointing to a significant mean difference between the two groups of children.

9.8 MRI data analysis of corpus callosum image

Albert Einstein's brain was removed shortly after his death (most likely without prior family consent), weighed, dissected, and photographed by a pathologist. Among other pictures, a digital scan of a picture of General Relativity creator's half brain taken at the autopsy is displayed subsequently. The *corpus callosum* (CC) connects the two cerebral hemispheres and facilitates interhemispheric communication. It is the largest white matter structure in the brain. We extracted the contour of the CC from this Einstein's brain image, the shape of which would be set at the center of a null hypothesis in our testing problem (see Figure 9.15).

Fletcher (2013) extracted contours of CC midsagittal sections from MRI images, to study possible age-related changes in this part of the human brain. His study points out certain age-related shape changes in the corpus callosum. Given that Einstein passed away at 76, we consider a subsample of corpus callosum brain contours from Fletcher (2013), in the age group 64–83, to test how far is the average CC contour from Einstein's. The data is displayed in Figure 9.16.

We consider contours, boundaries of 2D topological disks in the plane. To keep the data analysis stable, and to assign a *unique* labeling, we make the *generic* assumption that across the population there is a unique anatomical or geometrical landmark starting point p_0 on such a contour of perimeter one, so that the label of any other point p on the contour is the "counterclockwise" travel time at constant speed from p_0 to p. A *regular contour* $\tilde{\gamma}$ is regarded as the range of a piecewise differentiable *regular* arclength parameterized function $\gamma : [0, L] \to \mathbb{C}, \gamma(0) = \gamma(L)$, that is one-to-one on $[0, L)$. Two contours $\tilde{\gamma}_1, \tilde{\gamma}_2$ *have the same direct similarity shape* if there is a direct similarity $S : \mathbb{C} \to \mathbb{C}$, such that $S(\tilde{\gamma}_1) = \tilde{\gamma}_2$.

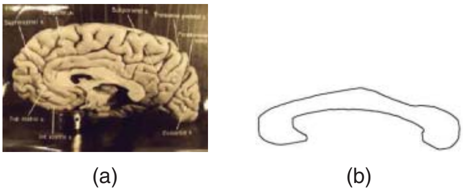


Figure 9.15 Right hemisphere of Einstein's brain including CC midsagittal section (a) and its contour (b).

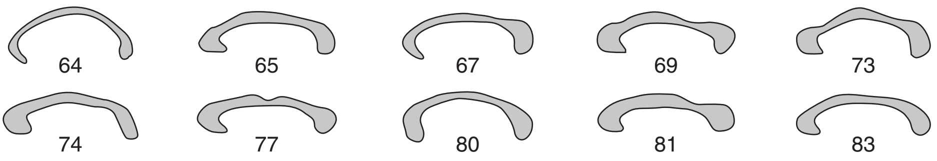
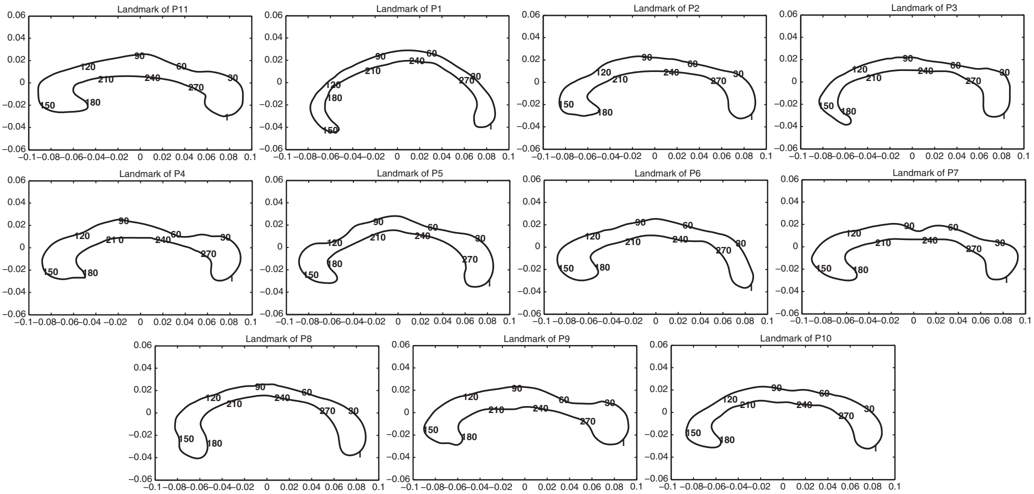


Figure 9.16 Corpus callosum midsagittal sections shape data, in subjects ages – 64–83.



Landmark of P11

Landmark of P1

Landmark of P2

Landmark of P3

Landmark of P4

Landmark of P5

Landmark of P6

Landmark of P7

Landmark of P8

Landmark of P9

Landmark of P10

Figure 9.17 Matched sampling points on midsagittal sections in for CC data (Einstein's is the upper left CC).

Two regular contours $\tilde{\gamma}_1, \tilde{\gamma}_2$ have the same similarity shape if their centered counterparts satisfy $\tilde{\gamma}_{2,0} = \lambda\tilde{\gamma}_{1,0}$, for some $\lambda \in \mathbb{C}\backslash 0$. Therefore Σ_2^{reg}, *set of all direct similarity shapes of regular contours,* is a dense and open subset of $P(\mathbf{H})$, the projective space corresponding to the Hilbert space $\mathbf{H}$ of all square integrable centered functions from S^1 to $\mathbb{C}$ (see Ellingson et al. 2013).

We will use the neighborhood hypothesis testing method on the manifold of planar contours to test if the average shape of the CC in a population of 64- to 83-year-old people is close to the shape of Einstein's CC in the sense of Ellingson et al. (2013). Data in Figure 9.16 was used to test the hypothesis that the mean CC shape is in a small ball of radius δ around the shape of Einstein's CC (see Qiu et al. 2014). Note that Fletcher (2013), from which we borrowed the MRI data, tacitly assumes that the similarity shape is preserved during the data acquisition. Likewise, frontal pinhole camera images of a planar scene are similarity preserving (see Mardia and Patrangenaru 2005); therefore, comparing similarity shapes from data collected using these two methods makes sense.

The closest representatives of the VW sample mean of the shapes of contours of the CC midsections compared to the shape of Einstein's CC midsection are displayed in Figure 9.17. The overlaps of the two contours are rare, which visually shows that the average CC contour shape is significantly different from Einstein's. The 95% bootstrap confidence region for the extrinsic mean CC contour (Figure 9.18), based on a conveniently selected icons, is given in Figure 9.19.

We set δ as the radius of the null hypothesis ball around Einstein's CC contour shape, as a point p_0 on $P(\mathbf{H})$. The maximum value for δ where the test is significant was found to be 0.1367, which is quite large taking into account the fact that the diameter of any finite dimensional complex projective space with the VW metric is $\sqrt{2}$. The result is explained by the fact that Einstein's brain halves had more interconnections than in an average 64- to 83-year-old individual. This is reflected in the thicker shape appearance of his CC midsection; when this shapes are regarded as points on the shape space $P(\mathbf{H})$, p_0 is a remote outlier of the cluster of shapes of CCs in the data because these are thinner.

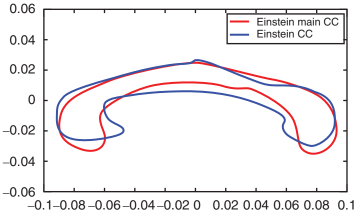


Figure 9.18 Registered icons for 2D direct similarity shapes of CC midsections : sample mean (light gray) versus Albert Einstein's (dark gray).

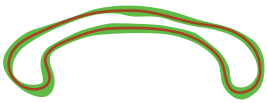


Figure 9.19 95% bootstrap confidence region for the extrinsic mean CC contour by 1000 resamples.

Acknowledgments

The authors acknowledge support from grants NSA-MSP- H98230-14-1-0135 and NSF-DMS-1106935. They are grateful to the referee for helpful comments that significantly improved the manuscript, and to E. R. Reisch (LSU) for providing the CT data.

References

Appleton B and Talbot H 2006 Globally minimal surfaces by continuous maxial flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(1), 106–118.

Bandulasiri A, Bhattacharya R and Patrangenaru V 2009 Nonparametric inference for extrinsic means on size-and-(reflection)-shape manifolds with applications in medical imaging. *Journal of Multivariate Analysis* **100**, 1867–1882.

Bhattacharya R and Patrangenaru V 2003 Large sample theory of intrinsic and extrinsic sample means on manifolds-part I. *Annals of Statistics* **31**(1), 1–29.

Bandulasiri A and Patrangenaru V 2005 Algorithms for nonparametric inference on shape manifolds *Proceedings of the Joint Statistical Meetings 2005*, pp. 1617–1622.

Bhattacharya R and Patrangenaru V 2005 Large sample theory of intrinsic and extrinsic sample means on manifolds-part II. *Annals of Statistics* **33**(3), 1211–1245.

Bookstein F 1991 *Morphometric Tools for Landmark Data, Geometry and Biology*. Cambridge University Press, Cambridge.

Bresson X, Esedoglu S, Vandergheynst P, Thiran J and Osher S 2005 Global minimizers of the active contour/snake model *FBP: Theory and Applications*.

Burgoyne C, Thompson HW, Mercante DE and Amin R 2000 Basic issues in the sensitive and specific detection of optic nerve head surface change within longitudinal LDT TOPSS images In *The Shape of Glaucoma, Quantitative Neural Imaging Techniques* Lemji HG and Schuman JS (eds), Kugler Publications, The Hague, The Netherlands, pp. 1–37.

Cabra LB, Cam N and Foran J 1994 Accelerated volume rendering and tomographic reconstruction using texture mapping hardware *Proceedings of ACM/IEEE Symposium on Volume Visualization*, pp. 91–98.

Caselles V, Kimmel R and Sapiro G 1997 Geodesic active contours. *International Journal of Computer Vision* **22**(1), 61–79.

Crane M and Patrangenaru V 2011 Random change on a Lie group and mean glaucomatous projective shape change detection from stereo pair images. *Journal of Multivariate Analysis* **102**, 225–237.

Derado G, Mardia K, Patrangenaru V and Thompson HW 2004 A shape based glaucoma index for tomographic images. *Journal of Applied Statistics* **31**, 1241–1248, http://www.tandfonline.com/doi/abs/10.1080/0266476042000285486.

Drebin R, Carpenter L and Harrahan P 1988 Volume rendering *SIGGRAPH '88*, pp. 665–674.

Dryden IL, Koloydenko A and Zhou D 2009 Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics* **3**(3), 1102–1123.

Dryden IL, Kume A, Lee H and Wood ATA 2008 A multi-dimensional scaling approach to shape analysis. *Biometrika* **95**(4), 779–798.

Dryden I and Mardia K 1998 *Statistical Shape Analysis*. John Wiley & Sons, Ltd: Chichester.

Efron B 1982 *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia, PA.

Ellingson L, Patrangenaru V and Ruymgaart F 2013 Nonparametric estimation of means on Hilbert manifolds and extrinsic analysis of mean shapes of contours. *Journal of Multivariate Analysis* **122**, 317–333.

Faugeras O 1992 What can be seen in three dimensions with an uncalibrated stereo rig? *Proceedings of European Conference on Computer Vision, LNCS 588*, pp. 563–578.

Fletcher TP 2013 Geodesic regression and the theory of least squares on riemannian manifolds. *International Journal of Computer Vision* **105**, 171–185.

Fréchet M 1948 Les elements aleatoires de nature quelconque dans un espace distancie (in French). *Annales de l'Institute Henri Poincare* **10**, 215–310.

Gibson S, Fyock C, Grimson E, Kanade T, Kikinis R, Lauer H, McKenzie N, Mor A, Nakajima S, Ohkami H, Osborne R, Samosky J and Sawada A 1998 Volumetric object modeling for surgical simulation. *Medical Image Analysis* **2**(2), 121–132.

Harris L and Camp J 1984 Display and analysis of tomographic volumetric images utilizing a vari-focal mirror. *Proceedings of SPIE* **507**, 38–45.

Hartley RI, Gupta R and Chang T 1992 Stereo from uncalibrated cameras *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761–764.

Hartley R and Zisserman A 2004 *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press.

Hawker MJ, Vernon S, Tattersall CL and Dua HS 2007 Linear regression modeling of rim area to discriminate between normal and glaucomatous optic nerve heads: the bridlington eye assessment project. *Glaucoma* **16**, 345–351.

Heffernan P and Robb R 1985 A new method of shaded surfaced displayed of biological and medical images. *IEEE Transactions on Medical Imaging* **4**, 26–38.

Herman G and Liu H 1977 Display of three-dimensional information in computed tomography. *Journal of Computer Assisted Tomography* **1**, 155–160.

Kass M, Witkin A and Terzopoulos D 1988 Snakes: active contour models. *International Journal of Computer Vision* **1**(4), 321–331.

Kaufman A, Cohen D and Yagel R 1993 Volume graphics. *IEEE Computer* **26**(7), 51–64.

Kendall D 1984 Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society* **16**, 81–121.

Kendall D, Barden D, Carne T and Le H 1999 *Shape and Shape Theory*. John Wiley & Sons, Inc., New York.

Kent JT 1994 The complex Bingham distribution. *Journal of the Royal Statistical Society Series B* **56**, 285–299.

Kent JT and Mardia KV 2012 A geometric approach to projective shape and the cross ratio. *Biometrika* **99**, 833–849.

Levoy M 1988 Display of surfaces from volume data. *IEEE Computer Graphics and Applications* **8**(5), 29–37.

Lorensen W and Cline H 1987 Matching cubes: a high-resolution 3-D surface construction algorithm. *Computer Graphics* **21**(3), 163–169.

Ma Y, Soatto S, Košecká S and Sastry SS 2006 *An Invitation to 3-D Vision*. Springer-Verlag.

Mardia KV, Kent JT and Bibby JM 1979 *Multivariate Analysis*. Academic Press.

Mardia KV and Patrangenaru V 2005 Directions and projective shapes. *Annals of Statistics* **33**, 1666–1699.

Milnor J 1963 *Morse Theory*. Princeton University Press.

Osborne DE 2012 Nonparametric data analysis on manifolds with applications in medical imaging. *Electronic Theses, Treatises and Dissertations*. Paper 5085. http://diginole.lib.fsu.edu/etd/5085.

Osborne D, Patrangenaru V, Ellingson L, Groisser D and Schwartzman A 2013 Nonparametric two-sample tests on homogeneous Riemannian manifolds, Cholesky decompositions and diffusion tensor image analysis. *Journal of Multivariate Analysis* **119**, 163–175.

Patrangenaru V, Liu X and Sugathadasa S 2010 Nonparametric 3D projective shape estimation from pairs of 2d images - I, in Memory of W.P. Dayawansa. *Journal of Multivariate Analysis* **101**, 11–31.

Patrangenaru V, Thompson H and Derado G 2000 Large sample and bootstrap methods on for 3D shape change with applications to detection of glaucomatous change in images of the optic nerve head *Abstracts of the Leeds Annual Statistics Research Workshop in honor of the 65th birthday of Professor K.V. Mardia*, http://www.maths.leeds.ac.uk/Statistics/workshop/leeds2000, pp. 30–33.

Pflesser B, Tiede U and Hohne K 1998 Specification, modelling and visualization of arbitrarily shaped cut surfaces in the volume model. *Medical Image Computing and Computer-Assisted Intervention-MICCAI* **1496**, 853–860.

Qiu M, Patrangenaru V and Ellingson L 2014 How far is the corpus callosum of an average individual from Albert Einstein's? *Proceedings of COMPSTAT-2014, The 21st International Conference on Computational Statistics*, Geneva, pp. 403–410.

Robb R 1994 *Three-Dimensional Biomedical Imaging-Principle and Practice*. VCH Publishers, New York.

Robb R 1996 Visualization methods for analysis of multimodality images In *Functional Neroimaging: Technical Foundations* (ed. Thatcher R, Hallett M, Zeffiro T, John E and Huerta M) Academics Press, San Diego, CA.

Robb R and Barillot C 1989 Interactive display and analysis of 3-d medical images. *IEEE Transactions on Medical Imaging* **8**(3), 217–226.

Sanfilippo PG, Cardini A, Hewitt AW, Crowston JG and Mackey DA 2009 Optic disc morphology–rethinking shape. *Progress in Retinal and Eye Research* **28**(4), 227–248.

Schwartzman A, Dougherty RF and Taylor JE 2008 False discovery rate analysis of brain diffusion direction maps. *Annals of Applied Statistics* **2**, 153–175.

Serra L, Nowinsk W, Poston T, Hern N, Meng L, Guan C and Pillay P 1997 The brain bench: virtual tools for stereotactic frame neurosurgery. *Medical Image Analysis* **1**(4), 317–329.

Sher L 1977 Graphics in space: see it now. *Proceedings of NCGA* **3**, 101–160.

Sughatadasa SM 2006 Affine and projective shape analysis with applications. *Ph.D. Dissertation*, Texas Tech University.

Unger M, Pock T, Trobin W, Cremers D and Bischof H 2008 Tvseg - interactive total variation based image segmentation. *Proceedings of the British Machine Vision Conference (BMVC)*.

Ziezold H 1977 On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. *Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 8th European Meeting of Statisticians*, Reidel, Dordrecht, pp. 591–602.

10

Some families of distributions on higher shape spaces

Yasuko Chikuse[1] **and Peter E. Jupp**[2]

[1] *Faculty of Engineering, Kagawa University, Takamatsu, Kagawa, Japan*

[2] *School of Mathematics and Statistics, University of St Andrews, St Andrews, UK*

10.1 Introduction

This chapter aims to enlarge the repertoire of useful distributions on the spaces, Σ_m^k, of shapes of ordered sets of k landmarks in $\mathbb{R}^m$ with $m > 2$. It is a pleasure to include it in a volume dedicated to Kanti Mardia, since he has been so influential in the development of shape analysis.

The shape of an object is usually understood as the geometrical information that remains when allowance is made for changes in location, scale and orientation. One standard construction of the shape space Σ_m^k is as follows. Every set of k (not totally coincident) labelled points $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in $\mathbb{R}^m$ can be centred and scaled to give a *pre-shape*

$$\mathbf{Z} = \left\{\mathrm{tr}\,(\mathbf{X}\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{X}^{\mathrm{T}})\right\}^{-1/2}\ \mathbf{X}\mathbf{H}^{\mathrm{T}},$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ and $\mathbf{H}$ is the $(k-1) \times k$ Helmert matrix, that is, the matrix having jth row

$$(\underbrace{h_j, \ldots, h_j}_{j}, -jh_j, 0, \ldots, 0), \qquad \text{with } h_j = -\left\{j(j+1)\right\}^{-1/2}$$

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

for $j = 1, \ldots, k-1$. Then $\mathbf{Z}$ is an $m \times (k-1)$ matrix which satisfies

$$\mathrm{tr}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}) = 1.$$

The set of such pre-shapes $\mathbf{Z}$ forms the *pre-shape space*, $\mathcal{S}_m^k$.

The *(similarity) shape space,* Σ_m^k, corresponding to sets of k labelled points in $\mathbb{R}^m$ is obtained from $\mathcal{S}_m^k$ by removing the effect of rotations. Thus

$$\Sigma_m^k = \mathcal{S}_m^k / SO(m),$$

where $SO(m)$ acts on $\mathcal{S}_m^k$ on the left by

$$\mathbf{Z} \mapsto \mathbf{U}\mathbf{Z} \qquad \mathbf{U} \in SO(m). \tag{10.1}$$

For a pre-shape represented by an $m \times (k-1)$ matrix $\mathbf{Z}$ in $\mathcal{S}_m^k$, the corresponding shape in Σ_m^k will be denoted by $[\mathbf{Z}]$. The shape spaces Σ_1^k and Σ_2^k can be identified with the sphere S^{k-2} and the complex projective space $\mathbb{C}P^{k-2}$, respectively; see Dryden and Mardia (1998, Section 4.1.9) or Kendall et al. (1999, Section 1.4). The uniform distribution on Σ_m^k is the distribution of $[\mathbf{Z}]$ when $\mathbf{Z}$ has the uniform distribution on $\mathcal{S}_m^k$. In this chapter, all probability densities of distributions on Σ_m^k are with respect to the uniform distribution. It is useful to identify distributions of $[\mathbf{Z}]$ on Σ_m^k with distributions of $\mathbf{Z}$ on $\mathcal{S}_m^k$ that are invariant under the action (10.1) of $SO(m)$.

The *reflection shape space*, $R\Sigma_m^k$, is obtained from $\mathcal{S}_m^k$ by removing the effect of rotations and reflections. Thus, $R\Sigma_m^k = \mathcal{S}_m^k / O(m)$, where $O(m)$ acts on $\mathcal{S}_m^k$ on the left by $\mathbf{Z} \mapsto \mathbf{U}\mathbf{Z}$ for $\mathbf{U} \in O(m)$, generalising (10.1). The function $\mathbf{Z} \mapsto \mathbf{Z}^T\mathbf{Z}$ on $\mathcal{S}_m^k$ provides an embedding of $R\Sigma_m^k$ in the space of symmetric $(k-1) \times (k-1)$ matrices having rank r with $1 \leq r \leq m$; see Chikuse and Jupp (2004) or Dryden et al. (2008). Distributions on $R\Sigma_m^k$ can be identified with distributions on $\mathcal{S}_m^k$ that are invariant under the action of $O(m)$.

10.1.1 Distributions on shape spaces

The main parametric families of distributions that have been used on shape spaces fall into the following groups:

(i) *offset shape distributions*, in which the observed landmarks $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in $\mathbb{R}^m$ are obtained by subjecting fixed ideal landmarks to appropriate random perturbations; see Dryden and Mardia (1991), Goodall and Mardia (1991, 1992, 1993), Kendall (1984), Mardia and Dryden (1989a, 1989b); Dryden and Mardia (1998, Section 6.6);

(ii) distributions on Σ_2^k that rely on the identification of Σ_2^k with $\mathbb{C}P^{k-2}$ obtained by identifying each real $2 \times (k-1)$ matrix $\mathbf{Z}$ in $\mathcal{S}_2^k$ with a unit vector $\mathbf{z}$ in $\mathbb{C}^{k-1}$. These distributions include

(a) the *complex Bingham* distributions of Kent (1994), having densities

$$f([\mathbf{z}]; \mathbf{A}) = c(\mathbf{A}) \exp\{\mathbf{z}^* \mathbf{A} \mathbf{z}\}, \tag{10.2}$$

where $\mathbf{A}$ is a $(k-1) \times (k-1)$ Hermitian matrix;

(b) the *complex Watson* distributions of the form (10.2) with $\mathbf{A}$ of rank one, so that the densities have the form

$$f([\mathbf{z}];\kappa,[\boldsymbol{\mu}]) = c(\kappa)\exp\{\kappa|\mathbf{z}^*\boldsymbol{\mu}|^2\},$$

where κ is a scalar and $\boldsymbol{\mu}$ is a unit vector in $\mathbb{C}^{k-1}$;

(c) the *complex Bingham quartic* distributions of Kent et al. (2006), having densities

$$f([\mathbf{z}];\mathbf{A},\mathbf{B},\boldsymbol{\mu}) = c(\mathbf{A},\mathbf{B},\boldsymbol{\mu})\exp\{\mathbf{z}^*\mathbf{A}\mathbf{z} + \Re\left((\mathbf{z}^*\boldsymbol{\mu})^2\mathbf{z}^{\mathrm{T}}\mathbf{B}\mathbf{z}\right)\}, \qquad (10.3)$$

where $\mathbf{A}$ and $\mathbf{B}$ are $(k-1)\times(k-1)$ complex matrices with $\mathbf{A}$ negative-definite Hermitian and $\mathbf{B}$ symmetric, and $\boldsymbol{\mu}$ is a unit vector in $\mathbb{C}^{k-1}$ with $\mathbf{A}\boldsymbol{\mu} = \mathbf{B}\boldsymbol{\mu} = \mathbf{0}$;

(d) the *complex angular central Gaussian* distributions of Kent (1994; 1997) on Σ_2^k, having densities

$$f([\mathbf{z}];\mathbf{A}) = |\mathbf{A}|\left(\mathbf{z}^*\mathbf{A}\mathbf{z}\right)^{-(k-1)}, \qquad (10.4)$$

where $\mathbf{A}$ is a positive-definite $(k-1)\times(k-1)$ Hermitian matrix;

(e) the rotationally symmetric distributions on Σ_2^k, having densities

$$f([\mathbf{z}];[\mathbf{w}],\kappa) = c_\phi(\kappa)^{-1}\exp\{-\kappa\phi(1-|z^*\mathbf{w}|^2)\},$$

where $\mathbf{w}$ is a unit vector in $\mathbb{C}^{k-1}$, κ is real and ϕ is a positive function; see Dryden and Mardia (1998, Section 6.5);

(iii) the *shape Bingham* distributions of Chikuse and Jupp (2004) on Σ_m^k, having densities

$$f([\mathbf{Z}];\mathbf{A}) = {}_1F_1\left(1/2; m(k-1)/2; \mathbf{A}\otimes\mathbf{I}_m\right)^{-1}\exp\{\mathrm{tr}(\mathbf{A}\mathbf{Z}^{\mathrm{T}}\mathbf{Z})\}, \qquad (10.5)$$

where $\mathbf{A}$ is a symmetric real $(k-1)\times(k-1)$ matrix. For $m=2$, the distributions (10.5) are the complex Bingham distributions (10.2). For $m>2$, Chikuse and Jupp (2004) showed that the distributions (10.5) are the Bingham distributions on $\mathcal{S}_m^k$ that are $O(m)$-invariant. The distributions with densities (10.5) are invariant under reflection, and so can be regarded as distributions on $R\Sigma_m^k$.

Thus, for $m>2$, only a few families of distributions on Σ_m^k have been explored. Of these, the shape Bingham distributions (10.5) have the disadvantage that they are concentrated near the shapes of collinear landmarks.

In Section 10.2, we introduce various families of distributions on Σ_m^k that are generalisations of the complex angular central Gaussian distributions (10.4). The shape Bingham distributions and the distributions of Section 10.2 are invariant under reflection, so in Section 10.3 we modify them to obtain shape distributions that allow departures from such symmetry. Section 10.4 proposes a test of symmetry under reflection.

10.2 Shape distributions of angular central Gaussian type

Hints on possible ways of generalising the complex angular central Gaussian distributions (10.4) from Σ_2^k to Σ_m^k come from the facts that (a) they are complex versions of the angular central Gaussian distributions of Tyler (1987) on real projective spaces $\mathbb{R}P^{p-1}$, having densities

$$f(\pm\mathbf{z};\mathbf{A}) = |\mathbf{A}|^{1/2}(\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z})^{-p/2} \qquad \mathbf{z} \in S^{p-1}, \tag{10.6}$$

where $\mathbf{A}$ is a $p \times p$ positive-definite matrix, (b) the densities (10.6) have been generalised by Chikuse (1990) to the matrix angular central Gaussian distributions on Grassmann manifolds, having densities

$$f(\mathbf{X}\mathbf{X}^{\mathrm{T}};\mathbf{A}) = |\mathbf{A}|^{r/2}|\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X}|^{-p/2} \qquad \mathbf{X} \in V_r(\mathbb{R}^p), \tag{10.7}$$

where $\mathbf{X}$ is a $p \times r$ matrix satisfying $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}_r$ and so representing an orthogonal r-frame in $\mathbb{R}^p$, that is, an element of the Stiefel manifold $V_r(\mathbb{R}^p)$, while $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ is the matrix representation of (the orthogonal projection onto) the subspace spanned by this frame, an element of the corresponding Grassmann manifold.

In this section, we present three families of distributions on Σ_m^k that are analogous to (10.6) and (10.7). Their densities are proportional to $|\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}|^{-a}$, $\{|\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}|/|\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|\}^{-a}$ and $\{\mathrm{tr}(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}})\}^{-a}$, respectively, where $\mathbf{A}$ is a $(k-1)\times(k-1)$ positive-definite matrix. Details are given in Sections 10.2.1–10.2.3, respectively.

10.2.1 Determinantal shape ACG distributions

The densities (10.7) of the matrix angular central Gaussian distributions suggest the family of *determinantal shape ACG distributions* on Σ_m^k, which have probability density functions of the form

$$f([\mathbf{Z}];\mathbf{A},a) = c(a,k,m)\left\{{}_2F_1(a, m/2; (k-1)/2; \mathbf{I}_{k-1} - \mathbf{A})\right\}^{-1}|\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}|^{-a}, \tag{10.8}$$

where $\mathbf{A}$ is a $(k-1)\times(k-1)$ positive-definite matrix and

$$c(a,k,m)^{-1} = \frac{\pi^{(k-1)m/2}}{\Gamma_m((k-1)/2)}\int_{\mathbf{T}>0,\mathrm{tr}\mathbf{T}=1}|\mathbf{T}|^{(k-m-2)/2-a}d\mathbf{T}, \tag{10.9}$$

the integral being over positive-definite $m \times m$ matrices $\mathbf{T}$ with $\mathrm{tr}\,\mathbf{T} = 1$. For $a < (k-m)/2$,

$$c(a,k,m)^{-1} = \frac{\pi^{(k-1)m/2}\Gamma_m((k-1)/2 - a)}{\Gamma_m((k-1)/2)\Gamma(m[(k-1)/2-a]+1)},$$

Γ_m being the multivariate Gamma function given by $\Gamma_m(t) = \pi^{m(m-1)/4}\prod_{i=1}^m \Gamma(t-(i-1)/2)$. Without loss of generality, we can assume that

$$\mathrm{tr}\,\mathbf{A} = 1. \tag{10.10}$$

If (10.10) holds, then the eigenvalues of $\mathbf{A}$ lie in (0, 1), so that ${}_2F_1(a, m/2; (k-1)/2; \mathbf{I}_{k-1} - \mathbf{A})$ is defined. The densities (10.8) have the equivariance property

$$f([\mathbf{Z}\mathbf{V}]; \mathbf{V}^{\mathrm{T}}\mathbf{A}\mathbf{V}, a) = f([\mathbf{Z}]; \mathbf{A}, a) \quad \mathbf{V} \in O(k-1), \tag{10.11}$$

and so form a composite transformation model under the right actions

$$[\mathbf{Z}] \mapsto [\mathbf{Z}\mathbf{V}] \quad \mathbf{V} \in O(k-1) \tag{10.12}$$

$$\mathbf{A} \mapsto \mathbf{V}^{\mathrm{T}}\mathbf{A}\mathbf{V} \quad \mathbf{V} \in O(k-1) \tag{10.13}$$

of $O(k-1)$ on Σ_m^k and the space of symmetric $(k-1) \times (k-1)$ matrices.

In the case $k = m + 1$, (10.8) becomes

$$f([\mathbf{Z}]; \mathbf{A}, a) = c(a)|\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|^{-a},$$

whatever the value of $\mathbf{A}$, where

$$c(a)^{-1} = \frac{\pi^{m^2/2}}{\Gamma_m(m/2)} \int_{\mathbf{T}>0,\, \mathrm{tr}\mathbf{T}=1} |\mathbf{T}|^{-1/2-a} d\mathbf{T}.$$

For $a < 1/2$,

$$c(a)^{-1} = \frac{\pi^{m^2/2}\Gamma_m(m/2 - a)}{\Gamma_m(m/2)\Gamma(m[m/2 - a] + 1)}.$$

If $a > 0$ then the density (10.8) is infinite at singular shapes, that is, those for which the landmarks lie in a proper affine subspace of $\mathbb{R}^m$, and so $\mathrm{rk}\,\mathbf{Z} < m$. Thus, the distribution is appropriate for modelling shapes that are clustered near a singular shape. A calculation based on the polar decomposition of $\mathbf{Z}$ and the spectral decomposition of $\mathbf{A}$ shows that if $a < 0$ then the density has a mode at $[\mathbf{Z}]$, where $m\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \Pi_{\mathbf{A},+}$ is the projection matrix of $\mathbb{R}^{k-1}$ onto the subspace spanned by the m dominant unit eigenvectors of $\mathbf{A}$.

Remark

Multiplying the determinantal ACG shape densities (10.8) by the shape Bingham densities (10.5) gives the model with densities

$$f([\mathbf{Z}]; \mathbf{A}, \mathbf{B}) = c(\mathbf{A}, \mathbf{B}, a)|\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}|^{-a} \exp\left\{\mathrm{tr}(\mathbf{Z}\mathbf{B}\mathbf{Z}^{\mathrm{T}})\right\}. \tag{10.14}$$

Some lengthy manipulation shows that

$$\begin{aligned} c(\mathbf{A}, \mathbf{B}, a)^{-1} = {} & \frac{\pi^{(k-1)m/2}}{\Gamma_m((k-1)/2)} \sum_{\ell_1=0}^{\infty} \sum_{\ell_2=0}^{\infty} \sum_{\lambda_1 \vdash \ell_1} \sum_{\lambda_1 \vdash \ell_2} \frac{(a)_{\lambda_1}}{\ell_1!\ell_2! C_{\lambda_2}(\mathbf{I}_m)} \\ & \times \left[\int_{\mathbf{T}>0, \mathrm{tr}\mathbf{T}=1} |\mathbf{T}|^{(k-m-2)/2-a} C_{\lambda_2}(\mathbf{T})\, d\mathbf{T} \right] \\ & \times \sum_{\phi \in \lambda_1 \cdot \lambda_2} \frac{(m/2)_\phi}{((k-1)/2)_\phi} \frac{C_\phi^{\lambda_1\lambda_2}(\mathbf{I}_m, \mathbf{I}_m)}{C_\phi(\mathbf{I}_m)} C_\phi^{\lambda_1\lambda_2}(\mathbf{I}_{k-1} - \mathbf{A}, \mathbf{B}), \end{aligned}$$

where the polynomials $C_\phi^{\lambda_1\lambda_2}$ of two matrix arguments are defined in e.g. Chikuse (2003, Appendix A.3). Because of the complexity of the normalising constant $c(\mathbf{A}, \mathbf{B}, a)$, we do not consider the distributions (10.14) any further here.

10.2.2 Modified determinantal shape ACG distributions

The singularity in the densities (10.8) (for $a > 0$) can be removed by replacing $|\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}|$ by $|\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}|/|\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|$. This yields the family of *modified determinantal shape ACG distributions* on Σ_m^k, which have probability density functions of the form

$$f([\mathbf{Z}]; \mathbf{A}, a) = c(\mathbf{A}, a) \left(\frac{|\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}|}{|\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|} \right)^{-a} \qquad \mathbf{A} > 0, \tag{10.15}$$

where

$$\begin{aligned} c(\mathbf{A}, a)^{-1} = & \frac{\pi^{(k-1)m/2}}{\Gamma_m((k-1)/2)} \int_{\mathbf{T}>0, \mathrm{tr}\mathbf{T}=1} |\mathbf{T}|^{(k-m-2)/2} d\mathbf{T} \\ & \times {}_2F_1(a, m/2; (k-1)/2; \mathbf{I}_{k-1} - \mathbf{A}). \end{aligned}$$

Without loss of generality, we can assume that (10.10) holds, and so $c(\mathbf{A}, a)$ is defined. The densities (10.15) have the equivariance property (10.11) and so form a composite transformation model under the actions (10.12) and (10.13) of $O(k-1)$. They are also invariant under reflection, and so can be regarded as densities on $R\Sigma_m^k$.

When $a = (k-1)/2$, (10.15) reduces to

$$f([\mathbf{Z}]; \mathbf{A}, (k-1)/2) = c_1(k, m) |\mathbf{A}|^{m/2} \left(\frac{|\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}|}{|\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|} \right)^{-(k-1)/2},$$

where

$$\begin{aligned} c_1(k, m)^{-1} &= \frac{\pi^{(k-1)m/2}}{\Gamma_m((k-1)/2)} \int_{\mathbf{T}>0, \mathrm{tr}\mathbf{T}=1} |\mathbf{T}|^{(k-m-2)/2} d\mathbf{T} \\ &= \pi^{(k-1)m/2} / \Gamma((k-1)m/2 + 1). \end{aligned}$$

In the case $k = m+1$, the density (10.15) is constant, i.e. the distribution is uniform, for all $\mathbf{A}$ and a.

A calculation shows that for $a > 0$ the density (10.15) has a mode at $[\mathbf{Z}]$, where $m\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \mathbf{\Pi}_{\mathbf{A},-}$, whereas for $a < 0$, the mode is at $[\mathbf{Z}]$, where $m\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \mathbf{\Pi}_{\mathbf{A},+}$. Here, $\mathbf{\Pi}_{\mathbf{A},-}$ and $\mathbf{\Pi}_{\mathbf{A},+}$ are the projection matrices of $\mathbb{R}^{k-1}$ onto the subspaces spanned by the m dominant unit eigenvectors of $\mathbf{A}^{-1}$ and $\mathbf{A}$, respectively.

In the case $a = (k-1)/2$, the maximum likelihood estimate $\hat{\mathbf{A}}$ of $\mathbf{A}$ based on observed shapes $[\mathbf{Z}_1], \ldots, [\mathbf{Z}_n]$ satisfies

$$\hat{\mathbf{A}}^{-1} = \frac{k-1}{m} \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i^{\mathrm{T}} \left(\mathbf{Z}_i \hat{\mathbf{A}} \mathbf{Z}_i^{\mathrm{T}} \right)^{-1} \mathbf{Z}_i,$$

which is similar to the equation for maximum likelihood estimation in the angular central Gaussian distributions; see Tyler (1987).

10.2.3 Tracial shape ACG distributions

An alternative method of producing distributions on Σ_m^k is to exploit (a) the identification of $SO(m)$-invariant distributions on $\mathcal{S}_m^k$ with distributions on Σ_m^k that identifies the distribution of $\mathbf{Z}$ in $\mathcal{S}_m^k$ with that of $[\mathbf{Z}]$ in Σ_m^k, (b) the identification of $\mathbf{Z}$ in $\mathcal{S}_m^k$ with $\mathbf{z} = \operatorname{vec}\mathbf{Z}$, where $\operatorname{vec}\mathbf{Z}$ is the vector obtained by writing the columns of $\mathbf{Z}$ above one another. The form (10.6) of the angular central Gaussian densities on $\mathbb{R}P^{p-1}$ suggests the use of densities on $\mathcal{S}_m^k$ that are proportional to $(\mathbf{z}^{\mathrm{T}}\mathbf{B}\mathbf{z})^{-a}$. Such a density is $SO(m)$-invariant precisely when $\mathbf{B} = \mathbf{A} \otimes \mathbf{I}_m$, so that the density of $[\mathbf{Z}]$ on Σ_m^k is

$$f([\mathbf{Z}]; \mathbf{A}, a) = c(\mathbf{A}, a)\{\operatorname{tr}(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}})\}^{-a} \qquad \mathbf{A} > 0. \tag{10.16}$$

These are the *tracial shape ACG distributions*. In the case $a = m(k-1)/2$, $\mathbf{z}$ has an angular central Gaussian distribution, and density (10.16) takes the form

$$f([\mathbf{Z}]; \mathbf{A}, m(k-1)/2) = |\mathbf{A}|^{m/2}\{\operatorname{tr}(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}})\}^{-m(k-1)/2}. \tag{10.17}$$

If $a \neq 0$, then identifiability of $\mathbf{A}$ in (10.16) can be ensured by condition (10.10). The densities (10.16) have the equivariance property (10.11) and so the family (10.16) is a composite transformation model under the actions (10.12)–(10.13) of $O(k-1)$. They are also invariant under reflection, and so can be regarded as densities on $R\Sigma_m^k$.

In the case $m = 2$, the family (10.17) consists of complex angular central Gaussian distributions (10.4) for which the Hermitian parameter matrix $\mathbf{A}$ is real.

The density (10.16) has a mode at $[\mathbf{Z}]$, where $m\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \mathbf{\Pi}_{\mathbf{A},-}$ if $a > 0$ but $m\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \mathbf{\Pi}_{\mathbf{A},+}$ if $a < 0$.

Both diffuse and concentrated tracial shape ACG distributions can be approximated by shape Bingham distributions. For a near 0 or $\mathbf{A} = \mathbf{I}_{k-1} + \mathbf{B}$ with $\mathbf{B}$ near $\mathbf{0}$,

$$\begin{aligned}\{\operatorname{tr}(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}})\}^{-a} &= \{1 + \operatorname{tr}(\mathbf{Z}\mathbf{B}\mathbf{Z}^{\mathrm{T}})\}^{-a} \\ &\simeq 1 - a\operatorname{tr}(\mathbf{Z}\mathbf{B}\mathbf{Z}^{\mathrm{T}}) \\ &\simeq \exp\{-a\operatorname{tr}(\mathbf{Z}\mathbf{B}\mathbf{Z}^{\mathrm{T}})\},\end{aligned}$$

and so densities of the form (10.16) with $\mathbf{A}$ almost a multiple of $\mathbf{I}_{k-1}$ can be approximated by densities of the form (10.5). On the other hand, for $\mathbf{A} = \mathbf{V}\operatorname{diag}(\kappa_1, \ldots, \kappa_{k-1})\mathbf{V}^{\mathrm{T}}$ with $\mathbf{V}$ in $O(k-1)$ and $\kappa_1 \geq \cdots \geq \kappa_{k-1} > 0$,

$$\operatorname{tr}(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}) = \kappa_1\left\{1 - \sum_{i=2}^{k-1}\left(1 - \frac{\kappa_i}{\kappa_1}\right)\|\mathbf{y}_i\|^2\right\},$$

where $\mathbf{Z}\mathbf{V} = (\mathbf{y}_1, \ldots, \mathbf{y}_{k-1})$. If κ_1/κ_2 is large, then with high probability $\|\mathbf{y}_2\|^2, \ldots, \|\mathbf{y}_{k-1}\|^2$ are small and so

$$\begin{aligned}\{\operatorname{tr}(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}})\}^{-a} &\simeq \kappa_1^{-a}\left\{1 + a\sum_{i=2}^{k-1}\left(1 - \frac{\kappa_i}{\kappa_1}\right)\|\mathbf{y}_i\|^2\right\} \\ &\simeq \kappa_1^{-a}\exp\left\{a\sum_{i=2}^{k-1}\left(1 - \frac{\kappa_i}{\kappa_1}\right)\|\mathbf{y}_i\|^2\right\} \\ &= \kappa_1^{-a}e^{a}\exp\{-(a/\kappa_1)\operatorname{tr}(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}})\}.\end{aligned}$$

Thus, concentrated densities of the form (10.16) can be approximated by densities of the form (10.5).

If $a > 0$, then the tracial shape ACG distributions (10.16), like the shape Bingham distributions (10.5), have a mode (which is unique if $\kappa_1 > \kappa_2$) at the 'collinear' shape $[\boldsymbol{\theta}\boldsymbol{\gamma}_1^{\mathrm{T}}]$, where $\boldsymbol{\gamma}_1$ is a dominant unit eigenvector of $\mathbf{A}$ and $\boldsymbol{\theta}$ is any unit vector in $\mathbb{R}^m$.

In the case $a = m(k-1)/2$, the maximum likelihood estimate $\hat{\mathbf{A}}$ of $\mathbf{A}$ based on observations $[\mathbf{Z}_1], \ldots, [\mathbf{Z}_n]$ can be obtained from

$$\hat{\mathbf{A}}^{-1} = \frac{k-1}{n} \sum_{i=1}^{n} \left\{ \mathrm{tr}\left(\mathbf{Z}_i \hat{\mathbf{A}} \mathbf{Z}_i^{\mathrm{T}} \right) \right\}^{-1} \mathbf{Z}_i^{\mathrm{T}} \mathbf{Z}_i,$$

which is reminiscent of the equation for maximum likelihood estimation in the angular central Gaussian distributions; see Tyler (1987).

10.3 Distributions without reflective symmetry

The shape Bingham distributions (10.5) and the shape ACG distributions (10.8), (10.15) and (10.16) are invariant under reflection, and so can be regarded as distributions on the reflection shape space, $R\Sigma_m^k$. Thus, they can be inappropriate for modelling in contexts in which the distinction between a shape and its reflection is important. In this section, we introduce and explore some distributions on Σ_m^k that need not have such symmetries. Our construction is to alter reflection-invariant densities by multiplying them by suitable non-invariant functions. This process of modulation of symmetric densities is inspired by the modulation of centrally symmetric densities on $\mathbb{R}^d$ described in Azzalini (2014, Section 1.2). The modulating functions that we consider exploit the fact that interchanging two columns of a determinant changes its sign. In Section, 10.3.1 the modulating functions are exponential functions of determinants involving pre-shapes; in Section 10.3.2, the modulating functions are linear functions of determinants.

10.3.1 Volume Fisher–Bingham distributions

Multiplying the shape Bingham density (10.5) by a modulating function of the form

$$\exp\left\{ \sum_{\alpha} b_\alpha |\mathbf{Z}_\alpha| \right\}$$

yields the density

$$f([\mathbf{Z}]; \mathbf{A}, \mathbf{B}) = c(\mathbf{A}, \mathbf{B}) \exp\left\{ \mathrm{tr}(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}) + \sum_{\alpha} b_\alpha |\mathbf{Z}_\alpha| \right\}, \qquad (10.18)$$

where $\mathbf{A}$ is a symmetric $(k-1) \times (k-1)$ parameter matrix, $\mathbf{B}$ is a set of skew-symmetric m-dimensional $(k-1) \times \cdots \times (k-1)$ arrays with entries b_α, the multi-index α runs through all $(j_1 \ldots j_m)$ with $1 \leq j_1 < \cdots < j_m \leq k-1$, and $|\mathbf{Z}_\alpha|$ is the determinant $|(\mathbf{z}_{j_1}, \ldots, \mathbf{z}_{j_m})|$, where $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{k-1})$. We may assume that $\mathbf{A}$ satisfies

$$\mathrm{tr}\, \mathbf{A} = 0.$$

The parameter $\mathbf{B}$ measures the asymmetry of (10.18) under reflection of $\mathbb{R}^m$. The model (10.18) is (a) a full exponential model of dimension $(k-2)(k+1)/2 + \binom{k-1}{m}$ with canonical statistic $[\mathbf{Z}] \mapsto (\mathbf{Z}^{\mathrm{T}}\mathbf{Z}, \{|\mathbf{Z}_\alpha|\})$, (b) a transformation model under action (10.12) and $(\mathbf{A}, \{b_\alpha\}) \mapsto (\mathbf{V}^{\mathrm{T}}\mathbf{A}\mathbf{V}, \{b_\alpha\})$. In the case $m = 2$, the family (10.18) is the family of complex Bingham distributions (10.2).

To obtain a more manageable model than (10.18), we consider the submodel in which the densities have the form

$$f([\mathbf{Z}]; \mathbf{A}, [\mathbf{M}], \kappa) = c(\mathbf{A}, [\mathbf{M}], \kappa) \exp\left\{\mathrm{tr}\,(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}) + \kappa\,|\mathbf{M}\mathbf{Z}^{\mathrm{T}}|\right\}, \tag{10.19}$$

where $\kappa \geq 0$ and $[\mathbf{M}]$ is in Σ_m^k. The parameter κ in (10.19) is a measure of asymmetry. When $\kappa = 0$, the densities (10.19) reduce to the shape Bingham densities (10.5). The submodel of (10.19) with $\kappa > 0$ and M of rank m has dimension $m(2k - m - 1)/2$. Because the determinant of a $m \times m$ matrix is the signed volume of the m-dimensional parallelepiped generated by its column vectors, we call the distributions with densities (10.19) 'volume Fisher–Bingham distributions' and those in the submodels with $\mathbf{A} = \mathbf{0}$ 'volume Fisher distributions'.

If $\mathbf{A} = \mathbf{0}$ then (10.19) has a mode at $[\mathbf{M}]$. As $\kappa \to \infty$, the distribution of $[\mathbf{Z}]$ becomes concentrated near $[\mathbf{M}]$.

Remarks

(i) If $k = m + 1$ then

$$c(\mathbf{A}, [\mathbf{M}], \kappa)^{-1} = \frac{\pi^{m^2/2}}{\Gamma_m(m/2)} \int_{\mathbf{T}>0, \mathrm{tr}\mathbf{T}=1} |\mathbf{T}|^{-1/2} \cosh\{\kappa|\mathbf{M}||\mathbf{T}|^{1/2}\} \times {}_0F_0^{(k-1)}(\mathbf{A}, \mathbf{T})d\mathbf{T},$$

where ${}_0F_0^{(k-1)}$ is a polynomial in two matrix arguments defined in for example, (A.6.6) of Chikuse (2003). In particular,

$$c(\mathbf{0}, [\mathbf{M}], \kappa)^{-1} = \frac{\pi^{m^2/2}}{\Gamma_m(m/2)} \int_{\mathbf{T}>0, \mathrm{tr}\mathbf{T}=1} |\mathbf{T}|^{-1/2} \cosh\{\kappa|\mathbf{M}||\mathbf{T}|^{1/2}\} d\mathbf{T}.$$

(ii) In the case $k = m + 2$, (10.18) reduces to

$$f([\mathbf{Z}]; \mathbf{A}, \mathbf{b}) = \exp\left\{\mathrm{tr}\,(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}) + |(\mathbf{b}, \mathbf{Z}^{\mathrm{T}})| - \kappa(\mathbf{A}, \mathbf{b})\right\},$$

where $\mathbf{b}$ is a parameter vector in $\mathbb{R}^{k-1}$. In this case, the models (10.18) and (10.19) have dimensions $(m^2 + 5m + 2)/2$ and $m(m + 3)/2$, respectively.

(iii) It might be interesting to explore models in which $|\mathbf{Z}_\alpha|$ in (10.18) is replaced by other symmetric functions of the eigen-values of $\mathbf{Z}_\alpha$.

10.3.2 Cardioid-type distributions

Multiplying the shape Bingham density (10.5) or the determinantal, modified determinantal, or tracial ACG shape densities (10.8), (10.15) or (10.16) by a modulating function of the form $1 + \kappa|\mathbf{M}\mathbf{Z}^T|$, where $\kappa \geq 0$ is a scalar and $[\mathbf{M}]$ is in Σ_m^k, gives probability density functions of the form

$$f([\mathbf{Z}]; \mathbf{A}, \kappa, [\mathbf{M}]) = {}_1F_1\left(1/2; m(k-1)/2; \mathbf{A} \otimes \mathbf{I}_m\right)^{-1} \times \exp\{\mathrm{tr}(\mathbf{A}\mathbf{Z}^{\mathrm{T}}\mathbf{Z})\}\left\{1 + \kappa|\mathbf{M}\mathbf{Z}^T|\right\}, \tag{10.20}$$

$$f([\mathbf{Z}]; \mathbf{A}, a, \kappa, [\mathbf{M}]) = c(a, k, m)\left\{{}_2F_1(a, m/2; (k-1)/2; \mathbf{I}_{k-1} - \mathbf{A})\right\}^{-1} \times |\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}|^{-a}\left\{1 + \kappa|\mathbf{M}\mathbf{Z}^T|\right\}, \tag{10.21}$$

$$f([\mathbf{Z}]; \mathbf{A}, a, \kappa, [\mathbf{M}]) = c(\mathbf{A}, a)\left[\frac{|\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}}|}{|\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|}\right]^{-a}\left\{1 + \kappa|\ M\mathbf{Z}^T|\right\}, \tag{10.22}$$

or

$$f([\mathbf{Z}]; \mathbf{A}, a, \kappa, [\mathbf{M}]) = c(\mathbf{A}, a)\{\mathrm{tr}(\mathbf{Z}\mathbf{A}\mathbf{Z}^{\mathrm{T}})\}^{-a}\left\{1 + \kappa|\mathbf{M}\mathbf{Z}^T|\right\}, \tag{10.23}$$

where $c(a, k, m)$ is given by (10.9) and $c(\mathbf{A}, a)$ is an appropriate normalising constant. Models (10.20)–(10.23) are transformation models under action (10.12) and $(\mathbf{A}, \kappa, \mathbf{M}) \mapsto (\mathbf{V}^{\mathrm{T}}\mathbf{A}\mathbf{V}, \kappa, \mathbf{M}\mathbf{V})$. A necessary and sufficient condition for any of (10.20)–(10.23) to be non-negative is that $\kappa m^{-m/2}\left|\mathbf{M}\mathbf{M}^{\mathrm{T}}\right|^{1/2} \leq 1$.

For $m = 1, k = 3$, the space Σ_m^k can be identified with the circle, and taking $\mathbf{A} = \mathbf{0}$ in densities (10.20)–(10.23) gives the cardioid distributions, having densities $f(\theta; \mu, \kappa) = (2\pi)^{-1}\left\{1 + \kappa\cos(\theta - \mu)\right\}$. It, therefore, seems appropriate to describe general distributions with densities (10.20)–(10.23) as being 'of cardioid type'.

Because pre-multiplication of $\mathbf{Z}$ by a reflection in $O(m)$ leaves densities (10.5), (10.8) and (10.14)–(10.16) unchanged, whereas it changes the sign of $|\mathbf{M}\mathbf{Z}^T|$, the normalising constants in (10.20)–(10.23) are the same as those in (10.5), (10.8) and (10.14)–(10.16). It then follows from the product form of the densities in (10.20)–(10.23) that inference can be carried out separately on the parameters $\mathbf{A}$ and $([\mathbf{M}], \kappa)$. In particular, $\mathbf{A}$ and $([\mathbf{M}], \kappa)$ are orthogonal.

Given a random sample $[\mathbf{Z}_1], \ldots, [\mathbf{Z}_n]$ from any of the models (10.20)–(10.23), κ and $[\mathbf{M}]$ can be estimated as follows. An intuitively reasonable estimator of $[\mathbf{M}]$ is $[\hat{\mathbf{M}}]$, where the columns of $\hat{\mathbf{M}}^{\mathrm{T}}$ are the largest m principal components of $\sum_{i=1}^n \mathbf{Z}_i^{\mathrm{T}}\mathbf{Z}_i$. The (partial) maximum likelihood estimator (given $[\hat{\mathbf{M}}]$), $\hat{\kappa}$, of κ satisfies

$$0 = \sum_{i=1}^n \frac{|\hat{\mathbf{M}}\mathbf{Z}_i^{\mathrm{T}}|}{1 + \hat{\kappa}|\hat{\mathbf{M}}\mathbf{Z}_i^{\mathrm{T}}|}$$

and is unique.

10.4 A test of reflective symmetry

A natural hypothesis on a distribution on Σ_m^k is that it is invariant under reflection of shapes. The following simple randomisation test is based on (signed) volumes of simplices formed from suitable subsets of $m + 1$ landmarks. Given a sample $[\mathbf{Z}_1], \ldots, [\mathbf{Z}_n]$ of shapes in Σ_m^k, it is intuitively reasonable to reject the null hypothesis of invariance under reflection if the value of

$$T = \sum_{\alpha} \left(\sum_{i=1}^{n} |\mathbf{Z}_{i,\alpha}| \right)^2 ,$$

is large. Here, α runs through all $(j_1 \dots j_m)$ with $1 \leq j_1 < \cdots < j_m \leq k-1$, and $|\mathbf{Z}_{i,\alpha}|$ is the determinant $|(\mathbf{z}_{i,j_1}, \dots, \mathbf{z}_{i,j_m})|$, where $\mathbf{Z}_i = (\mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,k-1})$. Significance of the value of T can be assessed by comparing it with its randomisation distribution, in which $|\mathbf{Z}_{i,\alpha}|$ is replaced by $\varepsilon_i |\mathbf{Z}_{i,\alpha}|$ and $(\varepsilon_1, \dots, \varepsilon_n)$ has the uniform distribution on $\{-1, 1\}^n$. The randomisation distribution can be enumerated for small n or simulated for large n. This test is consistent against all alternatives to reflective symmetry in models (10.18), (10.19) and (10.20)–(10.23).

10.5 Appendix: derivation of normalising constants

The normalising constants of distributions (10.8), (10.14) and (10.15) can be derived using the polar decomposition $\mathbf{Z}^{\mathrm{T}} = \mathbf{H}\mathbf{T}^{1/2}$ of $\mathbf{Z}$ (which has rank m with probability 1), in which $\mathbf{H} \in V_m(\mathbb{R}^{k-1})$ and $\mathbf{T} > 0$. Calculations based on (A.6.4), (A.6.6) and (A.2.7) of Chikuse (2003) show that $|\mathbf{H}^{\mathrm{T}}\mathbf{A}\mathbf{H}|^{-a} = {}_1F_0(a, (\mathbf{I}_m - \mathbf{H}^{\mathrm{T}}\mathbf{A}\mathbf{H}))$ and that

$$\begin{aligned} \int_{\mathbf{H} \in V_m(\mathbb{R}^{k-1})} \{|\mathbf{H}^{\mathrm{T}}\mathbf{A}\mathbf{H}|\}^{-a} \, d\mathbf{H} &= {}_1F_0^{(k-1)}(a; \mathbf{I}_{k-1} - \mathbf{A}, \mathbf{I}_m) \\ &= {}_2F_1(a, m/2; (k-1)/2; \mathbf{I}_{k-1} - \mathbf{A}), \end{aligned}$$

which leads to the normalising constant of (10.8). The normalising constants of the other distributions can be obtained similarly.

References

Azzalini A 2014 *The Skew-Normal and Related Families*, *IMS Monographs Series*. Cambridge University Press.

Chikuse Y 1990 The matrix angular central Gaussian distribution. *Journal of Multivariate Analysis* **33**, 265–274.

Chikuse Y 2003 *Statistics on Special Manifolds*, *Lecture Notes in Statistics*, vol. 174. Springer-Verlag.

Chikuse Y and Jupp P 2004 A test of uniformity on shape spaces. *Journal of Multivariate Analysis* **88**, 163–176.

Dryden I, Kume A, Le H and Wood A 2008 A multi-dimensional scaling approach to shape analysis. *Biometrika* **95**, 779–798.

Dryden I and Mardia K 1991 General shape distributions in a plane. *Advances in Applied Probability* **23**, 259–276.

Dryden I and Mardia K 1998 *Statistical Shape Analysis*. John Wiley & Sons.

Goodall C and Mardia K 1991 A geometrical derivation of the shape density. *Advances in Applied Probability* **23**, 496–514.

Goodall C and Mardia K 1992 The non-central Bartlett decompositions and shape densities. *Journal of Multivariate Analysis* **40**, 94–108.

Goodall C and Mardia K 1993 Multivariate aspects of shape theory. *Annals of Statistics* **21**, 848–866.

Kendall D 1984 Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society* **16**, 81–121.

Kendall D, Barden D, Carne T and Le H 1999 *Shape and Shape Theory*. John Wiley & Sons.

Kent J 1994 The complex Bingham distribution and shape analysis. *Journal of the Royal Statistical Society, Series B* **56**, 285–299.

Kent J 1997 Data analysis for shapes and images. *Journal of Statistical Planning and Inference* **57**, 181–193.

Kent J, Mardia K and McDonnell P 2006 The complex Bingham quartic distribution and shape analysis. *Journal of the Royal Statistical Society, Series B* **68**, 747–765.

Mardia K and Dryden I 1989a Shape distributions for landmark data. *Advances in Applied Probability* **21**, 742–755.

Mardia K and Dryden I 1989b The statistical analysis of shape data. *Biometrika* **76**, 271–281.

Tyler D 1987 Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika* **74**, 579–589.

11

Elastic registration and shape analysis of functional objects

Zhengwu Zhang, Qian Xie, and Anuj Srivastava
Department of Statistics, Florida State University, Tallahassee, FL, USA

11.1 Introduction

Professor Kanti Mardia and his colleagues have led the advancement of ideas and tools in the field of statistical shape analysis of objects for more than two decades. This progress has been triggered by a confluence of tools from geometry, statistics, computing, and imaging, and has continued in several interesting directions. One area that has seen an increasing focus is the joint solution to registration and shape comparison problems. Traditionally, shape analysis has been performed on finite point sets that have been labeled or registered, that is, one is given a correspondence between points across the sets (Dryden and Mardia 1998; Mardia and Dryden 1989). However, in many real applications, especially those involving image data, this correspondence may not be available. Thus, one has to solve for the registration problem as a part of shape analysis. While some early efforts took a sequential approach, where one registers the objects first and then uses this registration in subsequent shape analysis, it quickly became clear that a more comprehensive joint solution is needed. Thus, the simultaneous registration and shape analysis of objects became an important goal in shape analysis. In this chapter, we summarize advances in *elastic shape analysis*, a class of Riemannian solutions that provide a metric-based framework for registration of points while using the same metric for shape comparisons.

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

The objects of interest in shape analysis can vary according to applications. While shapes of planar, closed contours are of prime interest in image analysis and computer vision, where objects' boundaries help us classify objects and their motions, there is also interest in other types of objects. Some problems require analyzing shapes of curves in more than two dimensions. An example is protein structure analysis where one studies shapes of protein backbones (Liu et al. 2010, 2011), as curves in $\mathbb{R}^3$. Another object of interest is the shape of surfaces as embeddings of spheres or discs in $\mathbb{R}^3$ (Kurtek et al. 2012a). This is useful, for instance, in medical imaging where one studies shapes of anatomical structures for diagnosing medical conditions. Shape analysis of surfaces that form boundaries of 3D objects has also found interest in computer graphics, 3D printing, and visualization. There are also problem areas that do not directly involve shapes but where the ideas and methods derived from shape considerations can contribute significantly. An example is the problem of alignment of real-valued functions, the so-called phase-amplitude separation in functional data analysis (FDA) in Tucker et al. (2013, 2012), that has benefited from metrics and procedures developed initially for shape analysis of curves. Such alignment problems also arise in image registration where a *metric-based* approach offers significant advantages in Xie et al. (2012). The extensions of shape analysis of Euclidean curves have also led to formal studies for comparisons and modeling of trajectories on Riemannian manifolds in Su et al. (2014).

11.1.1 From discrete to continuous and elastic

As mentioned earlier, a large majority of past statistical analyses of shapes use discrete, point-set representations, while the more recent trend is to study continuous objects. Since continuous objects, such as parameterized curves and surfaces, are represented by coordinate functions and functional spaces are typically infinite-dimensional, this change introduces an additional complexity of infinite dimensionality. So, the question arises: Are we making the problem unnecessarily complicated by using functional representations? Let us study the options more carefully. Say we are given two sets, each set contains a finite number of unregistered points, and our goal is to register them and to compare their shapes. Now the problem of registration is a combinatorial one and adds considerable computational complexity to the solution. On the other hand, let us assume that the original objects are parameterized curves: $t \mapsto (f_1(t), f_2(t))$, for $t \in D$ where D is an appropriate domain. The interesting part in this approach is the following. For each t, the pair of points, $f_1(t)$ and $f_2(t)$ are considered registered. In order to change the registration, one simply has to re-parameterize one of the objects. In other words, find a re-parameterization γ of f_2 such that $f_1(t)$ is now registered to $f_2(\gamma(t))$. Thus, we can find optimal registration (or alignment) of curves by optimizing over the variable γ under a proper objective function. If this objective function is a metric that is invariant of all shape-preserving transformations, then we simultaneously achieve a joint solution for registration and shape comparison. Thus, parameterization controls registration between curves and an optimal registration can be found using algorithms with complexity much smaller than those encountered in combinatorial solutions. Similar arguments can be made for higher dimensional parameterized objects, such as surfaces and images, as well. The optimization over parameterization variability in shape analysis of objects, under a metric with proper invariance properties, leads to a framework called *elastic shape analysis*. In this chapter, we summarize the progress in elastic shape analysis of different types of continuous objects and point out some fundamental issues – theoretical and computational – in these areas.

11.1.2 General elastic framework

Here onward we focus exclusively on parameterized objects – functions, curves, surfaces, images, trajectories – and use parameterizations to control registrations.

For different types of objects, the choice of mathematical representations and domains will be different. In the case of FDA and shape analysis of curves, the domain of interest is $D = [0, 1]$; for analyzing shapes of surfaces, it is $D = \mathbb{S}^2$, and for performing registration of 2D images, it is $D = [0, 1]^2$. The re-parameterization is chosen to be a direction- and boundary-preserving diffeomorphism from D to itself, and Γ is the set of all such diffeomorphisms. For instance, in the case of FDA, Γ is the set of all positive diffeomorphisms of [0,1] such that $\gamma(0) = 0$ and $\gamma(1) = 1$. Similarly, for shape analysis of surfaces Γ includes all orientation-preserving diffeomorphisms of $\mathbb{S}^2$ to itself. An interesting property of Γ is that it forms a group action under composition, with the identity element given by the function $\gamma_{id}(t) = t$. Therefore, for any two γ_1, γ_2, the composition $\gamma_1 \circ \gamma_2$ is also a valid re-parameterization, and so is the inverse γ^{-1} for any γ.

The next issue is to decide the objective function so that an optimal re-parameterization can be found in a variational framework. A seemingly natural idea of performing alignment using the criterion $\inf_\gamma \|f_1 - f_2 \circ \gamma\|$, where $\|\cdot\|$ denotes the $\mathbb{L}^2$ norm, turns out to be problematic. The main issue is that it allows degeneracy, that is, one can reduce this cost arbitrarily close to zero even when the two functions may be quite different. This is commonly referred to as the *pinching problem* in Ramsay and Silverman (2005). Pinching implies that a severely distorted γ is used to eliminate (or minimize) those parts of f_2 that do not match with f_1; this can be done even when f_2 is mostly different from f_1. Another way to state the problem is that one can easily manipulate $\|f \circ \gamma\|$ into a broad range of values, by choosing an appropriate γ. Of course, one can avoid the pinching problem by imposing a roughness penalty on γ, thus avoiding a severe distortion of γs, but it leads to other issues including asymmetry. A related problem from the registration perspective is that: $\|f_1 - f_2\| \neq \|f_1 \circ \gamma - f_2 \circ \gamma\|$ in general. Why is this problematic? Observe that if we warp two functions by the same γ: earlier $f_1(t)$ matches with $f_2(t)$, and now $f_1(\gamma(t))$ matches with $f_2(\gamma(t))$. Each point-wise registration remains unchanged, but their $\mathbb{L}^2$ norm changes. Hence, the $\mathbb{L}^2$ norm is not a proper objective function to help solve the registration problem.

The solution comes from deriving an elastic-metric based objective function that is better suited for registration and shape comparison. While the discussion of the underlying elastic Riemannian metric is complicated, we directly move on to a simplification that is based on certain square-root transforms of data objects. Denoted by q, these objects take different mathematical forms in different contexts, as explained in later sections. The important mathematical property of these representations is that $\|q_1 - q_2\| = \|(q_1, \gamma) - (q_2, \gamma)\|$, for all γ, where q_is represent the objects f_is and (q_i, γ) represents the re-parameterized object $(f_i \circ \gamma)$. This property allows us to define a solution for all important problems:

$$\inf_\gamma \|q_1 - (q_2, \gamma)\| = \inf_\gamma \|(q_1, \gamma) - q_2\|. \tag{11.1}$$

Not only does the optimal γ help register the object f_2 to f_1, but also the infimum value of the objective function is a proper metric for shape comparison of the two objects. (In the case of shape analysis of curves and surfaces, one needs to perform an additional rotation alignment for shape comparisons.) This metric enables statistical analysis of shapes. One can compute mean shapes and the dominant modes of variations given a set of shape samples, develop statistical models for capturing observed shape variability, and use these models in performing hypothesis tests. While we focus on static shapes in this chapter, these ideas can also be naturally extended to dynamic shapes.

In the next few sections, we demonstrate applications of this elastic framework in the contexts of FDA, shape analysis of parameterized curves, shape analysis of surfaces and 2D image registration.

11.2 Registration in FDA: phase-amplitude separation

Recent years have seen an increasing involvement of functional data in statistical analyses (Ramsay and Silverman 2005; Kneip and Ramsay 2008; Ramsay and Li 1998; Tang and Muller 2008). The variables of interest here are functions on certain intervals, and one is interested in using these variables in a variety of problems, including modeling, prediction, and regression. Examples of functional data include growth curves, mass spectrometry data, bio-signals, human activity data, and so on. These observations are typically treated as square-integrable functions, with the resulting set of functions forming an infinite-dimensional Hilbert space. The standard $\mathbb{L}^2$ inner-product, $\langle f_1, f_2 \rangle = \int f_1(t) f_2(t) dt$, provides the Hilbert structure for comparing and analyzing functions. For example, one can perform function principal component analysis (FPCA) of a given set $\{f_i\}$ using this Hilbert structure. Similarly, a variety of ideas, such as the functional linear regressions, partial least squares, have been proposed for working with functional data.

A difficulty arises when the observed functions exhibit variability in their arguments. In other words, instead of observing a function $f(t)$ on an interval, say $[0, 1]$, one observes a "time-warped" function $f(\gamma(t))$, where γ is a time-warping function. This extraneous effect, termed *phase variability*, has the potential to add artificial variance in the observed data and needs to be accounted for in statistical analysis. Let $\{f_i\}$ be a set of observations of a functional variable f. Then, for any time t, the observations $\{f_i(t)\}$ have some inherent variability. However, if we observe $\{f_i \circ \gamma_i\}$ instead, for random warpings γ_is, then the resulting variability in $\{f_i(\gamma_i(t))\}$ has been enhanced due to random γ_is. The problem of registration of functional data, also called *phase-amplitude separation*, is an important one (Srivastava et al. 2011b; Tucker et al. 2013). Given a set of functions $\{f_i\}$ on a common interval, say $[0, 1]$, the goal is to find a set of warping functions $\{\gamma_i\}$, such that $\{f_i \circ \gamma_i\}$ are aligned/registered. Let Γ denote the set of all warping functions (positive diffeomorphisms from $[0, 1]$ to itself) .

We illustrate a solution to this problem based on a Riemannian metric that has origins in information geometry. This metric can be viewed as an extension of the classical Fisher–Rao metric, or rather its nonparametric version, from pdfs to a more general class of functions as mentioned in Srivastava et al. (2011b). While the original form of this metric is quite complicated, a simplification results from a simple change of variable. For a function $f : [0, 1] \to \mathbb{R}$, define a new function called the *square-root slope function* (SRSF) according to:

$$q : [0, 1] \to \mathbb{R}, \quad q(t) = \text{sign}\{\dot{f}(t)\}\sqrt{|\dot{f}(t)|}$$

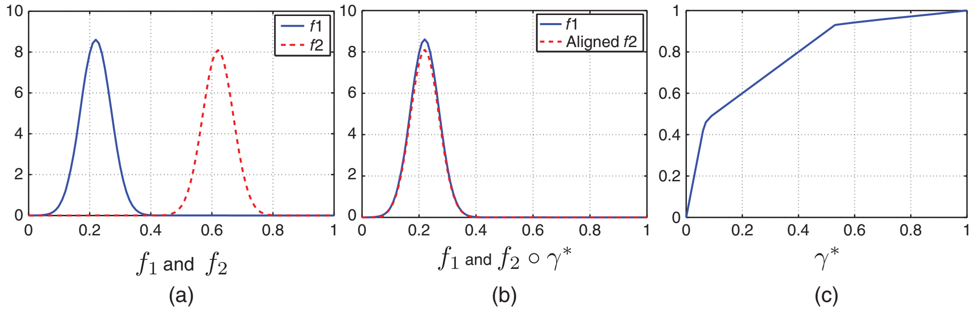


Figure 11.1 Alignment of two functions: align f_2 to f_1. The middle panel shows the aligned result.

If the original f is absolutely continuous, then the resulting q is square integrable. Srivastava et al. (2011b) has shown that the Fisher–Rao metric becomes the $\mathbb{L}^2$ metric under the change of variable $f \longrightarrow q$. Let f_1, f_2 be two functions that need to be registered and let q_1, q_2 be their SRSFs. Then, the registration problem is solved by:

$$\inf_{\gamma \in \Gamma} \|q_1 - (q_2 \circ \gamma)\sqrt{\dot{\gamma}}\| = \inf_{\gamma \in \Gamma} \|q_2 - (q_1 \circ \gamma)\sqrt{\dot{\gamma}}\|. \tag{11.2}$$

The optimization is performed using a numerical procedure called the *dynamic programming algorithm*. Figure 11.1 shows an example of this alignment between two Gaussian density functions. After optimization, the two functions are nicely aligned, as shown in (b), and the resulting optimal warping γ^* is shown in (c).

In case we have multiple functions that need to be aligned, we can extend the previous pairwise alignment as follows. We use the fact that the quantity in Equation (11.2) is actually a proper metric in a certain quotient space and use it to define a mean function. This mean function serves as a template for aligning other functions, that is, each function is aligned to this mean function. In fact, the problem of multiple alignment and mean computation are formulated and solved jointly using an iterative procedure: initialize the mean function μ and iteratively solve for

$$\gamma_i = \arg\inf_{\gamma \in \Gamma} \|\mu - (q_i \circ \gamma)\sqrt{\dot{\gamma}}\|, i = 1, 2, \ldots, n, \text{and}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(q_i \circ \gamma_i)\sqrt{\dot{\gamma}_i}. \tag{11.3}$$

A synthetic example of multiple functions alignment is shown in Figure 11.2. Figure 11.2(a) shows a number of bimodal functions in which the heights and locations of peaks are different. The aligned functions are shown in Figure 11.2(b), and the optimal warping functions γ_is are shown in 11.2(c).

Next we show one example of the multiple functions alignment in a real data set: the Berkeley growth data set, which contains 54 female and 39 male subjects. To better illustrate, we analyze the first derivatives of the growth curves. The results are shown in Figure 11.3. Figure 11.3(a) shows the alignment result for 54 female subjects and (b) shows the alignment result for 39 male subjects. The last column shows the mean

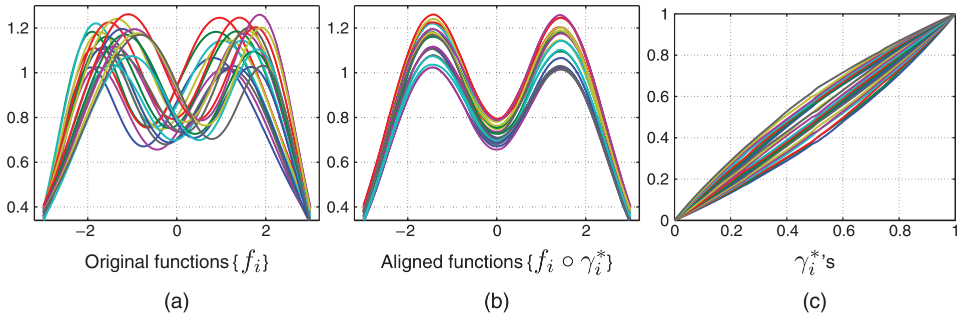


Figure 11.2 Multiple functions alignment. (a) A set of functions which have different height and peak locations. (b) The aligned result. (c) The optimal warping function γ_i^*'s.

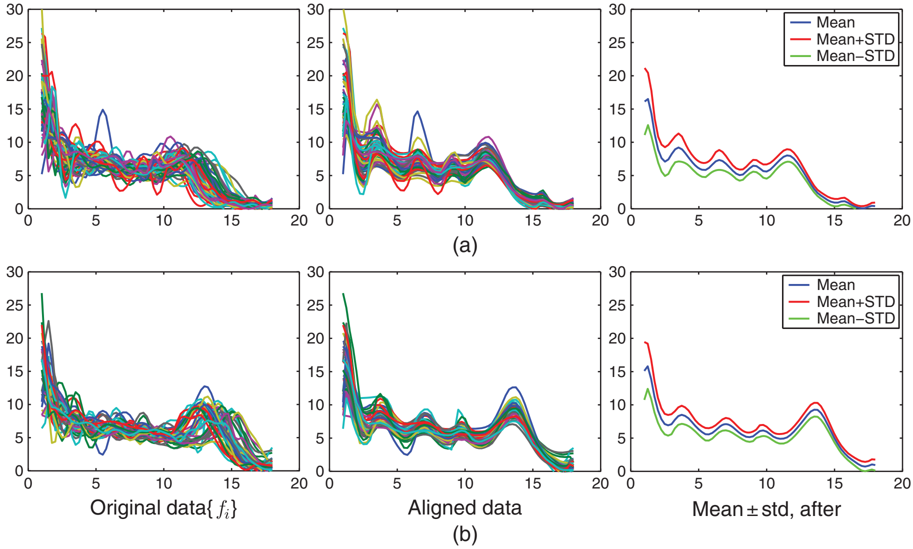


Figure 11.3 Analysis of growth data. (a) The growth data for 54 female subjects. (b) The growth data for 39 male subjects.

± (cross-sectional) standard deviation plot after the alignment. From the result, one can see that while the growth spurts for different individuals occur at slightly different times, there are some underlying patterns to be discovered.

11.3 Elastic shape analysis of curves

The framework for function alignment can be easily extended to perform shape analysis of parameterized curves. Here, the objects of interest are given by parameterized curves $f : [0, 1] \to \mathbb{R}^n$. (Note that in the case of closed curves, it is natural to use $\mathbb{S}^1$ as the parameterization domain, rather than an interval.) The $\mathbb{L}^2$ metric is given by $\langle f_1, f_2\rangle = \int_0^1 \langle f_1(t), f_2(t)\rangle\, dt$ and the resulting norm $\|f_1 - f_2\| = \int_0^1 |f_1(t) - f_2(t)|^2 dt$, where $|\cdot|$ denotes the vector norm. The mathematical representation of curves is in the form of the *square-root velocity function* (SRVF) given by Srivastava et al. (2011a) and Kurtek et al. (2012b):

$$q : [0,1] \to \mathbb{R}^n, \quad q(t) = \frac{\dot{f}(t)}{\sqrt{|\dot{f}(t)|}} .$$

The re-parameterization group here is the set of all positive diffeomorphisms of $[0,1]$. If q is the SRVF of a curve f, then the SRVF of the re-parameterized curve $f \circ \gamma$ is given by $(q \circ \gamma)\sqrt{\dot{\gamma}}$; we will denote this by (q, γ). Other simple representations of planar curves have been presented in Bauer et al. (2013).

From the perspective of shape analysis, a rigid motion (or translation), re-parameterization, rotation, and scaling of a curve do not alter its shape. The translation has been removed by the SRVF representation automatically. An illustration of different parameterizations of a curve is shown in Figure 11.4. The shape of f is exactly the same as the shape of $f \circ \gamma$, for any γ. The same holds for the rigid rotation of a curve. For any $O \in SO(n)$, the rotated curve $Of(t)$ has the same shape as the original curve. If we do not consider the scaling for the moment, this leads to formulation of equivalence classes, or orbits, of representations that all correspond to the same shape. Let $[f]$ denote all possible translations, rotations, and re-parameterizations of a curve f. The corresponding set in SRVF representation is given by $[q] = \{O(q,\gamma) | O \in SO(n), \gamma \in \Gamma\}$. Each such class represents a shape uniquely and shapes are compared by computing a distance between the corresponding orbits.

As mentioned earlier, the SRVF representation satisfies the property that $\|q\| = \|(q,\gamma)\|$, and $\|q_1 - q_2\| = \|(q_1,\gamma) - (q_2,\gamma)\|$ for all $\gamma \in \Gamma$ and all $q, q_1, q_2 \in \mathbb{L}^2$. Using this property, the shape distance between any two shapes is given by

$$d([q_1],[q_2]) = \inf_{\gamma \in \Gamma, O \in SO(n)} \|q_1 - O(q_2,\gamma)\| = \inf_{\gamma \in \Gamma, O \in SO(n)} \|O(q_1,\gamma) - q_2\|. \tag{11.4}$$

This optimization emphasizes the joint nature of our analysis – on the one hand, we optimally register points across two curves using re-parameterization and rotation, and on the other hand, we obtain a metric for comparing shapes of the two curves. The optimization over $SO(n)$ and Γ is performed using coordinate relaxation – optimizing over one variable while fixing the other. The optimization over $SO(n)$ uses the Procrustes method while the optimization over Γ uses the dynamic programming algorithm (Srivastava et al. 2011a). In the absence of any other constraints on the curves,

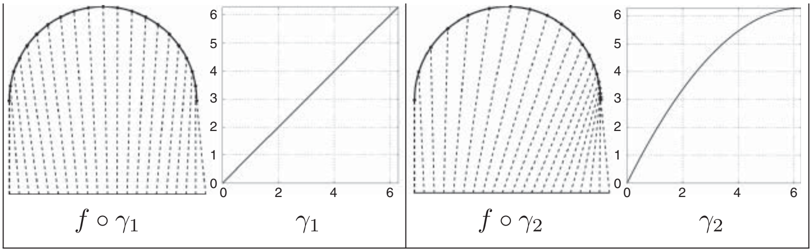


Figure 11.4 An illustration of re-parameterization curve in domain $D = [0, 2\pi]$.

a straight line between q_1 and the registered q_2, that is, $O^*(q_2, \gamma^*)$, with these quantities being the minimizers in Equation (11.4), forms the desired geodesic. However, if we rescale the curves to be of unit length and/or restrict ourselves to only closed curves, then the underlying space becomes nonlinear and requires additional techniques for computing geodesics. We have developed a path-straightening algorithm for computing geodesics in the shape space of closed curves under the elastic metric, as described in Srivastava et al. (2011a). Figure 11.5 shows some examples of geodesic paths between several pairs of closed curves taken from the MPEG7 data set (Jeannin and Bober 1999). One can see that this joint framework deforms one shape to another in a natural way – the features are better preserved across shapes and deformations are smooth.

11.3.1 Mean shape and modes of variations

This framework is amenable to the development of tools for statistical analysis of shapes. For example, given a set of observations of curves, we may want to calculate the sample mean and modes of variations. Furthermore, we are interested in capturing the variability associated with the shape samples using probability models. The notion of a sample mean on a nonlinear manifold is typically defined using the Karcher mean (Karcher 1977). Let $f_1, f_2, \ldots, f_n$ be the observed sample shapes and $q_1, q_2, \ldots, q_n$ be the corresponding SRVFs. The Karcher mean is defined as a quantity that satisfies $[\mu] = \text{argmin}_{[q]} \sum_{i=1}^{n} d([q], [q_i])^2$, where $d([q], [q_i])$ is calculated using Equation (11.4), and μ is the SRVF representation of the mean shape $\bar{f}$. The search for the optimal mean shape $\bar{f}$ can be solved using an iterative gradient-based algorithm (Karcher 1977; Srivastava et al. 2005; Kurtek et al. 2012b). Figure 11.6 shows some sample mean shapes calculated using this approach.

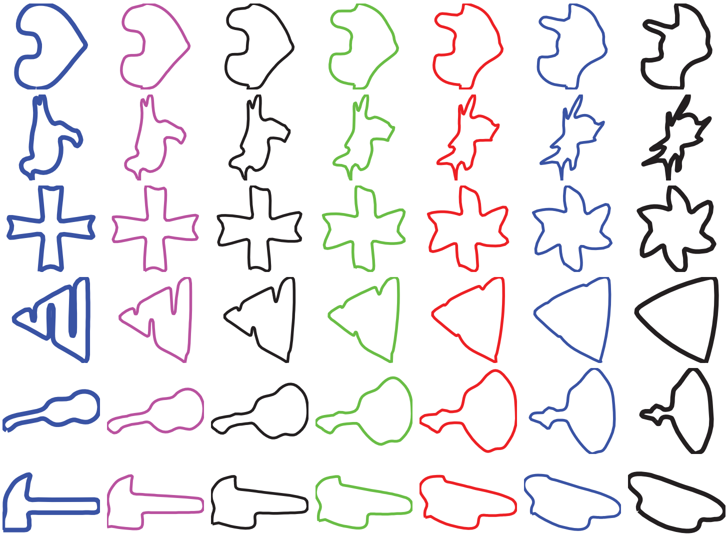


Figure 11.5 Each row shows an example of geodesic path between the starting and ending shapes under the elastic framework.

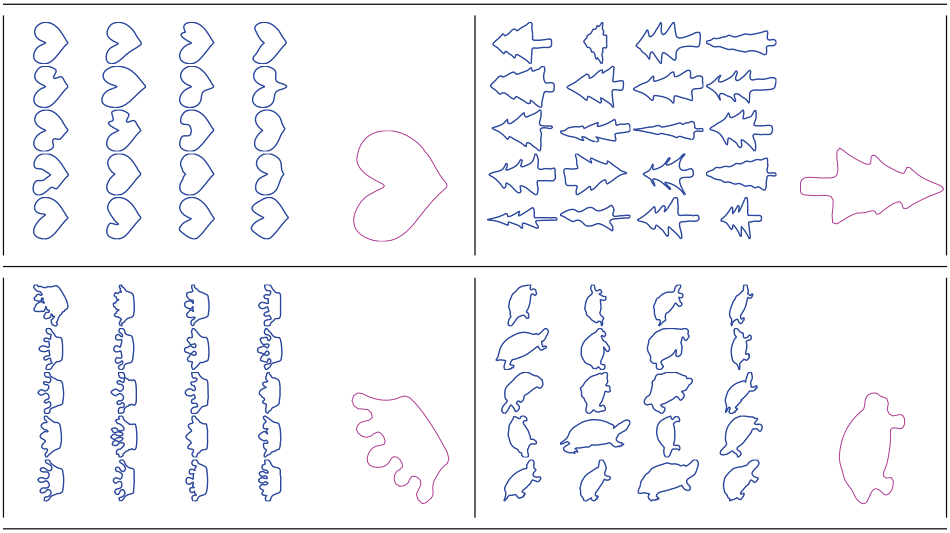


Figure 11.6 Mean shapes of four different classes of shapes. Each mean shape (shown in bottom right) is calculated from shapes on its left.

In addition to the Karcher mean, the Karcher covariance and modes of variation can be calculated to summarize the given sample shapes. Since the shape space is nonlinear, we can use the tangent space at the mean shape μ, which is a standard vector space, to perform the statistical analysis. We first map each sample shape onto the tangent space using inverse exponential map: $v_i = \log_\mu(q_i)$, then we define the covariance matrix to be: $C = \frac{1}{n-1}\sum_{i=1}^{n} v_i v_i^t$. Using principal component analysis (PCA) of C, we can get the modes of shape variation. If PC$_k$ denotes the kth principal direction, then the exponential map $\exp_\mu(t\text{PC}_k s_k)$ as a function of t shows the shape variation in PC$_k$ principal direction with standard deviation s_k. Figure 11.7 shows the modes of variations for different classes of shapes in Figure 11.6.

11.3.2 Statistical shape models

After obtaining the mean and covariance, we develop probability models to capture the distribution of given sample shapes. It is challenging to directly impose a probability density on the nonlinear shape space. A common solution is to impose a distribution on a finite subspace of the tangent vector space. For example, one can restrict to principal subspace of the tangent space at mean μ. Then, we can impose a multivariate Gaussian distribution on the principal subspace with zero mean and covariance matrix obtained from the sample shapes. Figure 11.8 shows the examples of random samples using means and covariance matrices estimated from shapes shown in Figure 11.6.

While traditional shape analysis removes the transformations resulting from rigid motions and global scaling in shape considerations, in elastic shape analysis we additionally remove the effects of re-parameterizations. In some situations, however, there is a need for removing other groups such as the affine and projective groups. For a discussion

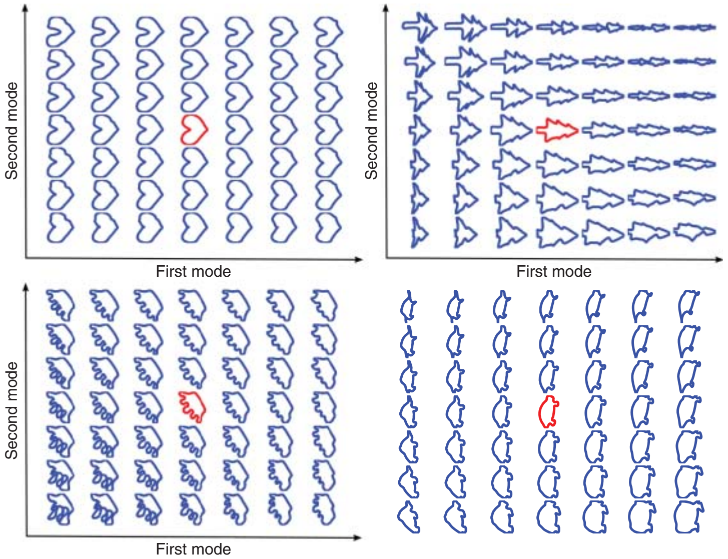


Figure 11.7 Modes of variations: for each class of shapes in mean shape examples (Figure 11.6), we show the variation along the first and second principal modes. Shape in the center is the mean shape.

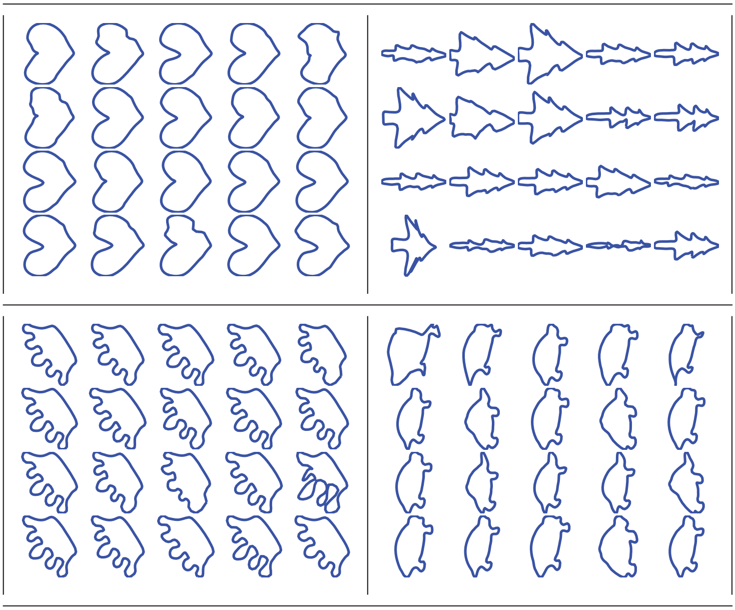


Figure 11.8 Random samples from the Gaussian shape distribution of different classes of shapes.

on the resulting affine-elastic shape analysis of planar curves, we refer the reader to the paper Bryner et al. (2014). This paper also describes a framework for projective-invariant shape analysis of planar objects but using point-set representations rather than continuous curves, using the ideas first proposed in Kent and Mardia (2012).

11.4 Elastic shape analysis of surfaces

The task of comparing shapes of 3D objects is of great interest in many important applications. For instance, the shapes of anatomical parts can contribute in medical diagnoses, including monitoring the progression of diseases (Samir et al. 2014; Grenander and Miller 1998; Kurtek et al. 2011). The main challenge in such shape analyses comes from the fact that image data are often collected from different coordinate systems and data registration becomes a critical part of the analysis. In the following discussion, we focus on surfaces that are embeddings of a unit sphere $\mathbb{S}^2$ in $\mathbb{R}^3$. In other words, the surfaces of interest can be parameterized using the sphere according to a mapping $f : \mathbb{S}^2 \to \mathbb{R}^3$. For any $s \in \mathbb{S}^2$, the vector $f(s) \in \mathbb{R}^3$ denotes the Euclidean coordinates of that point on the surface. The domain of interest is $D = \mathbb{S}^2$, and the $\mathbb{L}^2$ metric is given by $\langle f_1, f_2 \rangle = \int_{\mathbb{S}^2} \langle f_1(s), f_2(s) \rangle \, m(ds)$, with $m(ds)$ denoting the Lebesgue measure on $\mathbb{S}^2$, and the resulting norm is $\|f_1 - f_2\| = \int_{\mathbb{S}^2} |f_1(s) - f_2(s)|^2 m(ds)$. Let (u, v) denote the local coordinates of a point $s \in \mathbb{S}^2$. Then, the vectors $\frac{\partial f}{\partial u}(s)$ and $\frac{\partial f}{\partial v}(s)$ span the two-dimensional space tangent to the surface at point $f(s)$ and

$$n(s) = \frac{\partial f}{\partial u}(s) \times \frac{\partial f}{\partial v}(s)$$

is a vector normal to the surface at $f(s)$. Its magnitude $|n(s)| = \sqrt{\langle n(s), n(s) \rangle}$ denotes infinitesimal area of the current parameterization at that point and the ratio $n(s)/|n(s)|$ gives the unit normal vector. The mathematical representation of surfaces, suitable for elastic shape analysis, termed *square-root normal field* (SRNF), is defined as (Jermyn et al. 2012; Xie et al. 2013):

$$q : \mathbb{S}^2 \to \mathbb{R}^3, \quad q(s) = \frac{n(s)}{\sqrt{|n(s)|}} .$$

The re-parameterization group here is the set of all positive diffeomorphisms of $\mathbb{S}^2$. If q is the SRNF of a surface f, then the SRNF of the re-parameterized surface $f \circ \gamma$ is given by $(q \circ \gamma)\sqrt{J_\gamma}$, where J_γ is the determinant of the Jacobian matrix of the mapping $\gamma : \mathbb{S}^2 \to \mathbb{S}^2$. We will denote this by (q, γ). Similar to the identities presented for previous two cases, this representation also follows the isometry conditions: for all surfaces f, f_1, and f_2, and the corresponding SRNFs q, q_1, and q_2, and all $\gamma \in \Gamma$, we have $\|q\| = \|(q, \gamma)\|$ and $\|q_1 - q_2\| = \|(q_1, \gamma) - (q_2, \gamma)\|$.

Once again, from the perspective of shape analysis, a re-parameterization and a rotation of a surface do not alter its shape. The shape of f is exactly same as that of $O(f \circ \gamma)$, for any $\gamma \in \Gamma$ and $O \in SO(3)$. This motivates the formulation of equivalence classes, or orbits, of representations that all correspond to the same shape. Let $[f]$ denote all possible rotations and re-parameterizations of a surface f. The corresponding set in SRNF representation is given by $[q] = \{O(q, \gamma) | O \in SO(3), \gamma \in \Gamma\}$. Each such class represents a shape uniquely, and shapes are compared by computing a distance between the corresponding orbits. Similar to curves, the joint registration and shape comparison of surfaces are performed according to:

$$\inf_{\gamma \in \Gamma, O \in SO(3)} \|q_1 - O(q_2, \gamma)\| = \inf_{\gamma \in \Gamma, O \in SO(3)} \|O(q_1, \gamma) - q_2\|. \tag{11.5}$$

While the optimization over $SO(3)$ is relatively straightforward, the optimization over Γ is much more difficult here than the curve case. We have developed a gradient-based approach that uses the geometry of the tangent space $T_{\gamma_{id}}(\Gamma)$. It uses a set of vector fields that incrementally deform the current grid on f_2, so as to minimize the cost function given in Equation (11.5).

Similar to the case of constrained curves, the task of computing geodesics between any two registered surfaces is not trivial and requires a path-straightening algorithm (see Kurtek et al. 2012a). More recently, Xie et al. (2014a) have developed an approximation that first computes a straight-line geodesic between any two registered surfaces in the SRNF representation space and then inverts each point along this geodesic to obtain a geodesic in the surface space. For more details, we refer the reader to these papers.

In Figure 11.9, we show some examples of geodesics between objects including human hands and animals. In Figure 11.10, we compare the geodesics between surfaces to the linear interpolation of surfaces. From the results, we can see that the tail part of the cat is distorted and inflated on the linearly interpolated path, but the tail part is better persevered along the geodesic path.

Using geodesics, we can define and compute the mean shape using a standard algorithm for computing Karcher mean. Furthermore, we can define and compute Karcher covariance, and perform PCA on the tangent space at the mean shape. Figure 11.11 displays

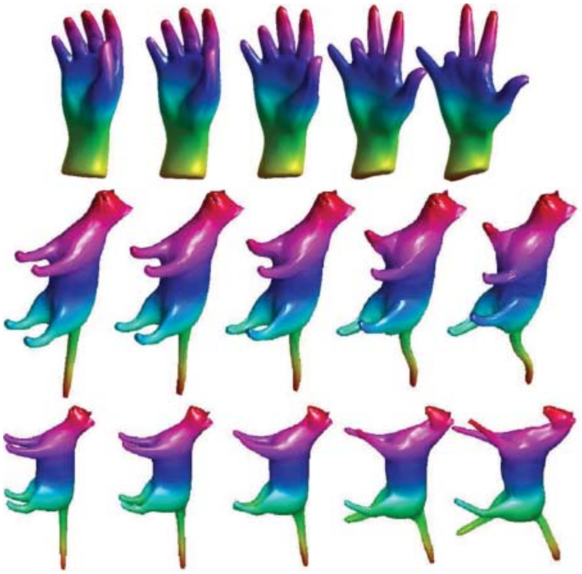


Figure 11.9 Each row shows an example of geodesic between a pair of objects (the starting and ending shapes) (*Source:* Xie et al. 2013, Figure 4, p. 870. Reproduced by permission of IEEE).

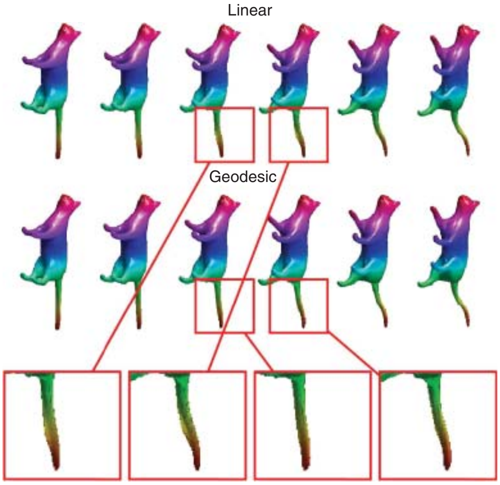


Figure 11.10 Comparing geodesic to linear interpolation (*Source*: Xie et al. 2013, Figure 5, p. 871. Reproduced by permission of IEEE).

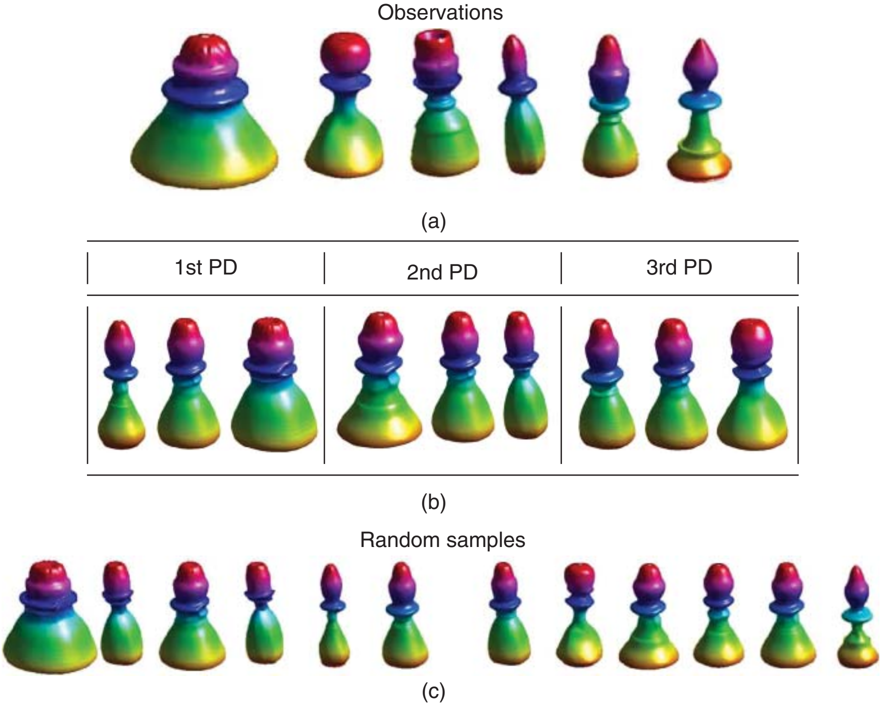


Figure 11.11 Computing mean shape, PC analysis and random samples under a Gaussian model. (a) Some observations of chess piece. (b) The three main principal components. (c) Several randomly sampled chess pieces using a Gaussian model are shown (*Source:* Xie et al. 2013, Figure 9, p. 872. Reproduced by permission of IEEE).

the observations and the kth principal directions (PD) by constructing principal geodesics $\exp_\mu(ts_k \cdot \mathrm{PC}_k)$, where $\mathrm{PC}_k \in T_\mu(\mathcal{F})$ is the kth principal component and s_k denotes the corresponding standard deviation. The PDs are displayed using the triples $\{\exp_\mu(-s_k \cdot \mathrm{PC}_k), \mu, \exp_\mu(s_k \cdot \mathrm{PC}_k)\}$. This analysis can be used to define a multivariate normal distribution on the principal coefficients. Assume that v is a random deformation of the mean surface, that is, $v \in T_\mu(\mathcal{F})$ according to the normal model. Then, we can use the shooting method to get a random sample of surfaces such that $f = \exp_\mu(v)$. Several randomly sampled chess pieces are shown in Figure 11.11c.

11.5 Metric-based image registration

In the problem of image understanding, especially in object recognition and classification using image data, it is important to perform registration of images during their analysis. The importance of image registration comes across clearly in many applications. For example, in developing image templates of different letters and numbers in human handwriting, for the purposes of automated handwriting recognition, it is important to align images of same objects before averaging. To improve performance, it is often necessary to perform a non-rigid alignment, that is, deform one image so as to match its pixel patterns with the other image as much as possible. While these deformations have been performed using various energy minimization methods in the past (Viola and Wells 1995; Collignon et al. 1997; Davies et al. 2002; Twining et al. 2004; Dupuis and Grenander 1998; Trouve 1998; Beg et al. 2005; Miller et al. 2002; Joshi et al. 2004; Lorenzen et al. 2005; Thirion 1998; Vercauteren et al. 2009; Bookstein 1989; Szeliski and Coughlan 1997; Eriksson and Astrom 2006), a novel idea is to use a proper metric for registration. As described in this section, there are several distinct advantages in this approach over the conventional ideas.

An image is treated as a function $f : D \to \mathbb{R}^n$, and the image space is $\mathcal{F} = \{f : D \to \mathbb{R}^n \mid f \in C^\infty(D)\}$. For a gray-level image, we have $n = 1$, and for a colored image, we have $n = 3$. The domain of interest is $D = [0, 1]^2$, and the $\mathbb{L}^2$ metric is given by $\langle f_1, f_2\rangle = \int_D \langle f_1(s), f_2(s)\rangle\, ds$. Let $\Gamma = \mathrm{Diff}^+(D)$ be a subgroup of Diff^+ (the orientation-preserving diffeomorphism group) that preserves the boundary of D. A registration of image f_1 to image f_2 is to find a diffeomorphism $\gamma \in \Gamma$ such that pixel values $f_1(s)$ and $f_2(\gamma(s))$ are optimally matched to each other.

In the elastic framework, the mathematical representation of any image is given by a square-root map (SRM): $q(s) = \sqrt{a(s)}f(s)$, where $a(s)$ is the "generalized area multiplication factor" of f at $s \in D$. It takes the form $a(s) = |\mathbf{J}f(s)|_V$ where $|\mathbf{J}f(s)|_V = \|\frac{\partial f}{\partial x^1} \wedge \frac{\partial f}{\partial x^2}\|$. Here, $\wedge$ denotes the wedge product, $(x^1, x^2) : D \to \mathbb{R}^2$ are the coordinates on (a chart of) D and $\mathbf{J}f(s)$ is the Jacobian matrix of f at s with the (j, i)th element as $\partial f^j/\partial x^i(s)$. The two special cases are as follows: if $n = 2$, then $a(s) = |\mathbf{J}f(s)|$; if and $n = 3$, then $a(s) = \|\frac{\partial f}{\partial x^1}(s) \times \frac{\partial f}{\partial x^2}(s)\|$. Note that this SRM, by definition, applies to images such that $n \geq 2$. In the case of gray-level images, one can use their gradient images, $(f_u, f_v)(s) \in \mathbb{R}^2$, to fit into this representation. Intuitively, the SRM leaves uniform regions as zeros while preserving edge information in such a way that it is compatible with change of variables, that is, stronger edges get higher values.

For $f \in \mathcal{F}$ and any γ, the SRM representation of $f \circ \gamma$ is given by $(q, \gamma) = \sqrt{|\mathbf{J}\gamma|}(q \circ \gamma)$. As mentioned earlier, under this representation, we have $\|q\| = \|(q, \gamma)\|$ and $\|(q_1, \gamma) - (q_2, \gamma)\| = \|q_1 - q_2\|$, for all q, q_1, q_2 and for all $\gamma \in \Gamma$. For the purpose of registration, we define an objective function between two images f_1 and f_2 by

$\mathcal{L}(f_1, (f_2, \gamma)) = \|q_1 - (q_2, \gamma)\|$. The registration of two images is then achieved by minimizing the objective function according to:

$$\gamma^* = \operatorname*{arginf}_{\gamma \in \Gamma} \mathcal{L}(f_1, (f_2, \gamma)) = \operatorname*{arginf}_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\|. \tag{11.6}$$

The optimization problem over Γ in Equation (11.6) forms the crux of our registration framework and is solved using a gradient descent method as in Kurtek et al. (2010 and Xie et al. 2012, 2014b).

This registration framework satisfies a list of fundamental properties such as: (i) it is invariant to simultaneous warping; (ii) it is inverse consistent (Xie et al. 2014b). Additionally, the optimal registration is not affected by scaling and translations of image pixels: let $g_1 = c_1 f_1 + d_1$ and $g_2 = c_2 f_2 + d_2$ with $c_1, c_2 \geq 0$ and $d_1, d_2 \in \mathbb{R}^n$, if $\gamma^* = \operatorname{arginf}_\gamma \mathcal{L}(f_1, (f_2, \gamma))$ then $\gamma^* = \operatorname{arginf}_\gamma \mathcal{L}(g_1, (g_2, \gamma))$ as well.

In Figure 11.12, we first present some results on synthetic images to demonstrate the use of the registration framework suggested in Equation (11.6). The images f_1 and f_2 are registered twice by first taking f_1 as the template image and estimating γ_{21} that optimally deforms f_2 using Equation (11.6). Then, the roles are reversed and f_2 is used as the template to obtain γ_{12}. We show the two converged energies, $\|(q_1, \gamma_{12}) - q_2\|$ and $\|q_1 - (q_2, \gamma_{21})\|$, associated with the optimal γ_{12} and γ_{21} to verify symmetry. The cumulative diffeomorphisms $\gamma_{21} \circ \gamma_{12}$ and $\gamma_{12} \circ \gamma_{21}$ are also used to demonstrate the inverse consistency of the proposed metric. The theory indicates that γ_{12} and γ_{21} are expected to be inverses of each other. We show the original images f_1 and f_2 with the matching warped images $f_2 \circ \gamma_{21}$ and $f_1 \circ \gamma_{12}$, respectively. The diffeomorphisms γ_{12} and γ_{21} learnt to register the images are also presented. By composing them in different orders, we expect the resulting diffeomorphisms to be the identity map. In order to better visualize that the composed diffeomorphisms are

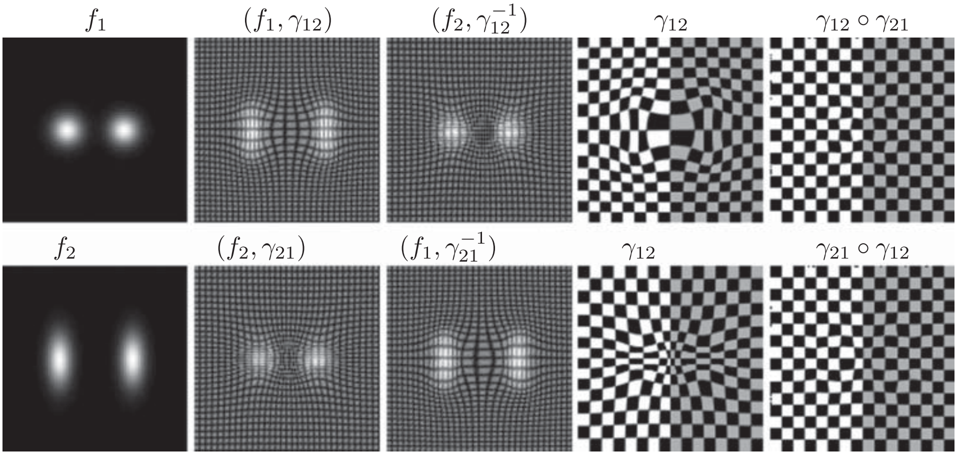

f_1
(f_1, γ_{12})
(f_2, γ_{12}^{-1})
γ_{12}
$\gamma_{12} \circ \gamma_{21}$
f_2
(f_2, γ_{21})
(f_1, γ_{21}^{-1})
γ_{12}
$\gamma_{21} \circ \gamma_{12}$

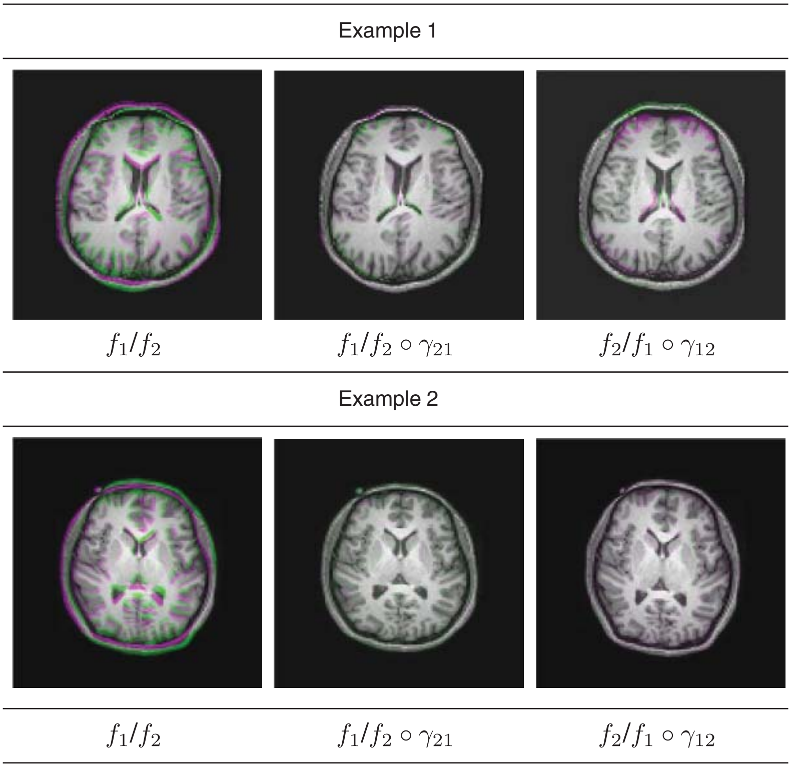
Figure 11.12 Registering synthetic smooth grayscale images. $\gamma_{12} = \operatorname{arginf}_{\gamma \in \Gamma} \|(q_1, \gamma) - q_2\|$ and $\gamma_{21} = \operatorname{arginf}_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\|$. $\|q_1 - q_2\| = 0.2312$, $\|q_1 - (q_2, \gamma_{21})\| = 0.0728$ and $\|(q_1, \gamma_{12}) - q_2\| = 0.0859$ (*Source:* adapted from Xie et al. 2014b, Figure 3, p. 246. Reproduced by permission of Springer).

close to identity, we apply them to checkerboard images. We observe that the composed diffeomorphisms $\gamma_{21} \circ \gamma_{12}$ and $\gamma_{12} \circ \gamma_{21}$ are close to the identity map.

In Figure 11.13, we present registration results using 2D brain MR images. In order to illustrate our method, in each of the two experiments, we show (i) the original images overlapped f_1/f_2 and (ii) overlapped images after registration ($f_1/f_2 \circ \gamma_{21}$ and $f_2/f_1 \circ \gamma_{12}$). The overlapped images show image pairs in a common canvas.

When objects in images have some specific landmarks, either provided by experts or some additional data analysis, they can provide some guidance in defining image correspondence. Automated registration methods routinely produce results that conflict with our contextual knowledge, and annotated landmarks provide a way to reconcile these two ideas. The framework can be further extended so that landmark information is incorporated during registration and all of the nice mathematical properties of the objective function are preserved.

Two pairs of 2D brain MR images are used to illustrate this procedure. In Figure 11.14, we want to register f_1 to f_2. Four landmark points are provided and are displayed in each image. The images are first registered using only the landmarks, with a kernel-based approach, and the resulting deformed image is f_1^{lm}. We further deform f_1^{lm} by applying



Example 1
f_1/f_2
$f_1/f_2 \circ \gamma_{21}$
$f_2/f_1 \circ \gamma_{12}$
Example 2
f_1/f_2
$f_1/f_2 \circ \gamma_{21}$
$f_2/f_1 \circ \gamma_{12}$

Figure 11.13 Two examples of brain MR image registration (each row as an example). First column shows overlapped original images f_1 and f_2; second column shows overlapped images f_1 and deformed f_2; third column shows f_2 and deformed f_1 (*Source:* adapted from Xie et al. 2014b, Figure 4, p. 247. Reproduced by permission of Springer).

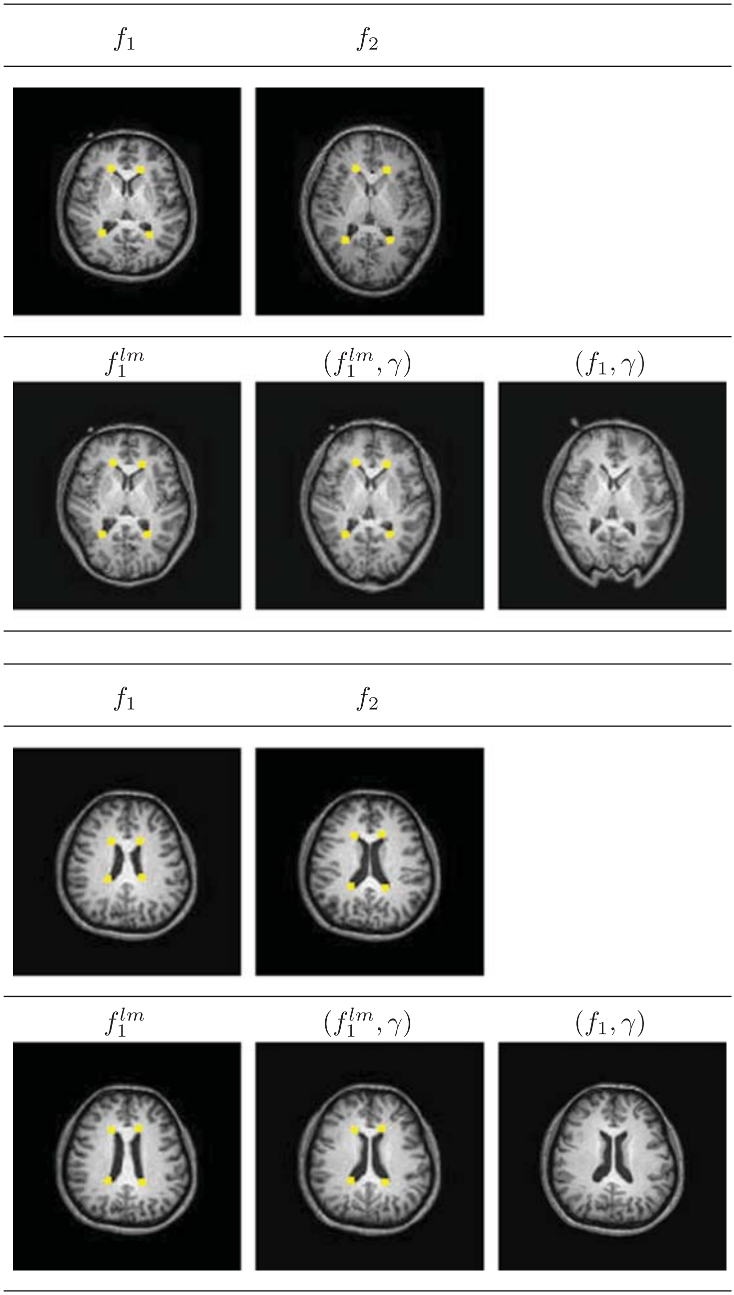


Figure 11.14 Two examples of brain image registration with landmarks. In each experiment, the top row shows the original images f_1 and f_2, and in the bottom row, the first column shows the deformed images f_1^{lm} using only landmarks; the second column shows the final deformed images (f_1^{lm}, γ) with f_1^{lm} as the initial condition; and the last column shows the registered images (f_1, γ) without involving landmarks (*Source:* adapted from Xie et al. 2014b, Figure 5, p. 248. Reproduced by permission of Springer).

our registration method as in Equation (11.6) with restricted vector fields, specified by a set of basis so that the landmark points remain intact. The final result is shown as (f_1^{lm}, γ) and are compared to registration without landmarks as f_1^{lm}. The optimally deformed f_1 without landmark information is displayed in the last column as (f_1, γ) as a baseline. Since the deformation in the skull is so large that our method gives a local solution. By adopting the landmark-aided registration, we at first get a deformed image f_1^{lm}, with the landmarks nicely matched and the skull deformed correspondingly. Then f_1^{lm} is further deformed to register the intensity details without moving the landmarks. The final result (f_1^{lm}, γ) matches f_2 with no artifacts around the skull. Generally, the registration with landmarks outperforms registration without landmarks.

11.6 Summary and future work

We have presented an overview of elastic shape analysis for several kinds of objects, including Euclidean curves, surfaces in $\mathbb{R}^3$, real-valued functions on $[0, 1]$, and 2D images. The analysis is characterized by a simultaneous registration of points across objects and comparisons of their shapes. The key idea is to restrict to parameterized objects and to use parameterization as a tool for registration under metrics that are invariant to all shape-preserving transformations, including re-parameterizations. The use of such metrics is facilitated by square-root mappings of original data because under these mappings, the original metrics become the standard $\mathbb{L}^2$ metric. For each of the data type considered, we present the corresponding square-root transformation and demonstrate some associated statistical tools.

In terms of future work, there are several questions associated with this framework that remain open. An important issue in choosing the elastic metric (or the related square-root representation) is the uniqueness. For instance, in the context of FDA and phase-amplitude separation, one can pose the question: Are there other transforms that allow, under the $\mathbb{L}^2$ metric, an appropriate framework for function registration? In fact, there exist other mappings, for example, $f(t) \mapsto G(f(t), \dot{f}(t))\sqrt{|\dot{f}(t)|}$, where G is an arbitrary function of f and $\dot{f}$, that leads to isometry under the $\mathbb{L}^2$ norm. However, their pros and cons in different situations need to be explored further.

Another important area in shape analysis of curves is the groups beyond the similarity transformations. We have already mentioned the paper by Bryner et al. (2014) that provides affine-invariant shape analysis of elastic curves. However, such an elastic analysis of planar curves that is invariant under projective transformation group, remains to be developed.

References

Bauer M, Bruveris M and Michor PW 2013 R-transforms for Sobolev H^2-metrics on spaces of plane curves. *eprint arXiv:1311.3526*.

Beg M, Miller M, Trouvé A and Younes L 2005 Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision* **61**, 139–157.

Bookstein FL 1989 Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(6), 567–585.

Bryner D, Klassen E, Le H and Srivastava A 2014 2D affine and projective shape analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(5), 998–1011.

Collignon A, Vandermeulen D, Marchal G and Suetens P 1997 Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging* **16**(2), 187–198.

Davies R, Twining C, Cootes T, Waterton J and Taylor C 2002 A minimum description length approach to statistical shape modeling. *IEEE Transactions on Medical Imaging* **21**(5), 525–537.

Dryden I and Mardia K 1998 *Statistical Shape Analysis*, Wiley Series in Probability and Statistics. John Wiley & Sons.

Dupuis P and Grenander U 1998 Variational problems on flows of diffeomorphisms for image matching. *Journal Quarterly of Applied Mathematics* **LVI**(3), 587–600.

Eriksson A and Astrom K 2006 Bijective image registration using thin-plate splines. *International Conference on Pattern Recognition* **3**, 798–801.

Grenander U and Miller MI 1998 Computational anatomy: an emerging discipline. *Quarterly of Applied Mathematics* **LVI**(4), 617–694.

Jeannin S and Bober M 1999 Shape data for the MPEG-7 core experiment ce-shape-1 @ONLINE.

Jermyn IH, Kurtek S, Klassen E and Srivastava A 2012 Elastic shape matching of parameterized surfaces using square root normal fields *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, Berlin, Heidelberg, pp. 804–817.

Joshi S, Davis B, Jomier BM and Guido Gerig B 2004 Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage* **23**, 151–160.

Karcher H 1977 Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics* **30**(5), 509–541.

Kent JT and Mardia KV 2012 A geometric approach to projective shape and the cross ratio. *Biometrika* **99**(4), 833–849.

Kneip A and Ramsay JO 2008 Combining registration and fitting for functional models. *Journal of the American Statistical Association* **103**, 1155–1165.

Kurtek S, Klassen E, Ding Z, Jacobson SW, Jacobson JB, Avison M and Srivastava A 2011 Parameterization-invariant shape comparisons of anatomical surfaces. *IEEE Transactions on Medical Imaging* **30**(3), 849–858.

Kurtek S, Klassen E, Ding Z and Srivastava A 2010 A novel Riemannian framework for shape analysis of 3D objects *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1625–1632.

Kurtek S, Klassen E, Gore JC, Ding Z and Srivastava A 2012a Elastic geodesic paths in shape space of parameterized surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(9), 1717–1730.

Kurtek S, Srivastava A, Klassen E and Ding Z 2012b Statistical modeling of curves using shapes and related features. *Journal of the American Statistical Association* **107**(499), 1152–1165.

Liu W, Srivastava A and Zhang J 2010 Protein structure alignment using elastic shape analysis *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, BCB '10. ACM, New York, NY, pp. 62–70.

Liu W, Srivastava A and Zhang J 2011 A mathematical framework for protein structure comparison. *PLOS Computational Biology* **7**(2), e1001075.

Lorenzen P, Davis B and Joshi S 2005 Unbiased atlas formation via large deformations metric mapping In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2005, Lecture Notes in Computer Science*, vol. 3750 (ed. Duncan J and Gerig G) Springer-Verlag Berlin Heidelberg, pp. 411–418.

Mardia KV and Dryden IL 1989 The statistical analysis of shape data. *Biometrika* **76**(2), 271–281.

Miller M, Trouve A and Younes L 2002 On the metrics and Euler-Lagrange equations of computational anatomy. *Annual Review of Biomedical Engineering* **4**, 375–405.

Ramsay JO and Li X 1998 Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(2), 351–363.

Ramsay JO and Silverman BW 2005 *Functional Data Analysis*, Springer Series in Statistics, 2nd ed. Springer-Verlag.

Samir C, Kurtek S, Srivastava A and Canis M 2014 Elastic shape analysis of cylindrical surfaces for 3D/2D registration in endometrial tissue characterization. *IEEE Transactions on Medical Imaging* **33**(5), 1035–1043.

Srivastava A, Joshi SH, Mio W and Liu X 2005 Statistical shape analysis: clustering, learning, and testing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(4), 590–602.

Srivastava A, Klassen E, Joshi SH and Jermyn IH 2011a Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(7), 1415–1428.

Srivastava A, Wu W, Kurtek S, Klassen E and Marron JS 2011b Registration of functional data using Fisher-Rao metric. *arXiv:1103.3817v2*.

Su J, Kurtek S, Klassen E and Srivastava A 2014 Statistical analysis of trajectories on Riemannian manifolds: bird migration, hurricane tracking and video surveillance. *Annals of Applied Statistics* **8**(1), 530–552.

Szeliski R and Coughlan J 1997 Spline-based image registration. *International Journal of Computer Vision* **22**(3), 199–218.

Tang R and Muller HG 2008 Pairwise curve synchronization for functional data. *Biometrika* **95**(4), 875–889.

Thirion J 1998 Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical Image Analysis* **2**(3), 243–260.

Trouve A 1998 Diffeomorphisms groups and pattern matching in image analysis. *International Journal of Computer Vision* **28**(3), 213–221.

Tucker J, Wu W and Srivastava A 2012 Analysis of signals under compositional noise with applications to sonar data *Oceans, 2012*, pp. 1–6.

Tucker JD, Wu W and Srivastava A 2013 Generative models for functional data using phase and amplitude separation. *Computational Statistics and Data Analysis* **61**, 50–66.

Twining C, Marsland S and Taylor C 2004 Groupwise non-rigid registration: the minimum description length approach. *Proceedings of the British Machine Vision Conference (BMVC)* **1**, 417–426.

Vercauteren T, Pennec X, Perchant A and Ayache N 2009 Diffeomorphic demons: efficient non-parametric image registration. *NeuroImage* **45**(Supplement 1), S61–S72.

Viola P and Wells III W 1995 Alignment by maximization of mutual information *15th International Conference on Computer Vision*, pp. 16–23.

Xie Q, Jermyn I, Kurtek S and Srivastava A 2014a Numerical inversion of SRNFs for efficient elastic shape analysis of star-shaped objects *European Conference on Computer Vision (ECCV)*.

Xie Q, Kurtek S, Christensen GE, Ding Z, Klassen E and Srivastava A 2012 A novel framework for metric-based image registration *Proceedings of the 5th International Conference on Biomedical Image Registration*, WBIR'12. Springer-Verlag, Berlin, Heidelberg, pp. 276–285.

Xie Q, Kurtek S, Klassen E, Christensen GE and Srivastava A 2014b Metric-based pairwise and multiple image registration *2014 European Conference on Computer Vision (ECCV)*.

Xie Q, Kurtek S, Le H and Srivastava A 2013 Parallel transport of deformations in shape space of elastic surfaces *2013 IEEE International Conference on Computer Vision (ICCV)*, pp. 865–872.

Part IV

SPATIAL, IMAGE AND MULTIVARIATE ANALYSIS

12

Evaluation of diagnostics for hierarchical spatial statistical models

Noel Cressie and Sandy Burden

National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, New South Wales, Australia

12.1 Introduction

In the twenty-first century, we can build large, complex statistical models that are very much similar to the scientific processes they represent. We use diagnostics to highlight inadequacies in the statistical model, and because of the complexity many different diagnostics are needed. This is analogous to the process of diagnosis in the medical field, where a suite of diagnostics is used to assess the health of a patient.

This chapter is focused on *evaluating* model diagnostics. In the medical literature, a structured approach to diagnostic evaluation is used, based on measurable outcomes such as Sensitivity, Specificity, Receiver Operating Characteristic (ROC) curves, and False Discovery Rate (FDR). We suggest using the same framework to evaluate model diagnostics for hierarchical spatial statistical models; we note that the concepts are the same in the nonspatial and nonhierarchical setting, although the specific proposals given in this chapter may be difficult to generalize.

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

12.1.1 Hierarchical spatial statistical models

The statistical models that we use to model a spatial process involve many sources of uncertainty, including uncertainty due to the observation process, uncertainty in the spatial process, and uncertainty in the parameters. A hierarchical spatial model allows us to express these uncertainties in terms of conditional probabilities that define, respectively, the data model, the process model, and the parameter model (e.g., Cressie and Wikle 2011, Chapter 2).

Suppose that $Y \equiv \{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ is a spatial process of scientific interest, where $\mathcal{D}$ is a known region in the d-dimensional Euclidean space $\mathbb{R}^d$. We use a spatial statistical model that depends on unknown parameters, $\boldsymbol{\theta}_p$, to quantify our uncertainty in the scientific process of interest, and we use a data model that depends on unknown parameters, $\boldsymbol{\theta}_d$, to quantify our uncertainty in the measurement process. The joint distribution of Y, given all possible parameters $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_d^T, \boldsymbol{\theta}_p^T)^T$, can be written as,

$$[Y|\boldsymbol{\theta}] = [Y|\boldsymbol{\theta}_p], \tag{12.1}$$

where $[A|B]$ is generic notation for the probability density or mass function of A given B. We call (12.1) the *process model*.

Due to measurement error and incomplete sampling, the scientific process is not directly observed. Instead, $\boldsymbol{Z} \equiv (Z(\boldsymbol{s}_1), \dots, Z(\boldsymbol{s}_n))^T$ is observed, whose uncertainty can be quantified through the *data model*,

$$[\boldsymbol{Z}|Y, \boldsymbol{\theta}] = [\boldsymbol{Z}|Y, \boldsymbol{\theta}_d]. \tag{12.2}$$

In a fully Bayesian model, uncertainty in the parameters is quantified through a parameter model,

$$[\boldsymbol{\theta}] = [\boldsymbol{\theta}_d, \boldsymbol{\theta}_p], \tag{12.3}$$

where recall that $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_d$ are the parameters from the process model and the data model, respectively.

The use of conditional distributions to specify a hierarchical statistical model is a powerful way to model complex dependence structures with many sources of uncertainty. Using Bayes' Rule, the posterior distribution for the process and the parameters, which forms the basis for inference in a Bayesian hierarchical model, is given by,

$$[Y, \boldsymbol{\theta}|\boldsymbol{Z}] = [\boldsymbol{Z}|Y, \boldsymbol{\theta}][Y|\boldsymbol{\theta}][\boldsymbol{\theta}]/[\boldsymbol{Z}]. \tag{12.4}$$

Statistical modeling is commonly undertaken to make inference on (i.e., predictions for) the spatial process Y. The usefulness of the hierarchical framework is demonstrated by comparison with a nonhierarchical-model specification. Bayesian, nonhierarchical statistical models implicitly integrate over the process model to obtain the posterior distribution, $[\boldsymbol{\theta}|\boldsymbol{Z}] = \int_Y [\boldsymbol{Z}|Y, \boldsymbol{\theta}][Y|\boldsymbol{\theta}][\boldsymbol{\theta}]/[\boldsymbol{Z}]dY$. When Y is not included in the model specification, the scientific relationships and the observation process are confounded. This has important implications for diagnostics because uncertainty in the measurement process is very different from uncertainty in the scientific process.

12.1.2 Diagnostics

Once we have specified a hierarchical spatial statistical model and fitted it to the data $\boldsymbol{Z}$, we use diagnostics to "stress-test" the model, to assess whether it is adequate for our purposes.

There is a wide range of diagnostics that we may use to do this, because the meaning of "adequate" depends on the purpose of fitting the model in the first place. Analogous to a medical diagnostic, each model diagnostic should be looking for something unusual, to indicate an inadequacy in the model.

The general features of common statistical-model diagnostics are well known and found in many statistical texts (e.g., Carlin and Louis 2009; Gelman et al. 2013; Huber-Carol et al. 2002), including those for hierarchical models (e.g., Banerjee et al. 2004; Cressie and Wikle 2011) and those for spatial data (e.g., Cressie 1993; Gelfand et al. 2010; Schabenberger and Gotway 2005). They include diagnostics to assess residuals (e.g., Belsley et al. 1980; Cook and Weisberg 1982; Cox and Snell 1968; Fox 1991; Kaiser et al. 2012), parameter estimates (e.g., Bousquet 2008; Evans and Moshonov 2006; Presanis et al. 2013), modeling assumptions (e.g., Goel and De Groot 1981; O'Hagan 2003; Scheel et al. 2011), and prior distributions (e.g., Hill and Spall 1994).

Many diagnostic criteria derive from probability measures (e.g., Crespi and Boscardin 2009; Meng 1994; Steinbakk and Storvik 2009), which may or may not be associated with an explicit hypothesis test. Alternatives include visualizing a diagnostic (e.g., Bradley and Haslett 1992; Massmann et al. 2014; Murray et al. 2013) and identifying "interesting" values heuristically or using an empirically derived "rule of thumb."

For hierarchical models, we typically wish to diagnose the adequacy of the model fitted to $[Y|\boldsymbol{\theta}_p]$. However, Y is not observed. Instead we observe data $\boldsymbol{Z}$, which includes measurement error and possible summarization and approximation. Loy and Hofmann (2013), Yan and Sedransk (2007), and Yuan and Johnson (2012) are general references, and an important class of hierarchical-model diagnostics is based on predictive distributions (e.g., Box 1980; Gelfand et al. 1992; Gelman et al. 1996; O'Hagan 2003).

Diagnostics for spatial statistical models (e.g., Anselin and Rey 2010; Christensen et al. 1992; Cressie 1993; Cressie and Wikle 2011; Gelfand et al. 2010; Glatzer and Müller 2004) are more complex due to spatial dependence between locations (e.g., Baddeley et al. 2005; Kaiser et al. 2012; Lee and Ghosh 2009). Global diagnostics applied to the fitted model give an indication of the overall adequacy of the model, but they do not assess the fit of the model at particular locations (e.g., Hering and Genton 2011). Here, local statistics can be powerful diagnostics (see Fotheringham 2009; Fotheringham and Brunsdon 1999, for a review of local analysis), although they can be computationally expensive. Examples include the Local Indicators of Spatial Association (LISA) (Anselin 1995; Getis and Ord 1992; Moraga and Montes 2011; Ord and Getis 1995), LICD, a LISA equivalent for categorical data (Boots 2003), the structural similarity index (SSM) (Robertson et al. 2014; Wang et al. 2004), the S-statistic (Karlström and Ceccato 2002), the local spatial heteroskedasticity statistic (LOSH) (Ord and Getis 2012; Xu et al. 2014) and local diagnostics based on the spatial scan statistic for identifying clusters (Kulldorff et al. 2006; Read et al. 2013).

12.1.3 Evaluation

Model diagnostics are widely used, and questions such as "How reliable are the results of the diagnostic?" and "What are the consequences of using a fitted model that a particular diagnostic deemed inadequate?" naturally arise. In the statistical literature, these questions are answered in ways that include reference to theoretical properties of the diagnostic (e.g., Gneiting 2011; Robins et al. 2000), the performance of the diagnostic on simulated data with known properties (e.g., Dormann et al. 2007), and the distribution of p-values (e.g., He et al.

2013). When a diagnostic is evaluated using the same data that were used to fit the model, the results are well known to be biased (Bayarri and Berger 2000; Efron 1986; Dahl 2006; Hjort et al. 2006). An alternative is to use cross-validation (Gelfand 1996; Le Rest et al. 2014; Stone 1974; Zhang and Wang 2010), where the model is fitted to $m < n$ observations and evaluated using the remaining $n - m$ observations. While cross-validation is considered a gold standard for diagnostics (Gelfand et al. 1992; Marshall and Spiegelhalter 2003; Stern and Cressie 2000), it is computationally expensive and may be impractical for very large data sets. Alternatives such as testing data sets (Efron 1983; 1986), importance sampling (Stern and Cressie 2000), simulation-based model checking (Dey et al. 1998), posterior predictive checks (Gelman et al. 1996; Marshall and Spiegelhalter 2007), and approaches that balance bias with the computational burden of cross-validation (Bayarri and Berger 2000; Bayarri and Castellanos 2007) may also be used.

For hierarchical spatial statistical models, an obvious class of diagnostics identifies those locations where the model is inadequate and those locations where it is adequate. However, in most cases the diagnostic will misclassify some locations. There is potentially a strong parallel here between spatial-model diagnostics and medical diagnostics (e.g., Moraga and Montes 2011; van Smeden et al. 2014), where a diagnostic test is used to identify unusual values (e.g., Pepe and Thompson 2000; Sackett and Haynes 2002). Two summary statistics that are routinely used to assess the performance of medical diagnostics are Sensitivity and Specificity (e.g., Akobeng 2007; Enøe et al. 2000; Hui and Zhou 1998). More recently, there has been a greater use of the FDR (e.g., Benjamini and Hochberg 1995, 1997; Efron 2004; Storey 2003; Storey and Tibshirani 2003; Genovese and Wasserman 2002), and the False Nondiscovery Rate (FNR) (e.g., Craiu and Sun 2008). FDR has been used with correlated data (Benjamini and Yekutieli 2001; Finner et al. 2007; Hu et al. 2010) and, for spatial data, generalized degrees of freedom and clustering may be used to increase the power of the FDR approach (Benjamini and Heller 2007; Shen et al. 2002).

In Section 12.2, we introduce a simple example of county-level sudden infant death syndrome (SIDS) (or cot death) to illustrate our ideas. In Section 12.3, we exploit a strong analogy between medical diagnosis and model diagnosis, and we define the summary measures of Specificity, Sensitivity, FDR, and FNR for evaluating a diagnostic. In Section 12.4, we use these ideas to define a Discovery curve that can be interpreted in an analogous way to the ROC curve. Finally, a discussion and our conclusions are given in Section 12.5.

12.2 Example: Sudden Infant Death Syndrome (SIDS) data for North Carolina

This section introduces an example that will be used to illustrate our proposal for the evaluation of model diagnostics. The data set includes the counts of SIDS for the 100 counties of North Carolina for the period July 1, 1974–June 30, 1978 (Cressie 1993; Cressie and Chan 1989; Symons et al. 1983), where the counties are numbered according to the alphabetical order of their county name. For each county, the data set also includes the number of live births, the spatial location of the county (here specified as the county centroid), and the adjacent counties (i.e., all pairs whose county seats are within 30 miles of each other); see Figure 12.1. The SIDS data have been extensively studied (e.g., Bivand 2014; Cressie 1993; Cressie and Read 1985; Cressie and Chan 1989; Sengupta and Cressie 2013), and

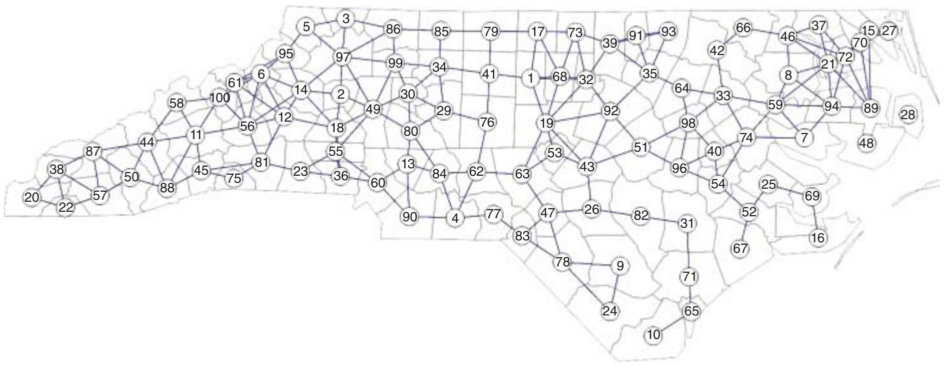


Figure 12.1 Map of the 100 counties in North Carolina, showing edges between the counties whose seats are within 30 miles of each other. The counties are numbered according to the alphabetical order of their county name (*adapted from Bivand 2014*).

they are widely available (e.g., in the spdep package in the R Statistical Software, Bivand 2014; R Core Team 2014).

Our purpose in this chapter is not to identify new diagnostics nor in this section to model the SIDS data in a new way. Instead, we shall model the data with a simple statistical model and then diagnose the fit of the model by using several established diagnostics. Using these results, we shall then evaluate the diagnostics for the model in the manner described in Sections 12.3 and 12.4. For this reason, we base our analysis on the results of previous exploratory analyses conducted by Cressie and Read (1985), Cressie and Chan (1989), and Cressie (1993, Sections 4.4, 6.2, and 7.6). These authors found that the Freeman–Tukey (square-root) transformation of the SIDS rates stabilizes the variance and results in a symmetrical distribution, so that an approximate Gaussian assumption can be made for the transformed data. Most analyses of this transformed data set are based on an auto Gaussian spatial model. We will follow this approach and fit a null statistical model that assumes a constant mean and Gaussian variation in the error. All 100 counties are included; note that in the past, Anson County (county 4) has been identified as an outlier and sometimes removed. Having fitted the model, we use the local Moran I statistic and the local Getis–Ord G^* statistic to assess the adequacy of the fitted model. The local statistics will be applied to the residuals to identify whether there is unusual spatial behavior after the model has been fitted.

In our study, recall that the seat of county i is used to define its location $\boldsymbol{s}_i; i = 1, \dots, 100$. Previous studies have found that the spatial correlation between the counties is close to zero at distances, $d_{ij} \equiv \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$, of 30 miles or more.

For $i = 1, \dots, 100$, let $N(\boldsymbol{s}_i)$ and $S(\boldsymbol{s}_i)$ denote the number of live births and the number of SIDS deaths, respectively, for county i. Its Freeman–Tukey transformed SIDS rate (per thousand live births) is given by

$$Z(\boldsymbol{s}_i) \equiv (1000S(\boldsymbol{s}_i)/N(\boldsymbol{s}_i))^{1/2} + (1000(S(\boldsymbol{s}_i) + 1)/N(\boldsymbol{s}_i))^{1/2} .$$

The null model for the transformed SIDS rate is defined as

$$Z(\boldsymbol{s}_i) = \mu_0 + \delta(\boldsymbol{s}_i), \tag{12.5}$$

where the mean transformed rate, μ_0, is assumed to be constant, and the error, $\delta(\boldsymbol{s}_i)$, is assumed to have a Gaussian distribution with mean zero and variance $\text{var}(\delta(\boldsymbol{s}_i)) = \sigma_\delta^2 V_\delta(\boldsymbol{s}_i)$, for $\sigma_\delta^2 > 0$ and $V_\delta(\boldsymbol{s}_i) \equiv N(\boldsymbol{s}_i)^{-1}$. We fitted this model by using weighted least squares, but not generalized least squares since initially $\delta(\cdot)$ is assumed to exhibit no spatial dependence. The estimate for the mean was 2.84 with a standard error of 0.075.

We would now like to determine whether there is any spatial clustering in the residuals after fitting the null model. To do this, we applied the local Moran I statistic (Anselin 1995), and the local Getis–Ord G^* statistic (Getis and Ord 1992) to the residuals from the model. For a spatial process $\{x_i : i = 1, \ldots, n\}$, the local Moran I statistic is given by

$$I_i = \frac{(x_i - \bar{x}) \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})}{n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{12.6}$$

where w_{ij} is a measure of the spatial dependence between observations i and j. In this example, the spatial-dependence matrix is given by $\boldsymbol{W} \equiv \{w_{ij} : i, j = 1, \ldots, 100\}$, where $w_{ii} = 0$; and for $i \neq j$, $w_{ij} = 1$ when $d_{ij} \leq 30$ miles, and $w_{ij} = 0$ otherwise.

The local Getis–Ord G^* statistic is given by

$$G_i^* = \frac{\sum_{j=1}^{n} c_{ij} x_j - \bar{x} \sum_{j=1}^{n} c_{ij}}{\left(n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 (n \sum_{j=1}^{n} c_{ij}^2 - (\sum_{j=1}^{n} c_{ij})^2)/(n-1)\right)^{1/2}}, \tag{12.7}$$

where the spatial-dependence matrix is given by $\boldsymbol{C} \equiv \{c_{ij} : i, j = 1, \ldots, 100\}$. In this example, $c_{ii} = 1$; and for $i \neq j$, $c_{ij} = 1$ when $d_{ij} \leq 30$ miles, and $c_{ij} = 0$ otherwise.

Values for the local Moran I statistic and the local Getis–Ord G^* statistic are shown in Figures 12.2 and 12.3; in each case, values of the statistic that are “statistically significant for $\alpha = 0.05$” are highlighted. The local Moran I statistic identifies 18 counties with significant spatial dependence. The local Getis–Ord G^* statistic identifies 12 counties with significant spatial dependence. Using the local Moran I statistic, we would conclude that our model is inadequate for four clusterings of counties in the study area. Using the local Getis–Ord G^* statistic, we would conclude that our model is inadequate for three clusterings of counties

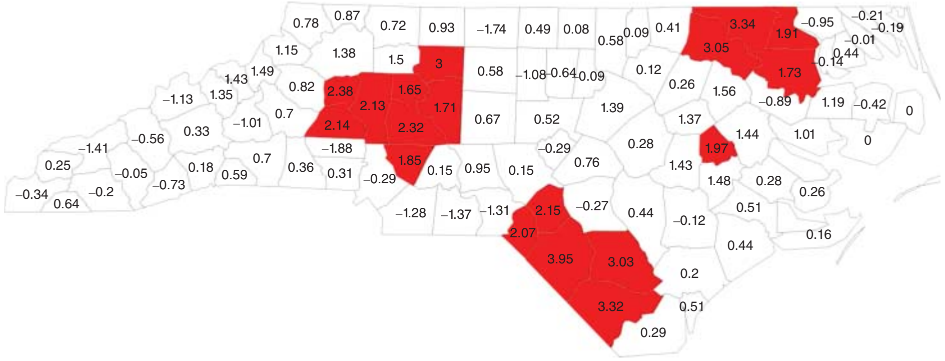


Figure 12.2 Local Moran I statistic for the residuals of the null model fitted using the transformed SIDS rates: Positive (i.e., unusually large) values are shaded.

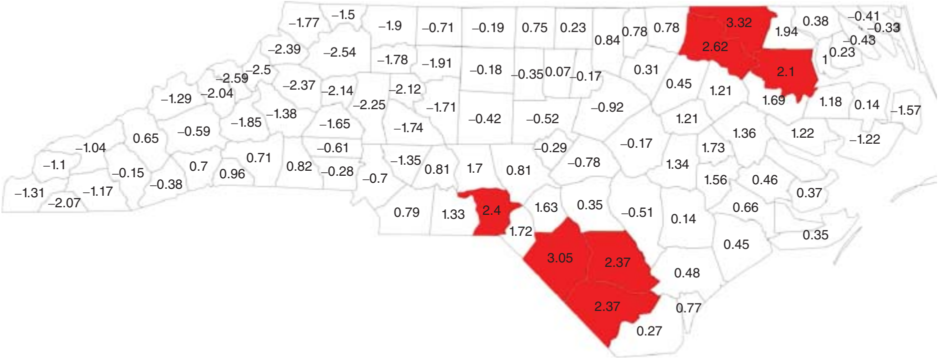


Figure 12.3 Local Getis–Ord G^* statistic for the residuals of the null model fitted using the transformed SIDS rates: Positive (i.e., unusually large) values are shaded.

in the study area. Both diagnostics identify two common spatial clusterings, but each also identifies additional spatial clusterings of counties.

12.3 Diagnostics as instruments of discovery

Whether diagnostics are applied to a spatial model, a hierarchical model, or really any statistical model, they are meant to highlight inadequacies (and adequacies) of the model. While one diagnostic might indicate no inadequacies with a model, it is perfectly plausible that another diagnostic might reveal inadequacies. And just because an inadequacy is found, it does not mean that it is truly an inadequacy. This latter statement may look different from the usual discussion about diagnostics, and it is something we shall pursue in this chapter.

We deem the declaration of an inadequacy of the model a “positive.” Likewise, the declaration of an adequacy is deemed a “negative.” This is clearest in the spatial setting where each datum $Z(\boldsymbol{s}_i)$ at spatial location $\boldsymbol{s}_i$, for $i = 1, \ldots, n$, is potentially a positive (model gives an inadequate fit) or a negative (model gives an adequate fit). If one thinks of diagnosing a model as an act of discovery, analogous to diagnosing a patient in a medical setting (see Section 12.1), then an indication by a diagnostic that something is unusual is seen as a positive.

Discovery of positives and negatives comes with its own uncertainty; a negative could either be a “true negative (TN)” or a “false negative (FN),” and a positive could either be a “false positive (FP)” or a “true positive (TP).” In the spatial setting, if we have n data points and we diagnose the adequacy of each one, then the number of positives (A_P) plus the number of negatives (A_N) equals n. From the aforementioned discussion, we have

$$\begin{aligned} A_{TN} + A_{FN} &= A_N, \\ A_{FP} + A_{TP} &= A_P, \end{aligned} \tag{12.8}$$

where $A_N + A_P = n$, and clearly A_{TN} is the number of *True negatives*, A_{FN} is the number of *False negatives*, A_{FP} is the number of *False positives*, and A_{TP} is the number of *True positives*.

Table 12.1 A 2 × 2 table resulting from our diagnostic evaluation based on a precise follow-up reanalysis.

	Negative	Positive	Total
Diagnostic negative	A_{TN}	A_{FN}	A_N
Diagnostic positive	A_{FP}	A_{TP}	A_P
Total	$A_{TN} + A_{FP}$	$A_{FN} + A_{TP}$	n

The way Equation (12.8) is written suggests Table 12.1, which is a 2 × 2 table, where the rows are classified according to the behavior of the diagnostic, negatives along the first row and positives along the second row. The columns are classified according to a precise "follow-up" reanalysis of each spatial datum; down the first column are the follow-up negatives and down the second column are the follow-up positives. Hence, the top left-hand corner gives the number of True negatives (since both row and column correspond to negatives); the top right-hand corner gives the number of False negatives (since the row is negative but the column shows it should actually be positive); and so forth.

This chapter is about *evaluating* diagnostics and is not directly concerned with defining a "better" diagnostic. Although once we have a yard-stick by which to compare diagnostics, there is a path forward to making them better and better. Our strategy is to take a given diagnostic, based on a particular fitted spatial model, and to determine how well it performs. Just as in the medical setting, we are interested in the diagnostic's False Discovery Rate (FDR), given by

$$\mathit{FDR} = A_{FP}/A_P = A_{FP}/(A_{FP} + A_{TP}), \tag{12.9}$$

and its False Nondiscovery Rate (FNR), given by

$$\mathit{FNR} = A_{FN}/A_N = A_{FN}/(A_{TN} + A_{FN}). \tag{12.10}$$

Notice that the FNR and FDR are obtained from the first and second *rows*, respectively, of the 2 × 2 table given by Table 12.1.

In our evaluation of a diagnostic, we treat it as an algorithm that acts on the n spatial data and, for better or for worse, separates $Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)$ into negatives and positives. A *summary* of this is captured by the counts, A_N and A_P (where recall $A_N + A_P = n$), but the full results of which datum is negative and which is positive are available and can be considered part of the output of the algorithm. Hence, for a *given algorithm* (i.e., diagnostic), the *row totals* A_N and A_P of Table 12.1 *are given*. Consequently, our statistical evaluation is derived from the distribution of A_{FN} and A_{FP}, given A_N and A_P.

Several statistics are routinely used to assess the performance of medical diagnostics, and a similar approach can be used here for model diagnostics. The *Specificity*, or True negative rate, is

$$\mathit{Sp} \equiv A_{TN}/(A_{TN} + A_{FP}), \tag{12.11}$$

which is obtained from the first *column* of Table 12.1. The denominator of (12.11) is the number (out of n) that are in fact negative, as determined by the precise follow-up reanalysis. In a hypothesis-testing setting, $1 - \mathit{Sp}$ is analogous to

$$\text{size} = \alpha \equiv \text{Type I error rate}.$$

The *Sensitivity*, or True positive rate, is

$$Se \equiv A_{TP}/(A_{FN} + A_{TP}), \tag{12.12}$$

which is obtained from the second *column* of Table 12.1. The denominator of (12.12) is the number (out of n) that are in fact positive, as determined by the precise follow-up reanalysis. In a hypothesis-testing setting, Se is analogous to

$$\text{power} = 1 - \beta \equiv 1 - \text{Type II error rate.}$$

In Section 12.4, we suggest alternatives to Sp and Se for assessing the performance of model diagnostics. These are the FDR and the FNR defined by (12.9) and (12.10), respectively.

Recall that we treat a model diagnostic as an algorithm that separates $Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)$ into negatives and positives, and hence A_N and A_P in (12.8) are given. We propose that the precise follow-up reanalysis of each spatial datum (to determine which of the negatives are True and which are False; and which of the positives are False and which are True) is obtained by *cross-validation* (e.g., Hastie et al. 2009, Section 7.10). The model diagnostic is based on a spatial model, and the cross-validation is, of course, based on the *same* spatial model. It is worth noting that cross-validation is typically very slow to implement and, hence, we are only proposing to use it in evaluation. This is analogous to the way a cheap and easy medical diagnostic might be used in the general population, but its evaluation typically involves expensive but precise laboratory analysis.

For cross-validation in the spatial setting, a datum $Z(\boldsymbol{s}_i)$ is held out, and the spatial model is fitted to $\boldsymbol{Z}_{-i} \equiv (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_{i-1}), Z(\boldsymbol{s}_{i+1}), \ldots, Z(\boldsymbol{s}_n))^T$. That model is then used to predict $Z(\boldsymbol{s}_i)$ from data $\boldsymbol{Z}_{-i}$, resulting in a predictor of $Z(\boldsymbol{s}_i)$ that we denote $\hat{Z}_{-i}(\boldsymbol{s}_i)$. Then, a negative at $\boldsymbol{s}_i$ is declared:

$$\begin{array}{rl} \text{True if} & |\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| \leq k_i, \\ \text{False if} & |\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| > k_i, \end{array} \tag{12.13}$$

and a positive at $\boldsymbol{s}_i$ is declared:

$$\begin{array}{rl} \text{False if} & |\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| \leq k_i, \\ \text{True if} & |\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| > k_i, \end{array} \tag{12.14}$$

where $\{k_i : i = 1, \ldots, n\}$ are thresholds determined by the variability in the cross-validation errors,

$$\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i); i = 1, \ldots, n. \tag{12.15}$$

Hence, given the negatives (whose number is A_N) and the positives (whose number is $A_P = n - A_N$), through (12.13) and (12.14), we can obtain all the numbers in Table 12.1. Consequently, we can compute the FDR given by (12.9), the FNR given by (12.10), the Sp given by (12.11), and the Se given by (12.12). We shall see in Section 12.4 how these quantities can be used to evaluate and compare spatial-model diagnostics. However, we first discuss the various entries in Table 12.1, for nonhierarchical models and then for hierarchical models.

12.3.1 Nonhierarchical spatial model

The concepts from which our diagnostic evaluation follows are clearest in the nonhierarchical case. Here, data $\boldsymbol{Z}$ are fitted directly to a spatial model without invoking a hidden model Y to deal with measurement error and “missingness.” The original geostatistical paradigm (Matheron 1963) makes no distinction between $\boldsymbol{Z}$ and Y, and we start with this case. In a sense, this nonhierarchical spatial model is a special case of the hierarchical model in (12.1) and (12.2), where the data model's error variance is zero (e.g., $\sigma_\delta^2 = 0$ for (12.5)). Then, at the location $\boldsymbol{s}_i$ where $Z(\boldsymbol{s}_i)$ is observed, the conditional distribution, $[Z(\boldsymbol{s}_i)|Y] = [Z(\boldsymbol{s}_i)|Z(\boldsymbol{s}_i)]$, is degenerate.

The missing data, which are at locations other than $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, represent unknowns in the model. For example, if there is no observation at $\boldsymbol{s}_0$, then we wish to predict $Z(\boldsymbol{s}_0)$ given $\boldsymbol{Z}$. Kriging (e.g., Cressie 1993, Chapter 3) is based on this. Thus, in the nonhierarchical case, we wish to obtain $[Z(\boldsymbol{s}_0)|\boldsymbol{Z}]$, sometimes called the predictive distribution, to make inference on the missing datum $Z(\boldsymbol{s}_0)$. We shall see in Section 12.3.2 that this goal generalizes to wishing to obtain $[Y(\boldsymbol{s})|\boldsymbol{Z}]$, for all $\boldsymbol{s}$ in the spatial domain of interest.

Cross-validation means that $Z(\boldsymbol{s}_i)$ is predicted from $[Z(\boldsymbol{s}_i)|\boldsymbol{Z}_{-i}]$. That predictor was notated $\hat{Z}_{-i}(\boldsymbol{s}_i)$ earlier, and a common example is

$$\hat{Z}_{-i}(\boldsymbol{s}_i) = E(Z(\boldsymbol{s}_i)|\boldsymbol{Z}_{-i}); \tag{12.16}$$

other predictors are possible. The cross-validation error (12.15) is substituted into (12.13) and (12.14) to determine which of the negatives and positives are True or False, and the counts are summarized in Table 12.1.

The SIDS example discussed in Section 12.2 involved two different diagnostics. The 2×2 table for each of them is given in Tables 12.2 and 12.3. The threshold k_i used for location $\boldsymbol{s}_i$ is given by

$$k_i = k\sigma_\delta / N(\boldsymbol{s}_i)^{1/2}, \tag{12.17}$$

where k is chosen so that

$$\text{Pr}(|N(0,1)| \leq k) = \text{Pr}(|N(0,1)| \geq k) = 0.5,$$

and $N(0, 1)$ is a standard normal random variable. This results in $k = 0.675$, which ensures that we give equal probability to being inside or outside the limit, assuming that the model fits. The map of positives given by cross-validation, namely, the counties where $|\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| > k_i$, for $i = 1, \ldots, n$, is shown in Figure 12.4.

Table 12.2 The 2×2 table given by Table 12.1, for the Local Moran I diagnostic applied to the transformed SIDS residuals after fitting the null model; cross-validation is abbreviated as CV.

	CV negative	CV positive	Total
Diagnostic negative	54	28	82
Diagnostic positive	2	16	18
Total	56	44	100

Table 12.3 The 2×2 table given by Table 12.1, for the Local Getis–Ord G^* diagnostic applied to the transformed SIDS residuals after fitting the null model; cross-validation is abbreviated as CV.

	CV negative	CV positive	Total
Diagnostic negative	53	35	88
Diagnostic positive	3	9	12
Total	56	44	100

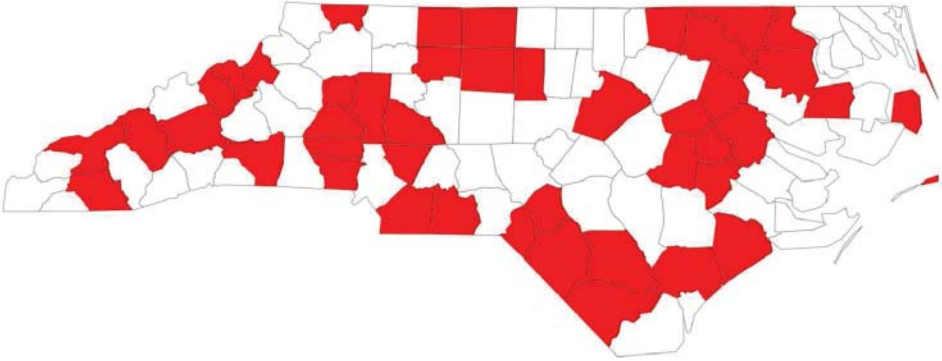


Figure 12.4 Cross-validation for the null model fitted to the transformed SIDS rates: Positive (i.e., unusually large) values are shaded.

Values of smaller k in (12.17) are of obvious interest because the precise follow-up reanalysis is then very stringent; and values up to $k = 1.96$ satisfy $Pr(|N(0,1)| \leq k) \leq 0.95$. Hence, we consider k to vary from small values near zero to values up to 2; in Section 12.4.1, it leads to a new type of curve that we call the *Discovery curve*.

12.3.2 Hierarchical spatial model

From the hierarchical model (12.1) and (12.2), there is a hidden process $Y(\cdot)$ that is to be inferred. In this case, the cross-validation error is

$$\hat{Y}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i); i = 1, \ldots, n, \tag{12.18}$$

where $\hat{Y}_{-i}(\boldsymbol{s}_i)$ is a predictor of $Y(\boldsymbol{s}_i)$ obtained from the predictive distribution, $[Y(\boldsymbol{s}_i)|\boldsymbol{Z}_{-i}]$. A common example is

$$\hat{Y}_{-i}(\boldsymbol{s}_i) = E(Y(\boldsymbol{s}_i)|\boldsymbol{Z}_{-i}).$$

Ideally, we would like to base the criterion for True/False negatives/positives on the error, $\hat{Y}_{-i}(\boldsymbol{s}_i) - Y(\boldsymbol{s}_i)$. However, $Y(\boldsymbol{s}_i)$ is unavailable.

In the hierarchical spatial model, (12.13) and (12.14) are modified, respectively, to: A negative at $\boldsymbol{s}_i$ is declared:

$$\begin{aligned} &\text{True if} \quad |\hat{Y}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| \leq m_i, \\ &\text{False if} \quad |\hat{Y}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| > m_i, \end{aligned} \tag{12.19}$$

and a positive at $\boldsymbol{s}_i$ is declared:

$$\begin{array}{rl} \text{False if} & |\hat{Y}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| \leq m_i, \\ \text{True if} & |\hat{Y}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| > m_i. \end{array} \tag{12.20}$$

The threshold m_i used for location $\boldsymbol{s}_i$ is determined as follows: From (12.1) and (12.2), $Z(\boldsymbol{s}_i) = Y(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i)$, and hence the cross-validation error given by (12.18) is

$$\hat{Y}_{-i}(\boldsymbol{s}_i) - Y(\boldsymbol{s}_i) - \epsilon(\boldsymbol{s}_i),$$

where $\epsilon(\boldsymbol{s}_i)$ is independent of $Y(\boldsymbol{s}_i)$ and $\hat{Y}_{-i}(\boldsymbol{s}_i)$. Its variance is

$$\text{var}(\hat{Y}_{-i}(\boldsymbol{s}_i) - Y(\boldsymbol{s}_i)) + \text{var}(\epsilon(\boldsymbol{s}_i)).$$

Thus, m_i is obtained in a similar manner to k_i with a modification to account for the measurement error, var$(\epsilon(\boldsymbol{s}_i)) \equiv \sigma_\epsilon(\boldsymbol{s}_i)^2$.

If a hierarchical model similar to that given by Cressie (1989) were fitted to the SIDS data in Section 12.2, we would have $\sigma_\epsilon(\boldsymbol{s}_i)^2 = \sigma_\epsilon^2/N(\boldsymbol{s}_i)$, and hence

$$\text{var}(Z(\boldsymbol{s}_i)) = (\sigma_\delta^2 + \sigma_\epsilon^2)/N(\boldsymbol{s}_i),$$

where we assume that σ_ϵ^2 is known (e.g., from spatial-sampling considerations). Consequently, (12.17) is modified to give the following threshold in (12.19) and (12.20):

$$m_i = k(\sigma_\delta^2 + \sigma_\epsilon^2)^{1/2}/N(\boldsymbol{s}_i)^{1/2}, \tag{12.21}$$

where once again $k = 0.675$ gives equal probability to being inside or outside the limit, assuming that the model fits. By varying k from small values near 0 to values up to 2, a Discovery curve for the hierarchical spatial case is obtained; see Section 12.4.2.

12.4 Evaluation of diagnostics

When evaluating medical diagnostics, biostatisticians often use the ROC curve (e.g., Metz 1978), which is a plot of Se (on the vertical axis) versus $1 - Sp$ (on the horizontal axis). It is well known from hypothesis testing that the Type I error rate (i.e., $1 - Sp$) and the Type II error rate (i.e., $1 - Se$) cannot both be kept small. Significance testing puts an upper bound on the Type I error rate (the level of significance) and uses tests whose 1−Type II error rate is large (preferably maximized). To evaluate a medical diagnostic, it is recognized that Sp and Se will co-vary, which is captured by an (x, y) curve in $[0, 1] \times [0, 1]$, where

$$x = 1 - Sp \qquad \text{and} \qquad y = Se.$$

This defines an *ROC curve*, and ideally it is confined to a region of the domain that is close to $(x, y) = (0, 1)$, or at the very least it maintains a consistently high Se for most values of $1 - Sp$. Furthermore, two diagnostics can be compared using their respective ROC curves, by ascertaining which values of $1 - Sp$ lead to a uniformly dominant Se value for one diagnostic over the other. A definitive ordering of several medical diagnostics can be obtained through the *areas* under their respective ROC curves (e.g., Fawcett 2006). In Table 12.1, the ROC curve computes rates with respect to each *column* and plots them. Craiu and

Sun (2008) propose another type of curve with $x = \text{FDR}$ and $y = 1 - Se$, which involves error rates from both a row and a column of Table 12.1.

When a medical diagnostic is applied many times over, error rates computed with respect to the two *rows* of the 2×2 table are more relevant. The analogy to spatial-model diagnostics is immediate, where each datum $Z(\boldsymbol{s}_i)$ at spatial location $\boldsymbol{s}_i$, for $i = 1, \ldots, n$, is potentially a positive or a negative. Thus, we propose to replace the ROC curve with something we call a *Discovery (DSC) curve*; it is an (x, y) curve in $[0, 1] \times [0, 1]$, where

$$x = \text{FDR} \qquad \text{and} \qquad y = 1 - \text{FNR},$$

for FDR and FNR given by (12.9) and (12.10), respectively.

The DSC curve captures the rate of False positives among all positives (plotted on the x-axis) and the rate of True negatives among all negatives (plotted on the y-axis). Ideally, the curve is confined to a region of the domain that is close to $(x, y) = (0, 1)$, or at the very least it maintains a consistently high $1 - FNR$ for most values of FDR. Hence, two diagnostics for a spatial model can be compared using their respective DSC curves, and a definitive ordering can be obtained through the areas under their respective curves.

In the next two subsections, we pursue the DSC-curve approach to evaluating diagnostics, first for nonhierarchical spatial models and then for hierarchical spatial models.

12.4.1 DSC curves for nonhierarchical spatial models

Table 12.1 is obtained from (12.13) and (12.14). If each entry in the table is seen as a function of $\boldsymbol{k} = (k_1, \ldots, k_n)^T$, then by varying $\boldsymbol{k}$, a DSC curve can be obtained. The SIDS example discussed in Section 12.2 and earlier in this section, has a 2×2 table that is determined by a single, normalized threshold k; see (12.17). By varying k from near 0 up to 2, we obtain a DSC curve for each of the two diagnostics. These are shown in Figure 12.5.

Recall the interpretation of these DSC curves; Figure 12.5 shows uniformly superior behavior of the local Moran I diagnostic compared with the local Getis–Ord G^* diagnostic.

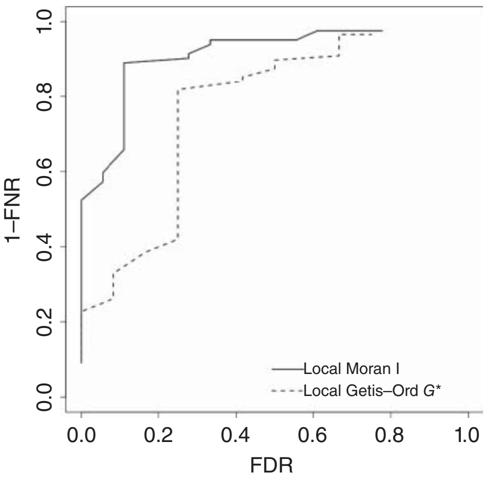


Figure 12.5 DSC curves for the SIDS data, for $0 < k < 2$ in (12.17).

12.4.2 DSC curves for hierarchical spatial models

Because a DSC curve depends on Table 12.1, if we can find such a 2×2 table for a hierarchical spatial model, then everything proceeds as in Section 12.4.1. From Section 12.3.2, we see that each entry in the 2×2 table can be seen as a function of the thresholds $\boldsymbol{m} = (m_1, \ldots, m_n)^T$. Then by varying $\boldsymbol{m}$ a DSC curve can be obtained.

If a hierarchical model similar to that given by Cressie (1989) were fitted to the SIDS data in Section 12.2, we have seen in Section 12.3.2 that $\boldsymbol{m}$ would depend only on a single k (Equation (12.21)) that could be varied from small values near 0 to values up to 2. This would result in a DSC curve for the hierarchical spatial model fitted to the SIDS data, representing the next step in this line of research.

12.5 Discussion and conclusions

This chapter explores the strong analogy between medical diagnostics and spatial-hierarchical model diagnostics. A spatial datum is analogous to an individual whose health is being diagnosed. Medical diagnostics can be evaluated with ROC curves, and in some applications they are investigated using the concept of false discovery rates. We have made the observation that a different curve, which we have called the Discovery (DSC) curve, gives another way to evaluate a diagnostic. For a spatial model, the True negatives and False positives are defined in our proposed evaluation procedure through cross-validation.

By its very nature, a spatial model describes statistical dependence between the data $\boldsymbol{Z}$. Hence, the cross-validation errors given by (12.15) or (12.18) are themselves spatially dependent. In future research, we wish to go beyond our descriptive, visual evaluation of a spatial-model diagnostic and address questions such as, “What is the confidence region for a given $(E(\text{FDR}), E(1 - \text{FNR}))$ pair?” and “Are two DSC curves significantly different?”

Cross-validation is almost always computationally expensive, which is why other diagnostics are preferred when data sets are massive. In this work on evaluation of a model diagnostic, we are willing to spend the computing resources to gauge a diagnostic’s “goodness” on benchmark data sets.

Cross-validation is just one way to define a precise follow-up reanalysis that is used to determine the counts in Table 12.1. Another way would be to base this reanalysis on “testing data sets” proposed by Efron (1983, 1986), which adapt well to the hierarchical-model setting.

Instead of Table 12.1 for nonhierarchical models, this chapter is really about a $2 \times 2 \times 2$ table for hierarchical models where the extra dimension captures a 2×2 table for the Z-process on top of a 2×2 table for the Y-process. The bottom table is hidden since Y is hidden, but it could be thought of as representing an “oracle” table. In this chapter, we have given ways to construct an appropriate 2×2 table and hence an appropriate DSC curve that recognizes the hierarchical nature (i.e., presence of a hidden process Y) of the spatial model, without appealing to the oracle table.

Acknowledgments

This research was partially supported by the US National Science Foundation (NSF) and the US Census Bureau under NSF grant SES-1132031, funded through the NSF-Census

Research Network (NCRN) program. The authors would like to express their appreciation to the editors for their helpful comments.

References

Akobeng AK 2007 Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatrica* **96**, 338–341.

Anselin L 1995 Local indicators of spatial association—LISA. *Geographical Analysis* **27**, 93–115.

Anselin L and Rey SJ 2010 *Perspectives on Spatial Data Analysis*. Springer-Verlag, Heidelberg and New York, NY.

Baddeley A, Turner R, Møller J and Hazelton M 2005 Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society, Series B* **67**, 617–666.

Banerjee S, Carlin BP and Gelfand AE 2004 *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton, FL.

Bayarri MJ and Berger JO 2000 P-values for composite null models. *Journal of the American Statistical Association* **95**, 1127–1142.

Bayarri MJ and Castellanos ME 2007 Bayesian checking of the second levels of hierarchical models. *Statistical Science* **22**, 322–343.

Belsley DA, Kuh E and Welsch RE 1980 *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, Inc., New York.

Benjamini Y and Heller R 2007 False discovery rates for spatial signals. *Journal of the American Statistical Association* **102**, 1272–1281.

Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Benjamini Y and Hochberg Y 1997 Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics* **24**, 407–418.

Benjamini Y and Yekutieli D 2001 The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.

Bivand R 2014 *spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.5–74.

Boots B 2003 Developing local measures of spatial association for categorical data. *Journal of Geographical Systems* **5**, 139–160.

Bousquet N 2008 Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics* **35**, 1011–1029.

Box GEP 1980 Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A* **143**, 383–430.

Bradley R and Haslett J 1992 High-interaction diagnostics for geostatistical models of spatially referenced data. *Statistician* **41**, 371–380.

Carlin BP and Louis TA 2009 *Bayesian Methods for Data Analysis*, 3rd ed. Chapman and Hall/CRC, Boca Raton, FL.

Christensen R, Johnson W and Pearson LM 1992 Prediction diagnostics for spatial linear models. *Biometrika* **79**, 583–591.

Cook RD and Weisberg S 1982 *Residuals and Influence in Regression*. Chapman and Hall, New York.

Cox DR and Snell EJ 1968 A general definition of residuals. *Journal of the Royal Statistical Society, Series B* **30**, 248–275.

Craiu RV and Sun L 2008 Choosing the lesser evil: trade-off between false discovery rate and non-discovery rate. *Statistica Sinica* **18**, 861–879.

Crespi CM and Boscardin WJ 2009 Bayesian model checking for multivariate outcome data. *Computational Statistics and Data Analysis* **53**, 3765–3772.

Cressie N 1989 Empirical Bayes estimation of undercount in the Decennial Census. *Journal of the American Statistical Association* **84**, 1033–1044.

Cressie N 1993 *Statistics for Spatial Data*, rev. edn. John Wiley & Sons, Inc., New York.

Cressie N and Chan NH 1989 Spatial modeling of regional variables. *Journal of the American Statistical Association* **84**, 393–401.

Cressie N and Read TRC 1985 Do sudden infant deaths come in clusters? *Statistics and Decisions, Supplement Issue* **2**, 333–349.

Cressie N and Wikle C 2011 *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Inc., Hoboken, NJ.

Dahl FA 2006 On the conservativeness of posterior predictive p-values. *Statistics and Probability Letters* **76**, 1170–1174.

Dey D, Gelfand A, Swartz T and Vlachos P 1998 A simulation-intensive approach for checking hierarchical models. *Test* **7**, 325–346.

Dormann CF, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, Davies RG, Hirzel A, Jetz W, Kissling WD, Kühn I, Ohlemüller R, Peres-Neto PR, Reineking B, Schröder B, Schurr FM and Wilson R 2007 Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30**, 609–628.

Efron B 1983 Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

Efron B 1986 How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**, 461–470.

Efron B 2004 The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* **99**, 619–642.

Enøe C, Georgiadis MP and Johnson WO 2000 Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine* **45**, 61–81.

Evans M and Moshonov H 2006 Checking for prior-data conflict. *Bayesian Analysis* **1**, 893–914.

Fawcett T 2006 An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874.

Finner H, Dickhaus T and Roters M 2007 Dependency and false discovery rate: asymptotics. *Annals of Statistics* **35**, 1432–1455.

Fotheringham AS 2009 The problem of spatial autocorrelation and local spatial statistics. *Geographical Analysis* **41**, 398–403.

Fotheringham AS and Brunsdon C 1999 Local forms of spatial analysis. *Geographical Analysis* **31**, 340–358.

Fox J 1991 *Regression Diagnostics*. Sage Publications, Newbury Park, CA.

Gelfand AE 1996 Model determination using sampling-based methods In *Markov Chain Monte Carlo in Practice* Gilks WR, Richardson S and Spiegelhalter DJ (eds), Chapman and Hall, London, pp. 145–161.

Gelfand AE, Dey DK and Chang H 1992 Model determination using predictive distributions with implementation via sampling-based methods In *Bayesian Statistics 4* Bernardo JM, Berger JO, Dawid AP and Smith A (eds), Oxford University Press, Oxford, pp. 147–167.

Gelfand AE, Diggle PJ, Fuentes M and Guttorp, P. (eds.) 2010 *Handbook of Spatial Statistics*. Chapman and Hall/CRC, Boca Raton, FL.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A and Rubin DB 2013 *Bayesian Data Analysis*, 3rd ed. Chapman and Hall/CRC, Boca Raton, FL.

Gelman A, Meng XL and Stern HS 1996 Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

Genovese C and Wasserman L 2002 Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B* **64**, 499–517.

Getis A and Ord JK 1992 The analysis of spatial association by use of distance statistics. *Geographical Analysis* **24**, 189–206.

Glatzer E and Müller WG 2004 Residual diagnostics for variogram fitting. *Computers & Geosciences* **30**, 859–866.

Gneiting T 2011 Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**, 746–762.

Goel PK and De Groot MH 1981 Information about hyperparameters in hierarchical models. *Journal of the American Statistical Association* **76**, 140–147.

Hastie T, Tibshirani R and Friedman JH 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer-Verlag, New York.

He D, Xu X and Liu X 2013 The use of posterior predictive p-values in testing goodness-of-fit. *Communications in Statistics - Theory and Methods* **42**, 4287–4297.

Hering AS and Genton MG 2011 Comparing spatial predictions. *Technometrics* **53**, 414–425.

Hill SD and Spall JC 1994 Sensitivity of a Bayesian analysis to the prior distribution. *IEEE Transactions on Systems, Man, and Cybernetics* **24**, 216–221.

Hjort NL, Dahl FA and Steinbakk GH 2006 Post-processing posterior predictive p-values. *Journal of the American Statistical Association* **101**, 1157–1174.

Hu JX, Zhao H and Zhou HH 2010 False discovery rate control with groups. *Journal of the American Statistical Association* **105**, 1215–1227.

Huber-Carol C, Balakrishnan N, Nikulin MS and Mesbah M 2002 *Goodness-of-Fit Tests and Model Validity*. Birkhäuser, Boston, MA.

Hui SL and Zhou XH 1998 Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* **7**, 354–370.

Kaiser MS, Lahiri SN and Nordman DJ 2012 Goodness-of-fit tests for a class of Markov random field models. *Annals of Statistics* **40**, 104–130.

Karlström A and Ceccato V 2002 A new information theoretical measure of global and local spatial association. *Jahrbuch für Regionalwissenschaft* **22**, 13–40.

Kulldorff M, Huang L, Pickle L and Duczmal L 2006 An elliptic spatial scan statistic. *Statistics in Medicine* **25**, 3929–3943.

Le Rest K, Pinaud D, Monestiez P, Chadoeuf J and Bretagnolle V 2014 Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography* **23**, 811–820.

Lee H and Ghosh SK 2009 Performance of information criteria for spatial models. *Journal of Statistical Computation and Simulation* **79**, 93–106.

Loy A and Hofmann H 2013 Diagnostic tools for hierarchical linear models. *Wiley Interdisciplinary Reviews: Computational Statistics* **5**, 48–61.

Marshall EC and Spiegelhalter DJ 2003 Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine* **22**, 1649–1660.

Marshall EC and Spiegelhalter DJ 2007 Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis* **2**, 409–444.

Massmann C, Wagener T and Holzmann H 2014 A new approach to visualizing time-varying sensitivity indices for environmental model diagnostics across evaluation time-scales. *Environmental Modelling and Software* **51**, 190–194.

Matheron G 1963 Principles of geostatistics. *Economic Geology* **58**, 1246–1266.

Meng XL 1994 Posterior predictive p-values. *Annals of Statistics* **22**, 1142–1160.

Metz CE 1978 Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**, 283–298.

Moraga P and Montes F 2011 Detection of spatial disease clusters with LISA functions. *Statistics in Medicine* **30**, 1057–1071.

Murray K, Heritier S and Müller S 2013 Graphical tools for model selection in generalised linear models. *Statistics in Medicine* **32**, 4438–4451.

O'Hagan A 2003 HSSS model criticism In *Highly Structured Stochastic Systems* Green PJ, Hjort NL and Richardson S (eds), Oxford University Press, Oxford, pp. 423–443.

Ord JK and Getis A 1995 Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* **27**, 286–306.

Ord JK and Getis A 2012 Local spatial heteroscedasticity (LOSH). *Annals of Regional Science* **48**, 529–539.

Pepe MS and Thompson ML 2000 Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123–140.

Presanis AM, Ohlssen D, Spiegelhalter DJ and de Angelis D 2013 Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statistical Science* **28**, 376–397.

R Core Team 2014 *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria.

Read S, Bath PA, Willett P and Maheswaran R 2013 New developments in the spatial scan statistic. *Journal of Information Science* **39**, 36–47.

Robertson C, Long JA, Nathoo FS, Nelson TA and Plouffe CCF 2014 Assessing quality of spatial models using the structural similarity index and posterior predictive checks. *Geographical Analysis* **46**, 53–74.

Robins JM, Van der Vaart A and Ventura V 2000 Asymptotic distribution of p-vlues in composite null models. *Journal of the American Statistical Association* **95**, 1143–1156.

Sackett DL and Haynes RB 2002 The architecture of diagnostic research. *British Medical Journal* **324**, 539–541.

Schabenberger O and Gotway CA 2005 *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL.

Scheel ID, Green PJ and Rougier JC 2011 A graphical diagnostic for identifying influential model choices in Bayesian hierarchical models. *Scandinavian Journal of Statistics* **38**, 529–550.

Sengupta A and Cressie N 2013 Empirical hierarchical modelling for count data using the spatial random effects model. *Spatial Economic Analysis* **8**, 389–418.

Shen X, Huang HC and Cressie N 2002 Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association* **97**, 1122–1140.

Steinbakk GH and Storvik GO 2009 Posterior predictive p-values in Bayesian hierarchical models. *Scandinavian Journal of Statistics* **36**, 320–336.

Stern HS and Cressie N 2000 Posterior predictive model checks for disease mapping models. *Statistics in Medicine* **19**, 2377–2397.

Stone M 1974 Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* **36**, 111–147.

Storey JD 2003 The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–2035.

Storey JD and Tibshirani R 2003 Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445.

Symons MJ, Grimson RC and Yuan YC 1983 Clustering of rare events. *Biometrics* **39**, 193.

van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KGM and de Groot JAH 2014 Latent class models in diagnostic studies when there is no reference standard - a systematic review. *American Journal of Epidemiology* **179**, 423–431.

Wang Z, Bovik AC, Sheikh HR and Simoncelli EP 2004 Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**, 600–612.

Xu M, Mei CL and Yan N 2014 A note on the null distribution of the local spatial heteroscedasticity (LOSH) statistic. *Annals of Regional Science* **52**, 697–710.

Yan G and Sedransk J 2007 Bayesian diagnostic techniques for detecting hierarchical structure. *Bayesian Analysis* **2**, 735–760.

Yuan Y and Johnson VE 2012 Goodness-of-fit diagnostics for Bayesian hierarchical models. *Biometrics* **68**, 156–164.

Zhang H and Wang Y 2010 Kriging and cross-validation for massive spatial data. *Environmetrics* **21**, 290–304.

13

Bayesian forecasting using spatiotemporal models with applications to ozone concentration levels in the Eastern United States

Sujit Kumar Sahu[1], Khandoker Shuvo Bakar[2] and Norhashidah Awang[3]

[1] *Mathematical Sciences and S^3RI, University of Southampton, Southampton, UK*

[2] *Department of Statistics, Yale University, New Haven, CT, USA*

[3] *School of Mathematical Sciences, Universiti Sains Malaysia, Penang, Malaysia*

13.1 Introduction

Bayesian forecasting in time and interpolation in space is a challenging task due to the complex nature of spatiotemporal dependencies that need to be modeled for better understanding and description of the underlying processes. The problem exacerbates further when the geographical study region, such as the one in the Eastern United States considered in this chapter, is vast and the training data set for forecasting, and modeling, is rich in both space and time. This chapter develops forecasting methods, and the associated computation methods using Markov chain Monte Carlo (MCMC), for three recently proposed hierarchical

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

Bayesian models for spatiotemporal data sets. A number of forecast calibration measures are also described and their computation methods developed to facilitate rigorous comparisons of Bayesian forecasting methods. The methods are illustrated with a test data set on daily maximum eight-hour average ozone concentration levels observed over a study region in the Eastern United States. Forecast validations, using several moving windows, find a model developed using an approximate Gaussian predictive process (GPP) to be the best, and it is the only viable method for large data sets when computing speed is also taken into account. The methods are implemented in a recently developed software package, spTimer, which is a publicly available contributed R package that has wider applicability.

Bayesian forecasting methods are very much in demand in many application areas in environmental monitoring and surveillance. Consequently, model-based forecasting has attracted much attention in the literature, see for example, Bauer et al. (2001), Damon and Guillas (2002), Feister and Balzer (1991), Huerta et al. (2004), Kumar and Ridder (2010), Mardia et al. (1998), McMillan et al. (2005), Sahu and Bakar (2012a), Sahu and Mardia (2005a, 2005b), Sahu et al. (2009, 2011), Sousa et al. (2009), Stroud et al. (2001), West and Harrison (1997) and Zidek et al. (2012). Some of these papers also consider space–time modeling for forecasting. However, the methods proposed in these articles are not able to handle the computational burden associated with large space–time data sets that we model in this chapter for forecasting purposes.

For point referenced spatial data from a large number of locations, exact likelihood-based inference becomes unstable and infeasible since it involves computing quadratic forms and determinants associated with a high-dimensional variance-covariance matrix Stein (2008). Besides the problem of storage (Cressie and Johannesson 2008), matrix inversion, at each iteration of the model fitting algorithm, such as the EM algorithm, is of $O(n^3)$ computational complexity, which is prohibitive, where n is a large number of modeled spatial locations. This problem also arises in evaluation of joint or conditional distributions in Gaussian process–based models under a hierarchical Bayesian setup; see for example, Banerjee et al. (2004). To tackle this problem, we develop a Bayesian forecasting method based on a model recently developed by Sahu and Bakar (2012b), using GPP approximation method for the underlying spatial surface, see Banerjee et al. (2008). Throughout this chapter, for convenience, we shall use the acronym GPP to also denote the modeling method based on the GPP approximation.

Forecasting using hierarchical Bayesian models is further limited by the lack of suitable software packages. There are a few available packages for forecasting using variants of the dynamic linear models (West and Harrison 1997), see for example, Petris et al. (2010). However, these packages do not allow incorporation of rich spatial covariance structure for the modeled data. On the other hand, `spBayes`, a recently developed spatial data analysis package, developed by Finley et al. (2007), can model short-length time series data by treating those as multivariate spatial data, but it is not really intended to handle large volume of spatiotemporal data that can be analyzed using the `spTimer` package developed by Bakar and Sahu (2014).

This chapter develops forecasting methods for three Bayesian hierarchical models that have been implemented in `spTimer`. The first of these is independent in time Gaussian process (GP) -based regression model, which is simple to implement and is often regarded as a starting model. The second is the hierarchical auto-regressive model developed by Sahu et al. (2007), which has been shown to be better in out-of-sample validation than some versions of dynamic linear models (Sahu and Bakar 2012a) and also a wide class of models (Cameletti et al. 2011). The third and final forecasting method is the one based on the GPP approximation method mentioned earlier. These methodological developments are then used to augment the `spTimer` package with the forecasting modules that can be used in a wide variety of applications in space–time data analysis.

Another objective of the chapter is to rigorously compare the Bayesian forecasts obtained from the three models. Toward this end, we develop MCMC implementation methods for several forecast calibration measures and diagnostic plots that have been proposed to compare the skills of the Bayesian forecast distributions, see for example, Gneiting et al. (2007). The measures include the continuous ranked probability score (CRPS), which is an integrated distance between the forecasts and the corresponding observations, the hit and false alarm rates and the empirical coverage. The diagnostic plots include the probability integral transform (PIT) and a marginal calibration plot (MCP) that is used to calibrate the equality of the forecast and the actual observations; see Section 13.4. These measures and plots enable us to compare the implied Bayesian forecast distributions fully – not just their specific characteristics, for example, the mean forecast, as would be done by simple measures such as the root-mean-square error (RMSE) and the mean absolute error (MAE).

A substantial application on an air pollutant, ground-level ozone, illustrates the forecasting methods of this chapter. Ground-level ozone is a pollutant that is a significant health risk, especially for children with asthma and vulnerable adults with respiratory problems. It also damages crops, trees, and other vegetation. It is a main ingredient of urban smog. Because of these harmful effects, air pollution regulatory authorities are required by law to monitor ozone levels, and they also need to forecast in advance, so that at-risk population can take necessary precaution in reducing their exposure. In the United States (US), a part of which is our study region in this chapter, the forecasts are issued, often, up to 24 hours in advance by various mass-media, for example, newspapers and also the website `airnow.gov`. However, ozone concentration levels, and also other air pollutants, are regularly monitored by only a finite number of sites. Data from these sparse network of monitoring sites need to be processed for developing accurate forecasts. In this chapter, we compare the forecasts of ground-level ozone, based on three models using a 3-week test data set on daily maximum ozone concentration levels observed over a large region in the Eastern United States.

The rest of this chapter is organized as follows: Section 13.2 describes the validation data set we use in this chapter with some summary statistics. In Section 13.3, we develop forecasting methods based on three recently proposed Bayesian spatiotemporal models. Section 13.4 discusses several useful and important forecast calibration methods and develops their MCMC implementation techniques. These are used to compare the forecasting methods with a smaller subset of the full validation data set in Section 13.5. This investigation finds that the GPP model is fast and it performs the best. Subsequently, this model is used in Section 13.6 to analyze and forecast for the full Eastern US data set. Finally, Section 13.7 concludes with a few summary remarks.

13.2 Test data set

The forecasting models proposed in this chapter will be tested using *daily* ozone concentration data for the 3-week period, June 24 to July 14 in 2010. A daily observation, measured in units of parts per billion (ppb), is the maximum of 24 averages in a day where each

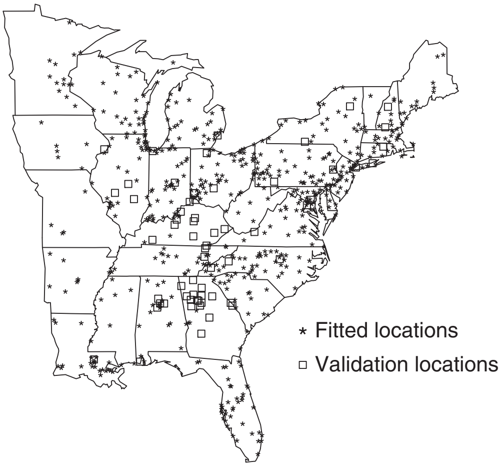


Figure 13.1 A plot of the 639 (62 validation and 577 model fitting) ozone monitoring sites in the Eastern United States.

average is based on hourly ozone concentration readings from eight consecutive hours. In this chapter, we use daily data from 639 monitoring sites in the Eastern United States. We aim to perform forecast validation for completely out-of-sample data from sites that we do not use for modeling at all. Hence, we set aside data from 62 randomly chosen sites (roughly 10%) for validation purposes. Figure 13.1 provides a map of these validation sites and the remaining 577 sites, data from which are used for modeling.

We perform forecast validation for seven moving windows of data from July 8 to July 14. July 8 is taken to be the earliest day for forecast validation that allows modeling of data for 14 days from June 24 to July 7. We also compare the next day forecasts based on modeling data from just seven previous days that complete a weekly cycle. Thus, for example, for forecasting for July 8 we use data from July 1–7.

Often, see, for example, `airnow.gov`, a deterministic model, known as the community multiscale air quality (CMAQ) model, is used for forecasting levels of ozone concentration and other air pollutants such as particulate matter. The CMAQ model in forecasting mode, known as Eta CMAQ, is based on emission inventories, meteorological information, and land use, and it produces gridded forecasts, up to two days in advance, for average ozone concentration levels at each cell of a 12 square-kilometer grid covering the whole of the continental US (Ching and Byun, 1999). However, these outputs are well known to produce biased forecasts, and to reduce this bias, in this chapter, we develop statistical models that can improve the Eta CMAQ forecasts by refining those in the light of the observed monitoring data. Incorporation of gridded CMAQ forecasts in a spatial model for point referenced monitoring data poses a spatial misalignment problem that is well known in the literature; see for example, Fuentes and Raftery 2005), Jun and Stein 2004, Lorence (1986). To incorporate the Eta CMAQ output, the hierarchical models are set up as spatiotemporal downscaler models, first implemented by Sahu et al. (2009) and then generalized by Berrocal et al. (2010b, 2010a) and Zidek et al. (2012). We use the forecasts for daily maximum 8-hour average CMAQ ozone concentration for the grid cell covering the monitoring site as the single covariate, following Sahu et al. (2009).

Many meteorological variables such as the daily maximum temperature are important predictors of ozone levels; see for example, Sahu et al. (2007). However, the meteorological variables no longer remain significant if the model for ozone levels also includes output of the CMAQ model; see for example, Sahu and Bakar (2012a). Moreover, direct inclusion of the meteorological variables in an ozone concentration forecasting model will also require forecasting of the meteorological variables in the first place. The models proposed in this chapter avoid this, although we note that the CMAQ forecasts already include future values of the meteorological variables that have been used as model inputs.

Out of the 13,419 observations from 639 sites for 21 days, 299 $(\approx 2.23\%)$ are missing. Our Bayesian models automatically estimate those by simulating from their full conditional distribution in each iteration of the Gibbs sampler. Table 13.1 provides the summary statistics for ozone levels and Eta CMAQ output, where it is seen that the Eta CMAQ forecasts are upwardly biased, although the medians seem to be close. Figure 13.2 investigates this further by providing side-by-side boxplots for each of 21 days for both the observed and the Eta CMAQ forecasted ozone levels. This figure also shows that the data set includes an episode of high ozone levels during days 12–16, which corresponds to July 5–9, just after the 4th of July celebrations in the United States. This episode of high ozone levels provides an opportunity to model and forecast when demand is likely to be higher than usual.

13.3 Forecasting methods

13.3.1 Preliminaries

We first define the generic notations that we need and use throughout the chapter. Let t denote the time where $t = 1, \dots, T$ and T is the total number of time units. Let $Y(\mathbf{s}_i, t)$ denote the observed point referenced data at location $\mathbf{s}_i$ and at time t for $i = 1, \dots, n$ where n is the total number of locations. Modeling the data on the original scale, as noted by many authors; see for example, Sahu et al. (2007), is prohibitive due to the instability in variance that often leads to negative forecasts. In this chapter, we model data on the square-root scale, denoted by $Z(\mathbf{s}_i, t)$, that encourages symmetry and normality; see for example, Sahu et al. (2007), but report all forecasts and predictions on the original scale, Y, for ease of interpretation by practitioners, although this may increase the mean square error of the forecasts. With this approach, negative forecasts on the square-root scale are conveniently truncated to zero, although we were never required to do this in our examples here. We also note that other variance stabilizing transformations such as log and the more general Box–Cox transformation can also be adopted depending on the nature of the problem, and finally, the methods we describe subsequently can also be used if a variance stabilizing transformation

Table 13.1 Summaries of the daily maximum ozone concentration levels and Eta CMAQ output for the test data set described in Section 13.2.

	Minimum	Mean	Median	Maximum
Ozone levels	0.00	50.62	50.99	113.00
CMAQ output	16.50	59.19	60.36	145.50

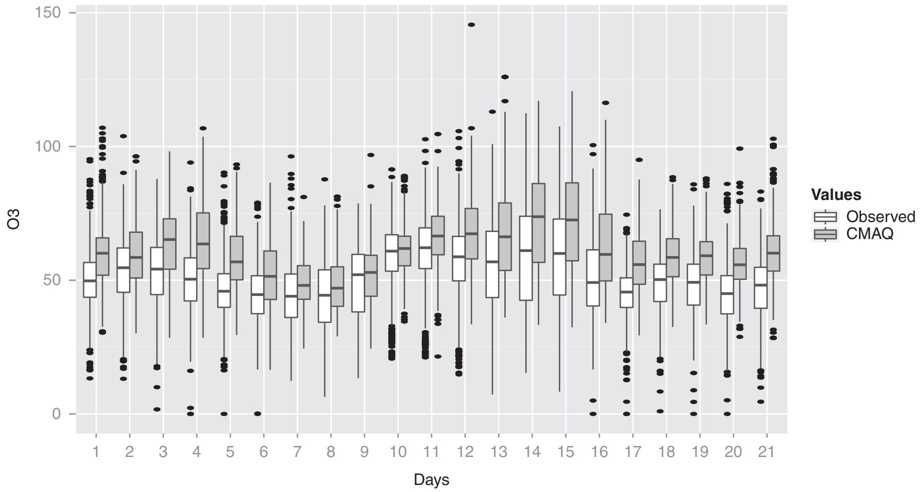


Figure 13.2 Side-by-side boxplots of the observed daily maximum ozone concentration levels and Eta CMAQ output for 21 days from all 639 sites in the eastern United States.

is not needed in the first place. MCMC methods enable us to estimate the uncertainties of the forecasts on the original scale.

Let $O(\mathbf{s}_i, t)$ be the true value corresponding to $Z(\mathbf{s}_i, t)$ at site $\mathbf{s}_i$, $i = 1, \ldots, n$ at time t. Let $\mathbf{Z}_t = (Z(\mathbf{s}_1, t), \ldots, Z(\mathbf{s}_n, t))\prime$ and $\mathbf{O}_t = (O(\mathbf{s}_1, t), \ldots, O(\mathbf{s}_n, t))\prime$. We shall denote that all the observed data by $\mathbf{z}$, and $\mathbf{z}^*$ will denote all the missing data. Similarly, $\mathbf{O}$ will denote all $\mathbf{O}_t$, for $t = 1, \ldots, T$. Let $N = nT$ be the total number of observations to be modeled.

For forecasting purposes, it is of interest to obtain the one-step ahead forecast distribution for noisy data $Y(\mathbf{s}_0, T + 1)$ on the original scale, and not for $O(\mathbf{s}_0, T + 1)$, since our objective is to compare the forecasting methods by validation of the noisy data itself, where $\mathbf{s}_0$ denotes any particular, monitored or unmonitored, site of interest. In the sequel, we shall obtain the marginal one-step ahead forecasts at a number of sites, say m. The joint one-step ahead forecast distribution for the m forecasts can also be developed for the models described subsequently, but are not of interest here.

We also assume that, in general, there are p covariates, including the intercept, denoted by the $n \times p$ matrix $\mathbf{X}_t$. Some of these covariates may vary in both space and time. The notation $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)\prime$ will be used to denote the $p \times 1$ vector of regression coefficients. In this chapter, we do not allow $\boldsymbol{\beta}$ to be dynamic, but it is possible to incorporate the dynamic models along the lines suggested by Mardia et al. (1998), and this will be considered elsewhere. We shall use the generic notation $\boldsymbol{\theta}$ to denote all the parameters.

13.3.2 Forecasting using GP models

The spatiotemporal linear regression model is defined by:

$$\mathbf{Z}_t = \mathbf{O}_t + \boldsymbol{\epsilon}_t, \tag{13.1}$$

$$\mathbf{O}_t = \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\eta}_t, \tag{13.2}$$

where $\boldsymbol{\epsilon}_t = (\epsilon(\mathbf{s}_1, t), \ldots, \epsilon(\mathbf{s}_n, t))\prime \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$ is the independently distributed white noise error with variance σ_ϵ^2 also known as the nugget effect, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. The term $\boldsymbol{\eta}_t = (\eta(\mathbf{s}_1, t), \ldots, \eta(\mathbf{s}_n, t))\prime$ is an independent, over time, realization of a GP with zero mean and the correlation function $\kappa(d; \phi, \nu)$, often assumed to be a member of the Matérn family; see for example, Banerjee et al. (2004), is allowed to depend on two unknown parameters ϕ and ν describing the correlation at distance d. In effect, this implies that the smooth process, $O(\mathbf{s}, t)$ is assumed to be isotropic and stationary. Note that this does not necessarily imply the same assumptions for the untransformed noisy data, Y since other hierarchical model components will contribute to the overall space–time correlation function.

Thus, we assume that $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \Sigma_\eta)$, where $\Sigma_\eta = \sigma_\eta^2 S_\eta$ and $(S_\eta)_{ij} = \kappa(||\mathbf{s}_i - \mathbf{s}_j||; \phi, \nu),\ i, j = 1, \ldots, n;\ \sigma_\eta^2$ is the site invariant common variance and $\kappa(.; \phi, \nu)$ is the spatial correlation that depends on spatial decay, ϕ, and smoothness, ν, parameters. For convenience, in this chapter, we use the exponential covariance function to model spatial dependence as

$$\Sigma_\eta = \sigma_\eta^2 S_\eta = \sigma_\eta^2 \exp(-\phi_\eta D),$$

where $\phi_\eta > 0$ is a spatial correlation decay parameter, and D is the matrix that has elements d_{ij}, that is the distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$, $i, j = 1, \ldots, n$. Here, and in the sequel, the matrix exponential is used to mean element-wise exponentiation, that is, $(\Sigma_\eta)_{ij} = \sigma_\eta^2 \exp(-\phi_\eta d_{ij}),\ i, j = 1, \ldots, n.$ The `spTimer` package provides options to implement using the full Matérn family. The error distributions of $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are assumed to be independent of each other. For future reference, let $\boldsymbol{\theta}$ denote all the parameters, $\boldsymbol{\beta}$, σ_ϵ^2, σ_η^2, and ϕ. We assume independent normal prior distribution with zero mean and a very large variance, 10^{10}, to achieve vague prior specification, for the components of $\boldsymbol{\beta}$. The inverse of the variance components σ_ϵ^2, σ_η^2 are given independent gamma distribution with mean a/b and variance a/b^2. Although any suitable values for a and b can be chosen, following Sahu et al. (2007) we have taken $a = 2$ and $b = 1$ to have a proper prior distribution for any variance component that will guarantee a proper posterior distribution. We assume discrete uniform prior distributions for the correlation parameters ϕ and ν, although many other choices are possible. Full details are provided in the `spTimer` package; see Bakar and Sahu (2014).

To obtain the one-step ahead forecast distribution of $Z(\mathbf{s}_0, T+1)$ at any unobserved location $\mathbf{s}_0$ at time $T+1$, we first note that

$$\begin{aligned} Z(\mathbf{s}_0, T+1) &= O(\mathbf{s}_0, T+1) + \epsilon(\mathbf{s}_0, T+1), \\ O(\mathbf{s}_0, T+1) &= \mathbf{x}'(\mathbf{s}_0, T+1)\boldsymbol{\beta} + \eta(\mathbf{s}_0, T+1). \end{aligned}$$

The one-step ahead forecast distribution is the posterior predictive distribution of $Z(\mathbf{s}_0, T+1)$ given $\mathbf{z}$ and is given by

$$\begin{aligned} \pi(Z(\mathbf{s}_0, T+1)|\mathbf{z}) = \int & \pi(Z(\mathbf{s}_0, T+1)|\boldsymbol{\theta}, \mathbf{O}, O(\mathbf{s}_0, T+1), \mathbf{z})\pi(O(\mathbf{s}_0, T+1)|\boldsymbol{\theta}, \mathbf{z}) \\ & \pi(\boldsymbol{\theta}, \mathbf{O}|\mathbf{z}) dO(\mathbf{s}_0, T+1) d\mathbf{O}\, d\boldsymbol{\theta}, \end{aligned} \tag{13.3}$$

where $\pi(\boldsymbol{\theta}, \mathbf{O}|\mathbf{z})$ denotes the joint posterior distribution of $\mathbf{O}$ and $\boldsymbol{\theta}$. Note that $\pi(Z(\mathbf{s}_0, T+1)|\boldsymbol{\theta}, \mathbf{O}, O(\mathbf{s}_0, T+1), \mathbf{z}) = \pi(Z(\mathbf{s}_0, T+1)|\boldsymbol{\theta}, \mathbf{O}, O(\mathbf{s}_0, T+1))$ due to the conditional independence of $Z(\mathbf{s}_0, T+1)$ and $\mathbf{Z}$ given $\mathbf{O}$. Similarly, $O(\mathbf{s}_0, T+1)$ does not depend on $\mathbf{Z}$ given $\boldsymbol{\theta}$, hence in the following development we replace $\pi(O(\mathbf{s}_0, T+1)|\boldsymbol{\theta}, \mathbf{z})$ by $\pi(O(\mathbf{s}_0, T+1)|\boldsymbol{\theta})$.

Now the one-step ahead forecast distribution (13.3) is constructed by composition sampling as follows. Assume that, at the jth MCMC iteration, we have posterior samples, $\boldsymbol{\theta}^{(j)}$ and $\mathbf{O}^{(j)}$. Then we first draw, $O^{(j)}(\mathbf{s}_0, T+1)$ from $N(\mathbf{x}\prime_{T+1}\boldsymbol{\beta}^{(j)}, \sigma_\eta^{2(j)})$. Finally, we draw $Z^{(j)}(\mathbf{s}_0, T+1)$ from $N(O^{(j)}(\mathbf{s}_0, T+1), \sigma_\epsilon^{2(j)})$.

Note that in the aforementioned paragraph, we use the marginal distribution instead of the conditional distribution because we have already obtained the conditional distribution given observed information up to time T at the observation locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, and at the future time $T+1$ there is no further new information to condition on except for the new regressor values $\mathbf{x}(\mathbf{s}_0, T+1)$ in the model. However, the conditional distribution can be used instead if it is so desired. To do this, we note that the joint distribution of $\mathbf{O}_{T+1} = (O(\mathbf{s}_1, T+1), \ldots, O(\mathbf{s}_n, T+1))\prime$ is simply given by $N(\mathbf{X}_{T+1}\boldsymbol{\beta}, \Sigma_\eta)$, according to (13.2). Similarly, we construct the joint distribution of $O(\mathbf{s}_0, T+1)$ and $\mathbf{O}_{T+1}$ from which we obtain the conditional distribution $\pi(O(\mathbf{s}_0, T+1)|\mathbf{O}_{T+1})$, that is Gaussian with mean

$$\mathbf{x}(\mathbf{s}_0, T+1)\boldsymbol{\beta} + S_{\eta,12}S_\eta^{-1}(\mathbf{O}_{T+1} - \mathbf{X}_{T+1}\boldsymbol{\beta})$$

and variance

$$\sigma_\eta^2(1 - S_{\eta,12}S_\eta^{-1}S_{\eta,21}),$$

where $S_{\eta,21}\prime = S_{\eta,12} = e^{-\phi\,\mathbf{d}_{12}}$ and $\mathbf{d}_{12} = (||\mathbf{s}_1 - \mathbf{s}_0||, \ldots, ||\mathbf{s}_n - \mathbf{s}_0||)\prime$.

For forecasting at any observed site $\mathbf{s}_i$ for any $i = 1, \ldots, n$ at time $T+1$ we note that

$$Z(\mathbf{s}_i, T+1) = O(\mathbf{s}_i, T+1) + \epsilon(\mathbf{s}_i, T+1),$$
$$O(\mathbf{s}_i, T+1) = \mathbf{x}\prime(\mathbf{s}_i, T+1)\boldsymbol{\beta} + \eta(\mathbf{s}_i, T+1).$$

These two identities make it clear that the one-step ahead forecast distribution of $Z(\mathbf{s}_i, T+1)$ given $\mathbf{z}$ can simply be constructed by iteratively sampling from the conditional distribution $O^{(j)}(\mathbf{s}_i, T+1) \sim N(\mathbf{x}\prime(\mathbf{s}_i, T+1)\boldsymbol{\beta}^{(j)}, \sigma_\eta^{2(j)})$ and then $Z^{(j)}(\mathbf{s}_i, T+1)$ from the normal distribution with mean $O^{(j)}(\mathbf{s}_i, T+1)$ and variance $\sigma_\epsilon^{2(j)}$. Finally, $Z^{(j)}(\mathbf{s}_i, T+1)$ values are transformed back to the original scale giving MCMC samples $Y^{(j)}(\mathbf{s}_i, T+1)$.

13.3.3 Forecasting using AR models

In this section, we briefly describe the forecasting method based on the hierarchical AR models proposed by Sahu et al. (2007; 2009). The model equations are given by

$$\mathbf{Z}_t = \mathbf{O}_t + \boldsymbol{\epsilon}_t, \tag{13.4}$$
$$\mathbf{O}_t = \rho\mathbf{O}_{t-1} + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\eta}_t, \tag{13.5}$$

where $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ have been previously specified, and ρ is a scalar denoting site-invariant temporal correlation. These auto-regressive models also need an initialization for $\mathbf{O}_0$ which we assume to be independently normally distributed with mean $\boldsymbol{\mu}$ and the covariance matrix $\sigma^2 S_0$, where the correlation matrix S_0 is obtained using the exponential correlation function with a new decay parameter ϕ_0. These additional parameters and initialization random variables are added to $\boldsymbol{\theta}$ and $\mathbf{O}$, respectively.

The temporal correlation, ρ in (13.5), for the smooth process $O(\mathbf{s}, t)$, has been assumed to be site invariant given the effects of the spatially and temporally varying covariates and the spatiotemporal intercepts $\eta(\mathbf{s}, t)$. A site-specific temporal correlation will perhaps be needed, though not pursued here, if only the last two terms are omitted from the model. We also assume, for stationarity, that $|\rho| < 1$.

We assume the same set of prior distributions for $\boldsymbol{\beta}$, the variance components σ_ϵ^2 and σ_η^2, and the correlation decay parameters ϕ as previously discussed in Section 13.3.2. For the additional ρ parameter, we again provide a normal prior distribution with zero mean and a large variance (10^{10} in our implementation), but we restrict the prior distribution in the range $|\rho| < 1$.

Under the AR models, the predictive distribution of $Z(\mathbf{s}_0, T+1)$ is determined by $O(\mathbf{s}_0, T+1)$. Following (13.5), we see that $O(\mathbf{s}_0, T+1)$ follows the normal distribution with site-invariant variance σ_η^2 and mean $\rho O(\mathbf{s}_0, T) + \mathbf{x}\prime(\mathbf{s}_0, T+1)\boldsymbol{\beta}$. This depends on $O(\mathbf{s}_0, T)$, and as a result, due to this auto-regressive nature, we have to determine all the random variables $O(\mathbf{s}_0, k)$, for $k = 0, \ldots, T$. In order to simulate, all these random variables, we first simulate from the conditional distribution of $O(\mathbf{s}_0, 0)$ given $\mathbf{O}_0$, which is a univariate normal distribution. Then, at the jth MCMC iteration, we sequentially simulate $O^{(j)}(\mathbf{s}_0, k)$ given $O^{(j)}(\mathbf{s}_0, k-1)$ for $k = 1, \ldots, T+1$ from the normal distribution with mean $\rho^{(j)} O^{(j)}(\mathbf{s}_0, k-1) + \mathbf{x}'(\mathbf{s}_0, k)\boldsymbol{\beta}^{(j)}$ and variance $\sigma_\eta^{2(j)}$. For forecasting at any observation location $\mathbf{s}_i$, we draw $Z^{(j)}(\mathbf{s}_i, T+1)$ from the normal distribution with mean $\rho^{(j)} O^{(j)}(\mathbf{s}_i, T) + \mathbf{x}'(\mathbf{s}, T+1)\boldsymbol{\beta}^{(j)}$ and variance $\sigma_\epsilon^{2(j)}$. For further details regarding prediction, see Sahu et al. (2007). Now these Z values are transformed back to the original scale, Y as in the case of GP models.

13.3.4 Forecasting using the GPP models

The models described in Section 13.3.3 assume the AR model for the true values of the modeled response $\mathbf{O}_t$. Sahu and Bakar (2012b) modified this model so that the modified version does not assume a true level $O(\mathbf{s}_i, t)$ for each $Z(\mathbf{s}_i, t)$ but instead assumes a space–time random-effect denoted by $\eta(\mathbf{s}_i, t)$. It then assumes an AR model for these space–time random effects. For a large number of spatial locations, the top level space–time random-effect term will lead to the estimation problem discussed in Introduction. Hence, we use the predictive process approximation technique (Sahu and Bakar 2012b). Here the main idea is to define the random effects $\eta(\mathbf{s}_i, t)$ at a smaller number of locations, m say, where $m \ll n$, called the knots, and then use kriging to predict those random effects at the data locations.

The top level model is written as follows:

$$\mathbf{Z}_t = \mathbf{X}_t\boldsymbol{\beta} + \tilde{\boldsymbol{\eta}}_t + \boldsymbol{\epsilon}_t, \; t = 1, \ldots, T, \tag{13.6}$$

where $\boldsymbol{\epsilon}_t$ has been previously specified. The space–time process $\tilde{\boldsymbol{\eta}}_t$ is specified by

$$\tilde{\boldsymbol{\eta}}_t = A\mathbf{w}_t \tag{13.7}$$

with $A = CS_w^{-1}$, where S_w is the correlation matrix of w_t with ijth element, which corresponds to two locations $\mathbf{s}_i$ and $\mathbf{s}_j$, is given by $\exp(-\phi_w||\mathbf{s}_i - \mathbf{s}_j||)$. The elements of the $n \times m$ matrix C are also calculated using this correlation function.

In the next stage of the modeling hierarchy, the AR model is assumed as

$$\mathbf{w}_t = \rho\, \mathbf{w}_{t-1} + \boldsymbol{\xi}_t, \tag{13.8}$$

where $\boldsymbol{\xi}_t \sim N(\mathbf{0}, \sigma_w^2 S_w)$. Again, we assume that $\mathbf{w}_0 \sim N(\mathbf{0}, \sigma^2 S_0)$, where the elements of the covariance matrix S_0 are obtained using the correlation function, $\exp(-\phi_0 d_{ij})$, which is the same correlation function used previously but with a different decay parameter ϕ_0. The Bayesian model specification here is completed by assuming the same set of prior distributions as noted in the previous two subsections.

At an unobserved location $\mathbf{s}_0$, the one-step ahead Bayesian forecast is given by the predictive distribution of $Z(\mathbf{s}_0, T+1)$, which we determine from equation (13.6) replacing t with $T+1$. Thus, the one-step ahead forecast distribution has variance σ_ϵ^2 and mean $\mathbf{x}'(\mathbf{s}_0, T+1)\boldsymbol{\beta} + \tilde{\eta}(\mathbf{s}_0, T+1)$, where $\tilde{\eta}(\mathbf{s}_0, T+1)$ is obtained analogous to (13.7) as

$$\tilde{\eta}(\mathbf{s}_0, T+1) = S_{w,12} S_w^{-1} \mathbf{w}_{T+1},$$

where $S_{w,12} = e^{-\phi_w \mathbf{d}_{12}}$ and $\mathbf{w}_{T+1}$ is obtained from (13.8).

Thus, at each MCMC iteration, we draw a forecast value $Z^{(j)}(\mathbf{s}_0, T+1)$ from this normal distribution. Forecasting at the observation sites $\mathbf{s}_1, \ldots, \mathbf{s}_n$ is performed by noting that, according to (13.6),

$$\mathbf{Z}_{T+1} = \mathbf{X}_{T+1}\boldsymbol{\beta} + \tilde{\boldsymbol{\eta}}_{T+1} + \boldsymbol{\epsilon}_{T+1},$$

with $\tilde{\boldsymbol{\eta}}_{T+1} = A\mathbf{w}_{T+1}$ and $\boldsymbol{\epsilon}_{T+1} \sim N(\mathbf{0}, \sigma_\epsilon^2 I_n)$. Thus, as before $\mathbf{w}_{T+1}$ is obtained from (13.8) and MCMC sampling from the forecast distribution of $Z(\mathbf{s}_i, T+1)$ for $i = 1, \ldots, n$ is straightforward. Again these Z samples are transformed back to the original scale Y, which we use for forecast calibration purposes.

13.4 Forecast calibration methods

The three model-based forecasting methods discussed in the previous section must be compared using suitable methods. Predictive Bayesian model selection methods are appropriate for comparing Bayesian models; see for example, Gelfand and Ghosh (1998). However, the main objective of this chapter is forecasting, and hence we compare the models on the basis of their forecasting performance. There is a large literature on forecast comparison and calibration methods; see for example, Gneiting et al. (2007) and the references therein. In the Bayesian context of this chapter, we need to compare the entire forecast predictive distribution, not just summaries such as the mean, since forecasting is the primary goal here.

To simplify notation, suppose that $y_i, i = 1, \ldots, m$ denote the m hold-out validation observations that have not been used in model fitting. Note that we use a single indexed notation y_i, instead of the more elaborate $y(\mathbf{s}, t)$ used previously. Clearly, some of these validation observations may be future observations at the modeling sites or completely at new sites – what's important here is that those must not have been used for model fitting. Let $F_i(y)$ denote the model-based forecast predictive distribution function of Y_i, the random variable whose realized value is y_i. Thus, $F_i(y)$ is one of the three forecast predictive distributions, corresponding to one of the three models: GP, AR, and GPP, described previously in Section 13.3. Let $G_i(y)$ be the true unknown forecast predictive distribution function, which the $F_i(y)$ is trying to estimate. The problem here is to calibrate $F_i(y)$ for $G_i(y)$,

$i = 1, \ldots, m$, conditional on the modeled data, $\mathbf{y}$ or equivalently its transformed value $\mathbf{z}$. Let $\hat{y}_i$ be the intended forecast for y_i, that is, $\hat{y}_i$ mean or median of the forecast distribution $F_i(y)$, estimated using the mean or median of the MCMC samples $y_i^{(j)}, j = 1, \ldots, J$, where J is a large number. In our implementation in Sections 13.5 and 13.6, we have taken $J = 15,000$ after discarding the first 5000 iterations; that was deemed to be adequate to mitigate the effect of initial values. Subsequently, we describe seven important forecast calibration and diagnostic methods and develop their computation methods by using MCMC.

1. The RMSE is defined by

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}.$$

It is perhaps the most popular forecast comparison criterion and the method with the smallest RMSE value is preferred.

2. Sometimes the MAE, defined by,

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

is preferred to the RMSE. Both the RMSE and the MAE are on the original unit of the data, and they provide a quick check on the magnitude of the errors in forecasts.

3. The CRPS is a proper scoring rule for comparing forecasts, (Gneiting et al. 2007) and is defined by

$$\text{crps}(F, y) = E_F|Y - y| - \frac{1}{2}E_F|Y - Y'|$$

where Y and $Y\prime$ are independent copies of a random variable with distribution function F and finite first moment. With m hold-out observations, we calculate the overall measure, given by

$$\text{CRPS} = \frac{1}{m}\sum_{i=1}^{m}\text{crps}(F_i, y_i).$$

We estimate the CRPS using J MCMC samples $y_i^{(j)}, j = 1, \ldots, J$, as follows. We first obtain,

$$\hat{\text{crps}}(F_i, y_i) = \frac{1}{J}\sum_{j=1}^{J}|y_i^{(j)} - y_i| - \frac{1}{2J^2}\sum_{j=1}^{J}\sum_{k=1}^{J}|y_i^{(j)} - y_i^{(k)}|,\ i = 1, \ldots, m,$$

and then the overall average CRPS is estimated as:

$$\hat{\text{CRPS}} = \frac{1}{m}\sum_{i=1}^{m}\hat{\text{crps}}(F_i, y_i).$$

Again, the model with the smallest CRPS value is the preferred choice.

4. The empirical coverage is defined by

$$\text{ECOV} = \frac{1}{m}\sum_{i=1}^{m} 1\left(l_i \leq y_i \leq u_i\right),$$

where l_i and u_i are, respectively, the lower and upper limits of a given predictive interval for y_i and $1(A) = 1$ if A is true and 0 otherwise. Good forecasting methods must have the empirical coverage close to their true value so that the uncertainties in the forecast distributions are correct, not only their central tendencies as measured by the RMSE or the MAE. In practice, the limits l_i and u_i are estimated using the appropriate quantiles of the MCMC samples $y_i^{(j)}, j = 1, \ldots, J$. For example, for 95% prediction intervals, these are estimated to be the 2.5th and 97.5th percentile of $y_i^{(j)}, j = 1, \ldots, J$, respectively.

5. The concentration of the forecast distribution is compared using the sharpness diagram. A sharpness diagram plots the widths of the (m) forecast intervals as side-by-side boxplots, where each boxplot is for a particular forecasting method. The forecasting method that produces narrower width forecast intervals, but with good empirical coverages, is preferred.

6. The hit and false alarm rates are also considered by many authors for forecast comparison purposes, see for example, Sahu et al. (2009). These rates are defined for a given threshold value y_0, which is often the value beyond which the pollutant is considered to be very dangerous. Hit is defined as the event where both the validation observation, y_i and the forecast, $\hat{y}_i$, for it are either greater or less than the threshold y_0. The false alarm, on the other hand, is defined as the event where the actual observation is less than y_0 but the forecast is greater than y_0. Thus, we define:

$$\text{Hit rate}(y_0) = \frac{1}{m}\sum_{i=1}^{m}\left\{1\left(y_i > y_0 \,\&\, \hat{y}_i > y_0\right) + 1\left(y_i < y_0 \,\&\, \hat{y}_i < y_0\right)\right\},$$

$$\text{False alarm}(y_0) = \frac{1}{m}\sum_{i=1}^{m} 1(y_i < y_0 \,\&\, \hat{y}_i > y_0).$$

Forecasting methods with high hit rates and low false alarm rates are preferred.

7. Many authors have proposed the PIT diagram as a necessary diagnostic tool for comparing forecasts. For each hold-out observation y_i, the PIT value is calculated as

$$p_i = F_i(y_i), i = 1, \ldots, m.$$

If forecasts are ideal, and F_i is continuous, then p_i has a uniform distribution. The PIT diagram is simply an histogram of the p_i's, $1, \ldots, m$. Using MCMC samples, p_i is estimated by:

$$\hat{p}_i = \frac{1}{J}\sum_{j=1}^{J} 1\left(y_i^{(j)} \leq y_i\right), \; i = 1, \ldots, m.$$

8. A MCP is used to calibrate the equality of the forecast and the actual value and is constructed as follows. First, take a grid, $y_k, k = 1, \ldots, K$, say, covering the domain of the forecast distribution. For each of those y_k values, calculate

$$\hat{G}(y_k) = \frac{1}{m} \sum_{i=1}^{m} 1\,(y_i \leq y_k)\,.$$

Now calculate

$$\bar{F}(y_k) = \frac{1}{m} \sum_{i=1}^{m} \hat{F}_i(y_k),$$

where

$$\hat{F}_i(y_k) = \frac{1}{J} \sum_{j=1}^{J} 1\left(y_i^{(j)} \leq y_k\right),\ i = 1, \ldots, m.$$

Now, the plot of the differences $\bar{F}(y_k) - \hat{G}(y_k)$ against y_k, for $k = 1, \ldots, K$ is the desired MCP. If the forecasts are good, only minor fluctuations about 0 are expected. Thus, a forecast distribution whose MCP stays closest to 0 will be the preferred choice.

13.5 Results from a smaller data set

The computation of all the forecast calibration methods for the whole eastern US data set is prohibitive because of the big-*n* problem as mentioned in Introduction; see also the next section. Due to this reason, we compare all three forecasting methods using a subset of the whole eastern US data, consisting of four states: Illinois, Indiana, Ohio, and Kentucky. There are 147 ozone monitoring sites in these states; see Figure 13.3. We set aside data from 20 randomly selected sites for validation purposes. As mentioned in Section 13.2, we validate for seven days from July 8 to 14.

For the GPP model, the knot size is taken as 107, which has been chosen from a sensitivity analysis similar to the ones reported in Sahu and Bakar (2012b). We have also performed a number of different sensitivity analysis with respect to the choice of the hyper-parameter

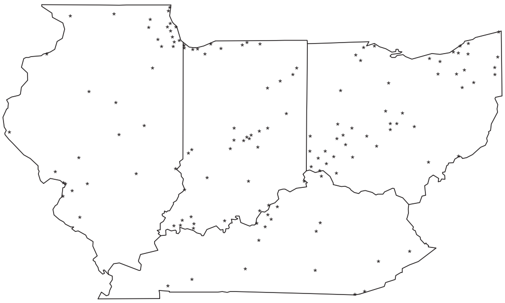


Figure 13.3 Map of the four states, Ohio, Indiana, Illinois, and Kentucky. A total of 147 ozone monitoring locations are superimposed.

values in the prior distribution, tuning of the MCMC algorithms and have also monitored convergence using trace plots and the package CODA (Plummer et al. 2006). We omit all those details for brevity.

All three models are fitted using the MCMC code developed within the `spTimer` package. As mentioned in Section 13.4, MCMC algorithms are run for a total of 20,000 iterations of which first 5000 are discarded to mitigate the effect of starting values. The algorithms run very fast taking only about 9, 16, and 3 minutes for the GP, AR, and GPP models, respectively, in a 2.6 GHz personal computer with 4GB of RAM running 32 bit Windows operating system. Thus, it is quite fast to fit the models and produce forecasts using all the models.

The RMSE and the MAE for the seven validation days are plotted in Figure 13.4. As expected, the RMSE and the MAE are very similar (compare the columns). But we do not see a large difference between modeling 7 and 14 days data (compare the rows). The RMSE and MAE of the GP and AR models are very similar, and they both have worse performance than the GPP model. This is also confirmed by the CRPS values; see Table 13.2. The actual coverages, of the 50% and 95% forecast intervals, provided in Table 13.3, however, are not able to compare the forecasting methods; but those show that all three methods are adequate. The average widths of the forecast intervals, see Table 13.4, clearly show that the GPP model is the best. This is also confirmed by the sharpness diagram; see Figure 13.5.

The hit and false alarm rates using all seven validation days data are provided in Table 13.5. All three models perform very well. The hit rate increases as the threshold value increases, and it is actually 100% when we use the threshold value of 85. The false alarm rate decreases to zero as

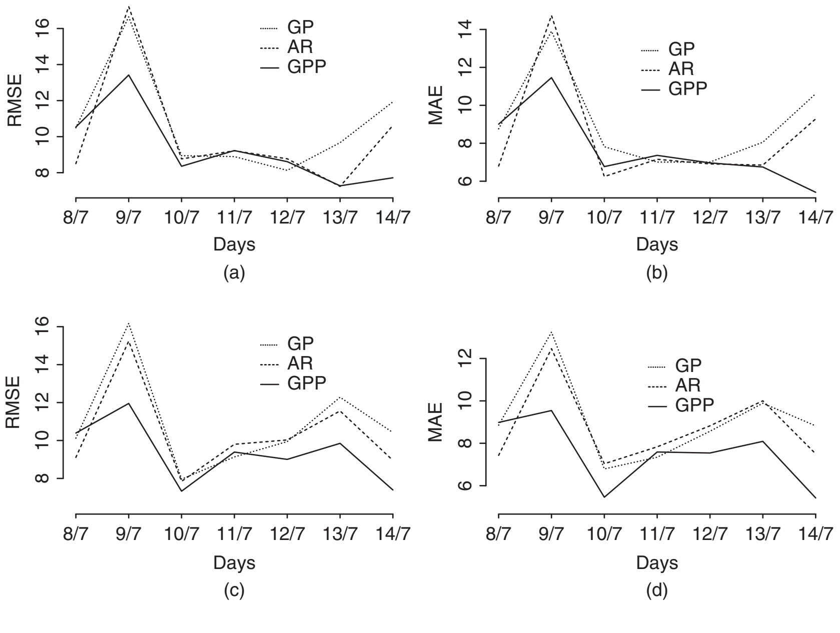


Figure 13.4 Plots of RMSE and MAE based on modeling 7 days data (a and b) and 14 days data (c and d).

Table 13.2 CRPS values from modeling data from four states during July 8 (denoted as 7/8) to 14.

	Values from modeling 7 days data							
Models	7/8	7/9	7/10	7/11	7/12	7/13	7/14	7/(8–14)
GP	6.12	10.22	5.04	5.05	4.78	5.70	6.95	6.27
AR	6.19	10.12	4.95	5.31	4.85	4.38	4.31	5.73
GPP	4.95	10.02	4.89	5.33	4.87	4.33	4.13	5.52
	Values from modeling 14 days data							
GP	6.14	9.82	5.33	5.42	5.21	5.64	6.29	6.27
AR	5.91	9.83	4.56	5.27	5.19	4.43	5.90	5.87
GPP	5.32	9.56	4.37	5.30	5.15	4.28	5.26	5.60

Table 13.3 Empirical coverages of the 50% and 95% forecast intervals for the one-step ahead forecasts at the 20 randomly chosen validation sites.

	Intervals			
	Using 7 days data		Using 14 days data	
Models	50%	95%	50%	95%
GP	51.43	95.71	55.00	95.71
AR	50.71	94.29	50.71	93.43
GPP	50.71	94.95	49.71	94.00

Table 13.4 Average width of the forecast intervals for the four states data set.

	Using 7 days data		Using 14 days data	
Models	50%	90%	50%	90%
GP	12.76	30.95	12.57	30.69
AR	13.51	32.95	13.36	32.28
GPP	11.54	28.11	9.58	23.47

the threshold value is increased from 65 to 75 ppb. These rates, however, do not discriminate between the three different forecasting methods.

The PIT diagrams for all three forecasting methods for the 14 days data modeling case are provided in Figure 13.6. Here also the GPP model is the preferred choice since its histogram is more uniform than the other two. The same diagrams based on modeling 7 days data showed similar patterns and hence have been omitted.

Figure 13.7 provides the MCPs of all three models using data for 7 and 14 days. Here also the GPP model performs better than its rivals, and the performance is differentiated better in the case of modeling data for 14 days. In addition, calibration improves toward the upper tail of the distribution that assures that the models are able to forecast high levels of ozone concentration quite accurately. In conclusion, we find that the GPP model is the best for forecasting among the three methods considered here.

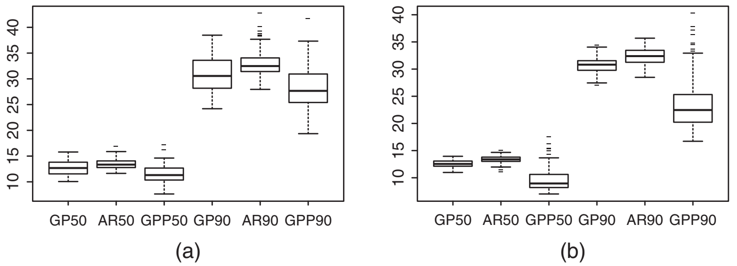


Figure 13.5 Sharpness diagram using: (a) 7 days data (b) 14 days data.

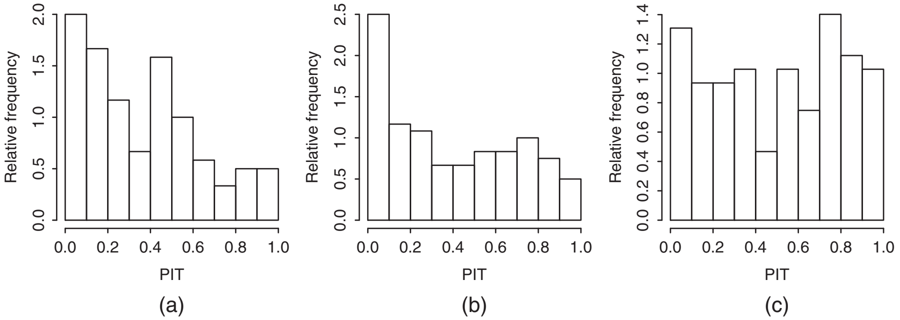


Figure 13.6 PIT diagrams for (a) GP, (b) AR, and (c) GPP models using 14 days data for modeling.

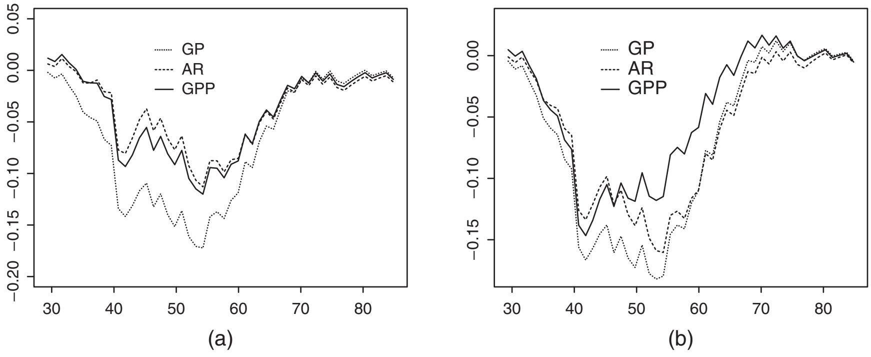


Figure 13.7 Marginal calibration plots for all the models using (a) 7 days data (b) 14 days data for modeling.

A further remark regarding the performances of the AR and GPP models is appropriate. As with any approximation, it can be expected that the approximate GPP model to perform worse than the full AR model. However, the GPP model in Section 13.3.4 cannot be seen as a true approximation for the AR model in Section 13.3.3 due to the inclusion of the auto-regressive term in two very different manners: one at the top level $\mathbf{O}_t$ in (13.5) and the other at the random-effect level $\mathbf{w}_t$ in (13.7). Thus, the AR and GPP models are very

Table 13.5 False alarm and hit rates for ozone threshold values of 65 and 75 for the four states data set.

		Using 7 days data		Using 14 days data	
Ozone levels	Model	False alarm	Hit rate	False alarm	Hit rate
65 ppb	GP	0.92	91.67	0.92	91.67
	AR	4.59	92.50	1.83	92.50
	GPP	3.67	91.67	2.75	91.67
75 ppb	GP	0.0	95.83	0.0	95.83
	AR	0.0	95.83	0.0	95.83
	GPP	0.0	96.67	0.0	97.50

different, and it is not surprising that we do not see any strict one-way performance ordering in our examples.

13.6 Analysis of the full Eastern US data set

As mentioned in Section 13.2, we use data from 577 sites to fit our models and the data from 62 sites are set aside for validation purposes. The implementation of the GPP model requires the selection of the number of knots. Using a similar sensitivity study that we have used in Sahu and Bakar (2012b), but with the forecast RMSE, as the criterion we compare the GPP model with 68, 105, 156, and 269 knots, which were all inside the land boundary

Table 13.6 Parameter estimates (mean and SD) for the models based on GPP approximation fitted with 14 days observations for the period June 24 (denoted as 6/24) to July 13, 2010 from the 577 modeling sites in the whole Eastern United States.

Fitted days		β_0	β_1	ρ	σ^2_ϵ	σ^2_w	ϕ
6/24–7/7	Mean	4.13	0.37	0.40	0.24	0.49	0.0046
	SD	0.20	0.03	0.04	0.005	0.04	0.0005
6/25–7/8	Mean	4.34	0.36	0.39	0.25	0.53	0.0042
	SD	0.23	0.02	0.03	0.004	0.04	0.0004
6/26–7/9	Mean	4.68	0.33	0.39	0.25	0.57	0.0041
	SD	0.33	0.03	0.04	0.006	0.05	0.0007
6/27–7/10	Mean	3.40	0.33	0.39	0.25	0.52	0.0046
	SD	0.22	0.03	0.04	0.005	0.04	0.0005
6/28–7/11	Mean	4.74	0.31	0.35	0.25	0.60	0.0031
	SD	0.17	0.02	0.04	0.005	0.05	0.0007
6/29–7/12	Mean	4.66	0.31	0.36	0.25	0.54	0.0037
	SD	0.20	0.02	0.04	0.005	0.04	0.0003
6/30–7/13	Mean	4.92	0.29	0.35	0.26	0.60	0.0032
	SD	0.30	0.03	0.04	0.005	0.07	0.0006

of the United States. The forecast RMSE improved with the increasing knot sizes, but only slightly when the size increased to 269 from 156. Henceforth, we adopt 156 as the knot size that implies a much smaller computational burden.

For the model fitting (a data set with 14 days data) and forecasting using 20,000 iterations, using the same personal computer as in the previous section, we have estimated that the GP model will take about 40 hours, while the AR model will take about 66 hours to run. This excludes the use of GP and AR models for forecasting next day ozone levels, which must be produced within 24 hours of computing time. The GPP model, on the other hand, takes only about 50 minutes to run the same experiment on the same personal computer and is the only feasible method that we henceforth adopt.

We compare the performance of the GPP model based with those obtained from a non-Bayesian linear regression model with the Eta CMAQ output as the only covariate, which is a simple method that does not require advanced modeling and computation techniques. We also illustrate parameter estimation and maps providing forecast surfaces.

We report the parameter estimates and their standard deviations in Table 13.6 for the model fitting cases with 14 days data. The estimates are broadly similar for different subsets of fitted data. The Eta CMAQ output always remains a significant predictor with very small standard deviation relative to the mean. The temporal correlation remained always near 20%. The random-effect variance σ_w^2 is always estimated to be larger than the nugget effect σ_ϵ^2. The estimate of the spatial decay parameter is 0.0024, which corresponds to an effective range of 1250 km. A similar table based on model fitting from 7 days data is omitted for brevity.

We now compare the GPP model-based forecasts with those from the linear regression model using the RMSEs based on validation data both from the 62 hold-out sites. The RMSE values, provided in Table 13.7, are smaller for the GPP model than the linear regression model. Moreover, the RMSE values are smaller when the forecasting model is trained with 14 days data than the same with 7 days data. The RMSE values for the forecasts made by the Eta CMAQ model are considerably higher, which justifies this additional statistical modeling effort.

The empirical coverages of the 95% forecast intervals, provided in Table 13.8, show that the uncertainty in the forecasts based on the GPP model is about right. However, the empirical coverages for the linear model-based forecasts are closer to 100%, which shows

Table 13.7 Values of the RMSE of the forecasts at the hold-out sites for the simple linear model and the GPP model based on modeling 7 and 14 days data for the whole of Eastern United States. The corresponding RMSE values for the Eta CMAQ output are also shown.

Forecast	CMAQ	7 days		14 days	
		Linear	GPP	Linear	GPP
7/8	20.52	12.16	10.34	10.97	10.30
7/9	19.68	12.25	10.79	11.59	10.04
7/10	16.36	9.87	8.59	9.49	8.13
7/11	15.51	8.55	8.17	8.69	7.98
7/12	13.12	8.99	8.67	8.44	8.17
7/13	20.36	12.70	10.85	13.95	9.83
7/14	18.10	9.64	9.20	10.25	9.05

Table 13.8 Empirical coverage of the 95% forecast intervals using the linear and GPP models and the CRPS values for the hold-out data for the GPP model for the whole Eastern US data set.

Forecast	7/8	7/9	7/10	7/11	7/12	7/13	7/14	7/(8–14)
Empirical coverage of the 95% forecast intervals using the linear model								
7 days	99.94	99.80	99.44	100.00	100.00	99.07	98.15	99.11
14 days	99.94	98.50	97.59	100.00	100.00	97.50	98.15	98.64
Empirical coverage of the 95% forecast intervals using the GPP model								
7 days	93.55	93.75	94.96	95.16	94.96	93.75	95.56	94.53
14 days	94.62	94.30	94.84	95.05	94.62	94.84	94.84	94.74
CRPS values								
7 days	10.05	7.98	6.52	6.79	7.12	7.18	7.11	7.54
14 days	9.43	7.25	5.89	6.80	6.93	6.94	6.74	7.15

that these forecast intervals are too wide and this method fails to reduce uncertainty in forecasts.

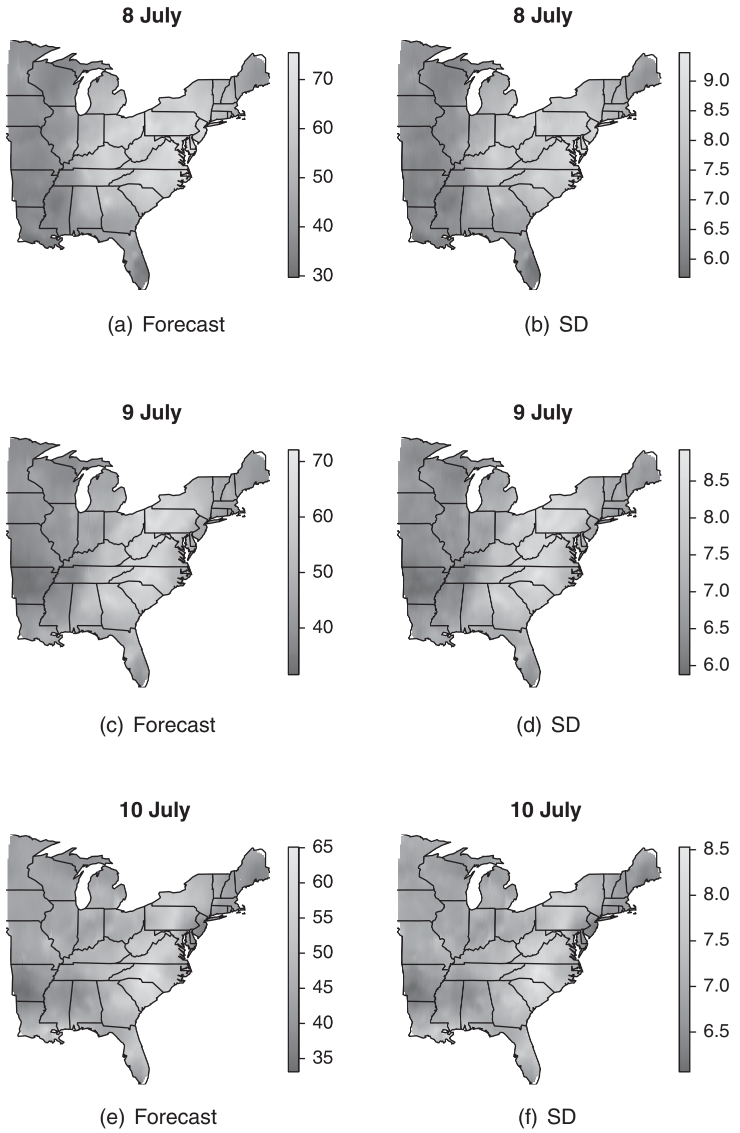
Table 13.8 also provides the CRPS values, which turn out to be slightly higher than the values presented in Table 13.2 for the four states data. This is not surprising since it is usually more difficult to extrapolate in larger spatial domains. We have also obtained the false alarm and hit rates of the forecasts from the GPP model, which are 0 and 95.33, respectively, when the threshold value is 75 ppb. Clearly, the GPP model is very accurate for forecasting, and hence, we do not consider the other diagnostics such as the PIT diagram and the MCPs. Instead, we proceed to illustrate the forecasts.

Figure 13.8 illustrates the forecast maps based on the GPP model along with their standard deviations for the three days, 8th, 9th, and 10th of July. Here, each forecast map has its own color scheme that enables us to show the full spatial variation of the forecasts. In addition, the maps of standard deviations reveal that higher ozone levels are associated with higher uncertainty levels, which is a common phenomenon in ozone concentration modeling.

13.7 Conclusion

This chapter has developed Bayesian forecasting methods using three recently published Bayesian hierarchical models for spatiotemporal data. MCMC methods have been developed to compute the Bayesian forecast distributions based on large space–time data. These methodological developments have enabled us to add the suite of forecasting routines in the contributed `R` software package, `spTimer` which is available from `CRAN` (http://cran.r-project.org/) and allows modeling of large space–time data sets.

The contribution of the chapter also includes development of methods for estimating several forecast calibration measures using output from the implemented Markov chain Monte Carlo algorithms. We have demonstrated that these measures are able to compare different Bayesian forecasting methods rigorously and conclusively. A forecasting method based on a space–time model developed using a GPP approximation has been shown to be fast and the best for the illustrative ozone concentration forecasting problem of the chapter.



8 July

8 July

(a) Forecast

(b) SD

9 July

9 July

(c) Forecast

(d) SD

10 July

10 July

(e) Forecast

(f) SD

Figure 13.8 Maps showing the forecasts and their standard deviations for July 8, 9 and 10 in 2010. Observed ozone levels are also superimposed on the forecast maps from a selected number of sites only, to avoid clutter.

References

Bakar KS and Sahu SK 2014 `spTimer`: spatio-temporal Bayesian modelling using R. *Journal of Statistical Software* **63**(15), 1–32.

Banerjee S, Carlin BP and Gelfand AE 2004 *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton, FL.

Banerjee S, Gelfand AE, Finley AO and Sang H 2008 Gaussian predictive process models for large spatial data sets. *Journal of Royal Statistical Society, Series B* **70**, 825–848.

Bauer G, Deistler M and Scherrer W 2001 Time series models for short term forecasting of ozone in the eastern part of Austria. *Environmetrics* **12**, 117–130.

Berrocal VJ, Gelfand AE and Holland DM 2010a A bivariate space-time downscaler under space and time misalignment. *Annals of Applied Statistics* **4**, 1942–1975.

Berrocal VJ, Gelfand AE and Holland DM 2010b A spatio-temporal downscaler for outputs from numerical models. *Journal of Agricultural, Biological and Environmental Statistics* **15**, 176–197.

Cameletti M, Ignaccolo R and Bande S 2011 Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics* **22**, 985–996.

Ching J ang Byun D 1999 Science algorithms of the EPA models-3 community multi-scale air quality (CMAQ) modeling system. National Exposure Research Laboratory, Research Triangle Park, NC, USA, EPA/ 600/R-99/030.

Cressie NAC and Johannesson G 2008 Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series: B* **70**, 209–226.

Damon J and Guillas S 2002 The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics* **13**, 759–774.

Feister U and Balzer K 1991 Surface ozone and meteorological predictors on a subregional scale. *Atmospheric Environment* **25**, 1781–1790.

Finley AO, Banerjee S and Carlin BP 2007 `spBayes`: an `R` package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software* **12**(4), 1–24.

Fuentes M and Raftery A 2005 Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61**, 36–45.

Gelfand AE and Ghosh SK 1998 Model choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

Gneiting T, Balabdaoui F and Raftery A 2007 Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B* **69**, 243–268.

Huerta G, Sanso B and Stroud JR 2004 A spatio-temporal model for Mexico City ozone levels. *Journal of the Royal Statistical Society, Series C* **53**, 231–248.

Jun M and Stein ML 2004 Statistical comparison of observed and CMAQ modeled daily sulfate levels. *Atmospheric Environment* **38**, 4427–4436.

Kumar U and Ridder KD 2010 GARCH modelling in association with FFT-ARIMA to forecast ozone episodes. *Atmospheric Environment* **44**, 4252–4265.

Lorence AC 1986 Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society* **112**, 1177–1194.

Mardia KV, Goodall C, Redfern EJ and Alonso F 1998 The Kriged Kalman filter (with discussion). *Test* **7**, 217–252.

McMillan N, Bortnick SM, Irwin ME and Berliner M 2005 A hierarchical Bayesian model to estimate and forecast ozone through space and time. *Atmospheric Environment* **39**, 1373–1382.

Petris G, Petrone S and Patrizia C 2010 *Dynamic Linear Models with R*. Springer-Verlag, Dordrecht.

Plummer M, Best N, Cowles K and Vines K 2006 CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**(1), 7–11.

Sahu SK and Bakar KS 2012a A comparison of Bayesian models for daily ozone concentration levels. *Statistical Methodology* **9**(1), 144–157.

Sahu SK and Bakar KS 2012b Hierarchical Bayesian auto-regressive models for large space time data with applications to ozone concentration modelling. *Applied Stochastic Models in Business and Industry* **28**, 395–415.

Sahu SK and Mardia KV 2005a A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society, Series C* **54**, 223–244.

Sahu SK and Mardia KV 2005b Recent trends in modeling spatio-temporal data *Proceedings of the Special meeting on Statistics and Environment*, Università Di Messina, pp. 69–83.

Sahu SK, Gelfand AE and Holland DM 2007 High-resolution space-time ozone modeling for assessing trends. *Journal of the American Statistical Association* **102**, 1221–1234.

Sahu SK, Yip S and Holland DM 2009 Improved space-time forecasting of next day ozone concentrations in the eastern U.S. *Atmospheric Environment* **43**, 494–501.

Sahu SK, Yip S and Holland DM 2011 A fast Bayesian method for updating and forecasting hourly ozone levels. *Environmental and Ecological Statistics* **18**, 185–207.

Sousa SIV, Pires JCM, Martins F, Pereira MC and Alvim-Ferraz MCM 2009 Potentialities of quantile regression to predict ozone concentrations. *Environmetrics* **20**, 147–158.

Stein ML 2008 A modelling approach for large spatial datasets. *Journal of the Korean Statistical Society* **37**, 3–10.

Stroud JR, Muller P and Sanso B 2001 Dynamic models for spatio-temporal data. *Journal of the Royal Statistical Society, Series B* **63**, 673–689.

West M and Harrison J 1997 *Bayesian Forecasting and Dynamic Models*, 2nd ed. Springer-Verlag, New York.

Zidek JV, Le ND and Liu Z 2012 Combining data and simulated data for space-time fields: application to ozone. *Environmental and Ecological Statistics* **19**, 37–56.

14

Visualisation

John C. Gower
Department of Mathematics and Statistics, The Open University, Milton Keynes, UK

14.1 Introduction

When I was young, 'visualisation' was something that went on in the mind's eye. Sometime in the 1960s, or thereabouts, visualisation came to mean the act of depicting things either on paper or on a computer VDU screen. Indeed, the coming of the VDU screen has driven developments to such an extent that some say that the rising use of visualisation is on a par with the invention of the printing press. Be that as it may, the idea of depicting things goes back to classical times and, if we include cave-paintings, into prehistory. The current craze with visualisation has many aspects, much of it having little connection with statistics. Statistical visualisation is concerned with conveying information about data, given in numerical or categorical form. Here remarks are confined to statistical visualisation and ignore other forms of visualisation such as technical drawings and advertising.

The work of Playfair, Minard and Florence Nightingale represent well-known nineteenth century milestones in the use of diagrams to convey statistical information. Original references are hard to find but for Playfair's 1786 atlas see the reissue edited by Wainer and Spence (Playfair 2007) and for Minard see Robinson (1967). Nightingale developed a form of circular histogram (the Nightingale rose diagram, or coxcomb), and made extensive use of coxcombs to present reports on the nature and magnitude of the conditions of medical care in the Crimean War. Such reports were addressed to Members of Parliament and civil servants who would have been 'unlikely to read or understand traditional statistical reports'. The work of all these pioneers was aimed at conveying information to officials and others who may have found visual summarisations more informative than numbers. It is interesting that Playfair, Minard and Florence Nightingale all used colour in their diagrams, a facility

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

rarely available in the 20th century publication; indeed, many journals were reluctant even to publish black and white diagrams. In the twenty-first century publishers seem to have caught up with the techniques of colour reproduction that are commonplace with today's personal computers.

Throughout most of the twentieth century, statistical visualisation had not been in much favour. This may have been dictated, at least in part, by publication costs, including the cost of making professionally drawn diagrams. Yet, in scientific laboratory work the main cost was in doing the experiments; the drawing of diagrams was a relatively minor issue. In multivariate analyses, computation was a significant cost but in addition the multidimensional nature of the data was a problem that called for the development of worthwhile visualisation tools. Initially one-dimensional scales sufficed for publication and these hardly needed graphical representation. In the early part of the twentieth century, especially in factor analysis, two factors were considered, but I have been unable to find when these were first presented as two-dimensional graphical visualisations. My guess is that it had to wait until computers enabled both the computations and the finished diagrams to be produced with ease. There were certainly forerunners, such as the canonical variate analysis of Rao (see Mahalanobis et al. 1949, especially Appendix 4, pp. 248–251, and Rao 1952), but my recollection is that it was only in the early 1960s that technical advances and associated software had developed sufficiently in the United States and it was some years after that that the United Kingdom had caught up.

Another reason for the lack of diagrams in the first part of the twentieth century, especially in those statistical journals that had mathematical pretensions, was that under the influence of Bourbaki, diagrams were infra dig. As Arnol'd (1990, p. 40) put it:

> Bourbaki writes with some scorn of Barrow [Newton's teacher] that in his book in a hundred pages of text there are about 180 drawings. Concerning Bourbaki's books it can be said that in a thousand pages there is not one drawing, and it is not at all clear which is the worse.

Even in the seventeenth century, the problem of excessive mathematisation had been recognised. In c1676 Leibniz wrote (see Leibniz 1993):

> Nothing is more alien to my mind than the scrupulous attention to minor details of some authors which imply more ostentation than utility. For they consume time, so to speak, on certain ceremonies, include more trouble than ingenuity and envelop in blind night the origin of inventions which is, as it seems to me, mostly more prominent than the inventions themselves.

Finally we have R. A. Fisher's attitude as reported by Mahalanobis (1938):

> The explicit statement of a rigorous argument interested him, but only on the important condition that such explicit demonstration of rigour was needed. Mechanical drill in the technique of rigorous statement was abhorrent to him, partly for its pedantry, and partly as an inhibition to the active use of the mind. He felt it was more important to think actively, even at the expense of occasional errors from which an alert intelligence would soon recover.

Fisher's papers rarely included formal mathematics though famously, he used geometrical ideas to derive distributions (chi-squared, Student's t, Correlation etc.). Whatever

Bourbaki said, it did not strike a chord with leaders of the statistical community. One of the few things I learned as an undergraduate was that even committed Bourbakistes were not above using diagrams to back their analytical developments, though never a hint of this would appear in their formal publications.

There is a well-recognised division between Numbers/Algebra people and Geometry people, but there is no reason why they cannot coexist. A famous mathematician said that beautiful mathematical theories are constructed by first erecting an elaborate scaffolding and then knocking it down once its purpose is fulfilled. I think that it is useful to keep the scaffolding. This is especially true for research into multivariate methods such as multidimensional scaling, biplots and Procrustes analysis, where appeals to multidimensional spaces are invaluable for visualising methodology by means of diagrams involving spaces, subspaces, intersection spaces, orthogonal spaces, projections, rotations and so on. When conveying the results of one's research to colleagues, all reference to diagrams that underpin theory could be suppressed and only algebraic results presented, but although Bourbaki might not approve, it seems ridiculous to deny oneself the possibility of using visualisations when describing methodological developments whose objective is to provide visualisations. Thus, one purpose of statistical visualisation is to help develop statistical methods and to present their visualisation for the convenience of colleagues and this may call on revealing some of the scaffolding (Stone 1987 is a good example of this approach.). But the primary purpose of visualisations is to present aspects of data in a form accessible to applied scientists and to the journal public. The presentation of visualisations in the press and other media, to members of the general public is also important. We shall not pursue their special problems here, except to point out that in some ways their position is similar to that of the previous group but there is the added hazard that some sections of the media sometimes produce visualisations that are deliberately aimed to misinform (by using false origins, questionable scalings and presenting linear information in volumetric form etc.). An excellent account of the aesthetics and misuse of visualisations is given by Tufte (1983). Our concern here is not with aesthetics, nor with misinformation but with lack of information.

14.2 The problem

The coming of the VDU has revolutionised the use of statistical visualisations, but it also has brought with it some problems. I wrote a short paper (Gower 2003) about shortcomings of the visualisations given in publications. In the following, I shall refer only to the deficient visualisations to be found only too often in much of the applied literature. I was made aware of the problem when an email came to me from Indonesia asking if I could help interpret a visualisation given by a well-known statistical package. The main concepts used in statistical visualisations are distance, angle, projection, area but there was no indication which, if any, of these were appropriate, although I could make some guesses. I could only sympathise with my enquirer's predicament and suggest that he might get more information from the package's supporting manual, though I was not very confident about this. I soon became aware that interpretation of visualisations was not an isolated problem and that many published visualisations are defective. Even our statistical colleagues may sometimes have problems with interpreting visualisations (I know that I do), so how much more of a problem is it with the users of statistical software when analysing their data. Because they may not be aware of all the possibilities that are familiar to statistical methodologists, users may feel more secure. They will note various patterns in the visualisations presented to them and will be happy to report them, whether or not the patterns they see are justifiable. This is dangerous.

In most cases, users see a set of points relating to objects, possibly supplemented by some directions relating to variables and scaled orthogonal coordinate axes. If we confine our attention to visualisations of multidimensional scaling and allied multidimensional methods, the most important interpretive tools are distance (usually Euclidean), isocontours, neighbourhoods, convex hulls, inner-products, sometimes angles, the meaning to be associated with any origin and area. Being knowledgeable, our colleagues know what are the relevant tools in any particular instance, and, if our colleagues are not to be misled, it is vital that the diagrams are properly scaled. Thus, if distance is important, a circle must be exhibited as a circle – what I would term isotropic scaling. Unfortunately, some software (and sometimes an editor) attempts to fit diagrams neatly onto printed paper or computer screen. Such diagrams may be elegant but they are not interpretable. The situation is particularly dangerous when users are not aware that extraneous scaling has been introduced by the software. Often the diagrams include a scaling of both x- and y-axes. These axes, and especially their associated scales, are rarely of interest in themselves, but if it is noted that one unit in the x-direction does not equal one unit in the y-direction, the user is at least forewarned that the scaling is anisotropic. Isotropic scaling is vital for preserving distance and angle, including projection interpretations, but some flexibility is available with areas and centroids. A change of scale in one direction may be compensated for by an inverse change of scale in the other direction, without affecting area. Similarly, centroids are self-compensating although the distance between centroids is not. Note that the mediancentre of a set of points depends on minimising the sum of distances to these points and hence requires isotropic scaling. The topological properties of convexity, including convex hulls, are preserved by anisotropic scaling. An isocontour is the locus of a point P that has a constant relationship with one or more fixed points. Thus, a circle is an isocontour for all points with the same Euclidean distance from a fixed point; a square (rotated at 45° to the axes) is the corresponding locus for the L_1 norm. With area representations of asymmetry, all points P generating the same area with two fixed points is a line through P parallel to the line joining the fixed points. In three-way analyses, each choice of triadic distance generates its own set of isocontours that fall into three main classes: elliptical (including circles), hyperbolic and figures of eight; some triadic distances give isocontours in disjoint segments (De Rooij and Gower 2003). To interpret visualisations based on triadic distances, users need to be aware of the appropriate shape of isocontours. If we use inner product models, as with linear biplots and correspondence analysis of a two-way table, there is another danger. It is valid to interpret two projections onto the same biplot axis but not projections onto different biplot axes. Even if one were clever enough to be able to evaluate $ab \cos \theta$ in one's head, the comparison would still be invalid. As explained by (Gower and Hand 1995), this is because each biplot axis has its own scale that, with appropriate calibration, allows the inner product to be read directly. If the scales are shown on the biplot axis, valid comparisons between projections onto different axes may be made without having to evaluate $ab \cos \theta$. However, biplots may be used for two different purposes – interpolation and prediction – each of which needs its own scale. Furthermore, prediction uses projection, while interpolation uses the process of vector summation. The interpreter of any biplot needs to know these things and know which form is being presented in any biplot visualisation.

Orthogonal projection of a point P is given by its nearest point in a subspace, often a coordinate axis. For a categorical variable, rather than a subspace, we have a set of category-level points, CLPs (Gower and Hand 1995) and the nearest CLP to P is required. The CLPs define a set of convex neighbour regions, which may be shown on visualisations and used to predict category levels in the same way that projection predicts quantitative variables. Ordered categorical variables define ordered neighbour regions. See Gower (2002) for further information. Other problems arise with approximations based on the singular value decomposition where singular value contributions may be assigned to rows and columns in a variety of ways. The relevant variant needs to be recognisable in any visualisation.

The considerations that underlay the interpretation of multidimensional scaling type diagrams differ markedly from those appropriate for plotting functional relationships in two dimensions. There, the scaling of the two variables may be chosen to optimise concepts such as the aspect ratio (Cleveland 1995). In multidimensional scaling, differences between measurement scales are accommodated by normalisation or a choice of distance function such as one of the many dissimilarity coefficients or chi-squared distance; thereafter Euclidean visualisations are isotropic and are usually rotationally invariant. Note, however, that the group-average configuration of INDSCAL is isotropic but not rotationally invariant. The root of this difference between visualisations of functions and multidimensional visualisations is the identification, or not, of a response variable.

14.3 A possible solution: self-explanatory visualisations

The previous section has identified that multidimensional visualisations are often incompletely described. It is true that a careful reading of manuals accompanying software may resolve some problems, but you have to know that there may be a problem in the first place. Also, in principle, all the information could be given in the legend describing a visualisation, but this would be incredibly tedious and repetitive. What is needed is some coded form of the information which would tell the initiated all that was required and would alert the uninitiated to the possibility of difficulties. A readily available document would give a detailed description of the code. The objective is to give a compact form of all the information needed to enable a correct interpretation of the visualisation. I called this a Self-Explanatory Visualisation and the compact form of the code a ‘cartouche’ although more recently we use the term ‘icon’.

Gower (2003) gave some preliminary suggestions of what information may be presented in the cartouche/icon form. Information would be needed on: Origin, Isocontours, Distances, Scales, Projection and inner product, Eigenvalue scaling, Neighbourhoods but only when relevant. As mentioned earlier the discussion was confined to applied multivariate analysis but it was recognised that if the idea were to be successful it should be extended to include the whole of statistics and even beyond! Of course, nothing happened but recently a group of colleagues tried again to raise interest in the problem. We were stimulated by the very poor examples of visualisations to be found in much of the applied literature and wrote a short account of our findings. To limit the extent of our investigations, we focused on the marketing literature. We had trouble in getting it published in any of the marketing journals, although most editors said that they had found the paper interesting, it was just that it was not the sort of paper that they published, or they did not think their readers would be interested or it was not methodological research and so on. Eventually the Journal of Food Quality and Preference (Gower et al. 2014) accepted it and it has just been published. We await responses. However, I am convinced that there is an in urgent need of some generally agreed method for making visualisations self-explanatory and I hope that others will be sufficiently interested to take the essential steps.

I am not sure what all this has to do with Kanti except in so far as good visualisations are in the interest of all statisticians. I am sure all Kanti's visualisations are good and I do not think he has Bourbaki tendencies.

References

Arnol'd VI 1990 *Huygens, Barrow, Newton and Hooke: Pioneers in Mathematical Analysis and Catastrophe theory from Evolvents to Quasi Crystals*. Birkhauser, Basel. translated from the Russian by E. J. F. Primrose.

Cleveland WS 1995 *The Elements of Graphing Data*. Wadsworth Publishing Company, Monterey.

De Rooij M and Gower JC 2003 The geometry of triadic distances. *Journal of Classification* **20**, 181–220.

Gower JC 2002 Categories and quantities In *Measurement and Multivariate Analysis* Nishisato S, Baba Y, Bozdogan H and Kanefuji K (eds), Springer-Verlag, Tokyo, pp. 1–12.

Gower JC 2003 Visualisation in multivariate and multidimensional data analysis. *Bulletin of the International Statistical Institute* **54**, 101–104.

Gower JC and Hand DJ 1995 *Biplots*. Chapman and Hall, London.

Gower JC, Groenen PJF, Van de Velden M and Vines K 2014 Better perceptual maps: introducing explanatory icons to facilitate interpretation. *Food Quality and Preference*, **36**, 61–69.

Leibniz GW 1993 *De quadratura arithmetica circuli ellipseos et hyperbolae cujus corollarium est trigonometria sine tabulis*. Vandenhoeck & Ruprecht, Göttingen. Edited and commented by Eberhard Knobloch.

Mahalanobis PC 1938 Professor Ronald Aylmer Fisher. *Sankhya* **4**, 265–272.

Mahalanobis PC, Majumdar DN, Yeatts MWM and Rao CR 1949 Anthropometric survey of the United Provinces, 1941: a statistical study. *Sankhyā* **9**, 89–324.

Playfair W 2007 *The Commercial and Political Atlas, Representing, by means of Stained Copper-Plate Charts, The Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the whole of the Eighteenth Century, and The Statistical Breviary; Shewing on a Principle entirely new, the resources of every state and kingdom in Europe; illustrated with Stained Copper-Plate Charts, representing the physical powers of each distinct nation with ease and perspicuity both*. Cambridge University Press, New York. Edited by Howard Wainer and Ian Spence.

Rao CR 1952 *Advanced Statistical Methods in Biometric Research*. John Wiley & Sons, Inc., New York.

Robinson AH 1967 The thematic maps of Charles Joseph Minard. *Imago Mundi* **21**, 95–108.

Stone M 1987 *Coordinate Free Multivariate Statistics*. Clarendon Press, Oxford.

Tufte ER 1983 *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.

15

Fingerprint image analysis: role of orientation patch and ridge structure dictionaries

Anil K. Jain and Kai Cao
Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

15.1 Introduction

Biometric traits, such as palmprints (Duta et al. 2002) and fingerprints (Jain et al. 1997), refer to distinctive anatomical and behavioral characteristics for automatic human identification. Fingerprints, which are ridge and valley patterns on the tip of a human finger, are one of the most important biometric traits due to their known uniqueness and persistence properties (Maltoni et al. 2009). Since the advent of fingerprints for identifying and tracing criminals in Argentina in 1893 (Hawthorne 2008), fingerprints have been primarily used as evidence in law enforcement and forensics. After the first paper on automated fingerprint matching was published by Mitchell Trauring (1963) in Nature in 1963, the Federal Bureau of Investigation (FBI) installed the first Automated Fingerprint Identification System (AFIS) in 1980. Now large-scale fingerprint recognition systems are not only used worldwide by law enforcement and forensic agencies, they are also beginning to be deployed in civilian applications, such as (i) the OBIM (formerly the US-VISIT) program by the Department of Homeland Security (Department of Homeland Security 2014) and (ii) India's Aadhar project (Planning Commission, Goverment of India 2014). In 2013, the TouchID system (Apple, Inc. 2014) in the Apple iPhone 5s for authenticating mobile phone

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

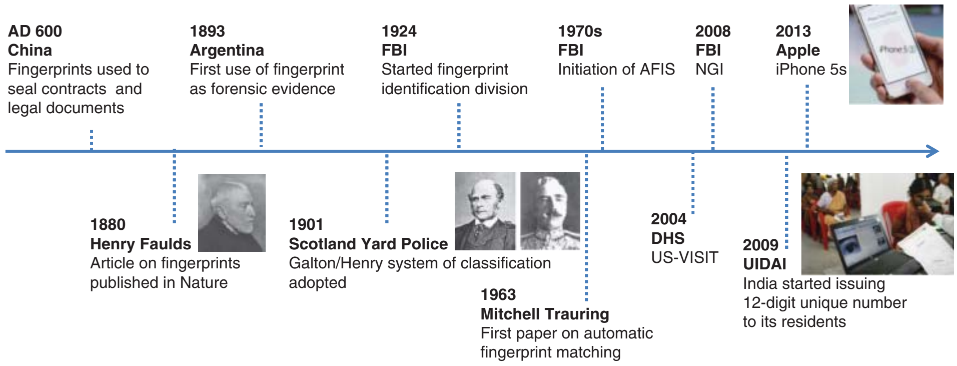

AD 600
China
Fingerprints used to seal contracts and legal documents
1893
Argentina
First use of fingerprint as forensic evidence
1924
FBI
Started fingerprint identification division
1970s
FBI
Initiation of AFIS
2008
FBI
NGI
2013
Apple
iPhone 5s
1880
Henry Faulds
Article on fingerprints published in Nature
1901
Scotland Yard Police
Galton/Henry system of classification adopted
1963
Mitchell Trauring
First paper on automatic fingerprint matching
2004
DHS
US-VISIT
2009
UIDAI
India started issuing 12-digit unique number to its residents

Figure 15.1 Some major milestones in fingerprint recognition.

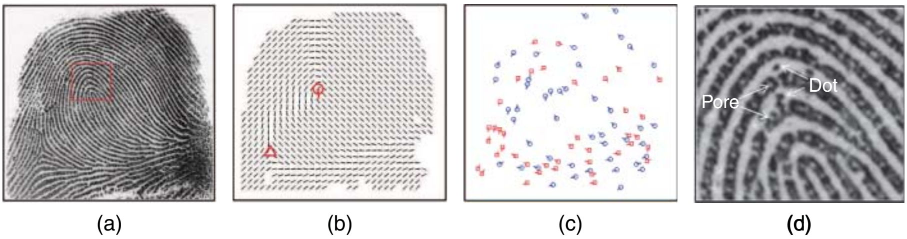


Figure 15.2 Illustration of fingerprint features at three different levels. (a) A gray-scale fingerprint image (NIST SD30, A002_01), (b) level 1 features: orientation field and singular points (core point shown as a circle and delta point shown as a triangle), (c) level 2 features: ridge ending minutiae (squares) and ridge bifurcation minutiae (circles), and (d) level 3 features: pores and dots.

users launched the application of fingerprint in mobile devices. Some major milestones in the history of fingerprint recognition are illustrated in Figure 15.1.

The purported uniqueness of fingerprints is characterized in terms of three levels of features (Maltoni et al. 2009) (see Figure 15.2). Level 1 features include the general ridge flow and pattern configurations such as pattern type, ridge orientation, and frequency fields, and singular points (core and delta points). While level 1 features are not sufficient for individualization, they can be used for exclusion (the outcomes of comparing a fingerprint pair are one of three possibilities: match, inconclusive, and exclusion). Level 2 features mainly refer to local ridge discontinuities, called minutia points; ridge endings and ridge bifurcations are the two most prominent types of minutia points. Level 3 features cover all other attributes at a fine level, such as width, shape, curvature, and edge contours of ridges, pores, and incipient ridges. Level 3 feature extraction requires that the fingerprint images be acquired at a 1000 ppi resolution. Among all these fingerprint features, the set of minutia points (called minutiae) is regarded as the most distinctive and, hence, is the most commonly used feature in fingerprint identification, both by human experts and AFIS.

Based on the fingerprint image acquisition method and their source, fingerprints can be classified into three types, namely rolled, plain, and latent (see Figure 15.3). Rolled fingerprints are obtained by taking the impression from “nail to nail” in order to capture the

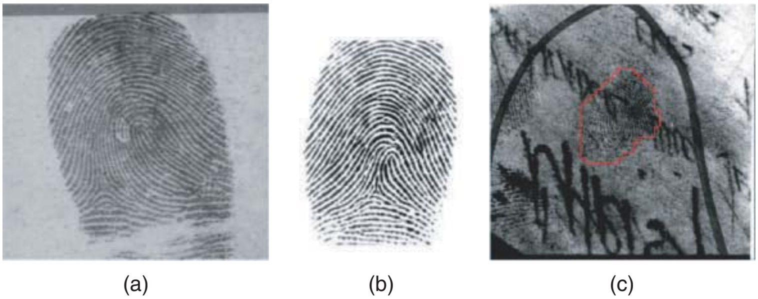


Figure 15.3 Three types of fingerprint images. (a) Rolled fingerprint (from NIST Special Database 4 2014), (b) plain fingerprint from (FVC2002 2002), and (c) latent fingerprint (*Source:* adapted from NIST Special Database 27 2014).

complete ridge details of a finger. Plain fingerprints are acquired by pressing a fingertip onto a flat surface of either a paper for the inking method or a flatbed of a live-scan device (Maltoni et al. 2009). Latent fingerprints (or simply latents) refer to fingerprints lifted from the surfaces of objects, which are inadvertently touched or handled by a person. Compared to rolled and plain fingerprints (or collectively called exemplar fingerprints), which are typically acquired in the presence of an officer or trained personnel, latents are generally of poor quality with incomplete ridge structure, background noise, and nonlinear distortion. Consequently, the accuracy of latent-to-exemplar matching is significantly lower than that of exemplar-to-exemplar matching. In NIST evaluations, the best performing AFIS achieved a true acceptance rate of 99.4% at a false acceptance rate of 0.01% for exemplar-to-exemplar fingerprint matching (Wilson et al. 2004). The best performing commercial latent matcher could only achieve a rank-1 identification rate of 63.4% in searching 1,114 latents against a background database containing 100,000 exemplar prints (Indovina et al. 2012). The search for the source of a latent is a challenging problem in terms of both the algorithmic efficiency and identification accuracy, especially when the reference or exemplar database (rolled or plain fingerprints) is extremely large. Figure 15.4 shows examples of rolled-to-rolled fingerprint matching and latent-to-rolled fingerprint matching.

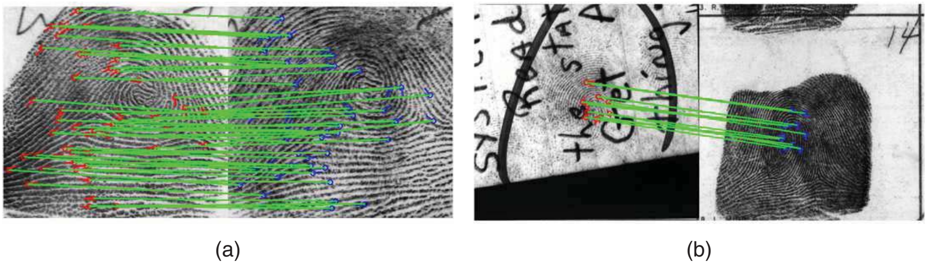


Figure 15.4 Examples of (a) rolled-to-rolled fingerprint matching and (b) latent-to-rolled fingerprint matching. Features in the rolled fingerprints shown here are extracted automatically by an AFIS, but features (minutiae, region of interest, and singular points) in the latent were manually marked.

Given the difficulty of automatic latent matching, human intervention is unavoidable in order to assess the value of latents as forensic evidence, mark features such as region of interest (ROI) and minutiae, and make a decision whether the latent has a match in the reference database given the candidate list (typically top 50 matches) generated by AFIS. Hence, latent examiners and AFIS work collaboratively in a framework called Analysis, Comparison, Evaluation, and Verification (ACE-V) (Ashbaugh 1999). However, human involvement in latent examination has raised some concerns related to repeatability and reliability (Ulery et al. 2011; 2012). Furthermore, when the comparison time (between latent and exemplar prints) is limited, latent examiners are more likely to make an inconclusive matching decision (Dror et al. 2011). One of the priorities in the FBI's Next Generation Identification (NGI) program is to support the development of a lights-out[1] capability for latent identification (FBI- NGI 2014). An essential component necessary for achieving the lights-out capability is automatic feature extraction from latent fingerprints, this is necessary to (i) increase the throughput of latent matching systems, (ii) achieve repeatability of latent feature extraction, and (iii) improve the compatibility between features extracted in the latents and in the exemplar prints (Feng et al. 2013).

In order to achieve reliable feature extraction from latents, the latent images need to go through two main preprocessing steps: (i) segmentation to separate friction ridges from noisy background and (ii) fingerprint enhancement to enhance ridge and valley structures. Directional filtering, such as Gabor filtering (Hong et al. 1998), can adaptively improve the clarity of ridge and valley structures; the filters are tuned based on the local ridge orientation and frequency. Therefore, for latent enhancement, it is essential to first obtain a good estimate of ridge orientation and frequency fields.

There is a rich body of literature on fingerprint segmentation (Chikkerur et al. 2007; Hong et al. 1998), orientation field estimation (Chikkerur et al. 2007; Hong et al. 1998; Mardia et al. 1997; Wang et al. 2007), and frequency field estimation (Chikkerur et al. 2007; Jiang 2000) for exemplar fingerprints. But these approaches do not work well on latent fingerprints since they did not consider (i) the presence of structured noise, such as lines, markings, characters, and speckles (see Figure 15.3(c)), which break the friction ridge pattern and hinder reliable feature extraction; and (ii) unclear fingerprint ridges in the foreground area. Some approaches have been proposed to specifically address the problem of latent fingerprint segmentation (Karimi-Ashtiani and Kuo 2008; Short et al. 2011; Zhang et al. 2013) and enhancement (Yoon et al. 2011). However, none of these approaches use the prior knowledge of ridge structure in fingerprints, resulting in only a marginal improvement in latent matching.

A dictionary, which is a set of words (or vectors) used to sparsely and linearly represent signals of the same dimension (called sparse coding), has been successfully applied to a number of signal processing problems, such as image denoising (Elad and Aharon 2006; Mairal et al. 2008b), classification (Lian et al. 2010; Mairal et al. 2008a), and face recognition (Liao et al. 2013; Wright et al. 2009). The dictionary learned from a set of training data is a collection of representative vectors of the training data. In this chapter, we investigate the use of dictionaries for the challenging problems in latent fingerprint image analysis, namely, latent fingerprint segmentation and enhancement. Given that fingerprint patterns can be represented at two different levels (i.e., coarse representation for fingerprint ridge flow or orientation field and fine representation for ridges and valleys), two dictionaries are

[1] Lights-out identification refers to an AFIS requiring minimal or no human assistance in which a query fingerprint image is presented as input, and the output consists of a short candidate list (Indovina et al. 2009).

developed: an orientation patch[2] dictionary (Feng et al. 2013) and a ridge structure dictionary (Cao et al. 2014). An orientation patch dictionary, which contains only the orientation information in patches, is proposed to estimate the orientation field for latent fingerprint enhancement. A ridge structure dictionary, which contains ridge and valley patterns, is proposed for latent fingerprint segmentation (locating friction ridge pattern) and enhancement by estimating orientation and frequency fields. Experimental results on public domain latent fingerprint databases show that the dictionaries, learned from a large number of fingerprints, capture a domain-specific knowledge, which is effective in improving the accuracy of latent fingerprint matching.

The rest of the chapter is organized as follows. In Section 15.2, the method of dictionary construction for fingerprint patterns at orientation level and ridge level is described. The orientation patch dictionary for orientation field estimation in latent fingerprints is presented in Section 15.3. The ridge structure dictionary for latent segmentation and enhancement is introduced in Section 15.4. Conclusions and future research directions are described in Section 15.5.

15.2 Dictionary construction

Orientation patch and ridge structure dictionaries are both constructed off-line from high-quality fingerprints to capture prior knowledge about fingerprint patterns.

15.2.1 Orientation patch dictionary construction

To construct a dictionary of reference orientation patches, we used a set of 50 high-quality fingerprints (referred to as reference fingerprints) in the NIST SD4 database (NIST Special Database 4 2014). All five major pattern types (plain arch, tented arch, left loop, right loop, and whorl) are covered by the reference fingerprints. The high-quality fingerprints are manually selected, and their orientation fields (with block size 16×16 pixels) are estimated using VeriFinger 6.2 SDK (Neurotechnology Inc. 2012). A number of training orientation patches are obtained by sliding a window (of size $b \times b$ blocks) across the orientation field and its mirrored version for each reference fingerprint, where an orientation patch consists of $b \times b$ orientation elements and an orientation element refers to the dominant orientation in a block of size 16×16 pixels. Each orientation patch is rotated by 21 different angles $\{i \cdot 5^{\circ}, -10 \leq i \leq 10\}$ to generate additional training orientation patches to cover all possible directions in the latent fingerprints.

Given the training orientation patches, the orientation patch dictionary (shown in Figure 15.5) is constructed by a *greedy* Algorithm, which is described as follows:

1. The first orientation patch in the training set is added to the dictionary, which is initially empty.
2. The next orientation patch in the training set, which is sufficiently different from all orientation patches in the dictionary, is added to the dictionary. The similarity measure between two orientation patches of size $b \times b$ is computed as n_s/b^2, where n_s denotes the number of orientation elements whose difference is less than 10°.
3. Repeat step 2 until all orientation patches have been considered.

[2] An orientation patch refers to the block-wise orientation field of a fingerprint patch.

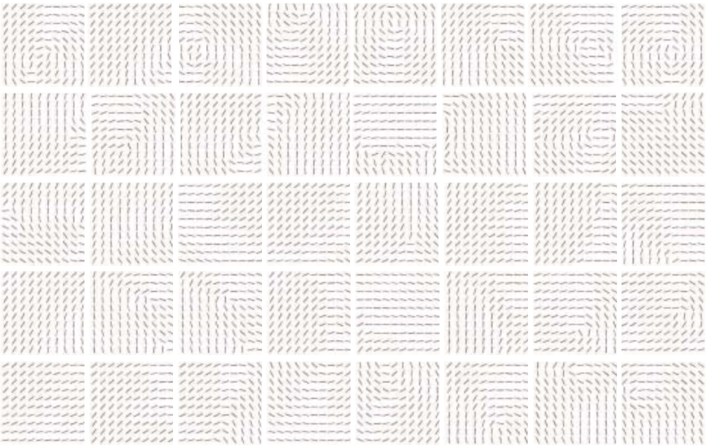
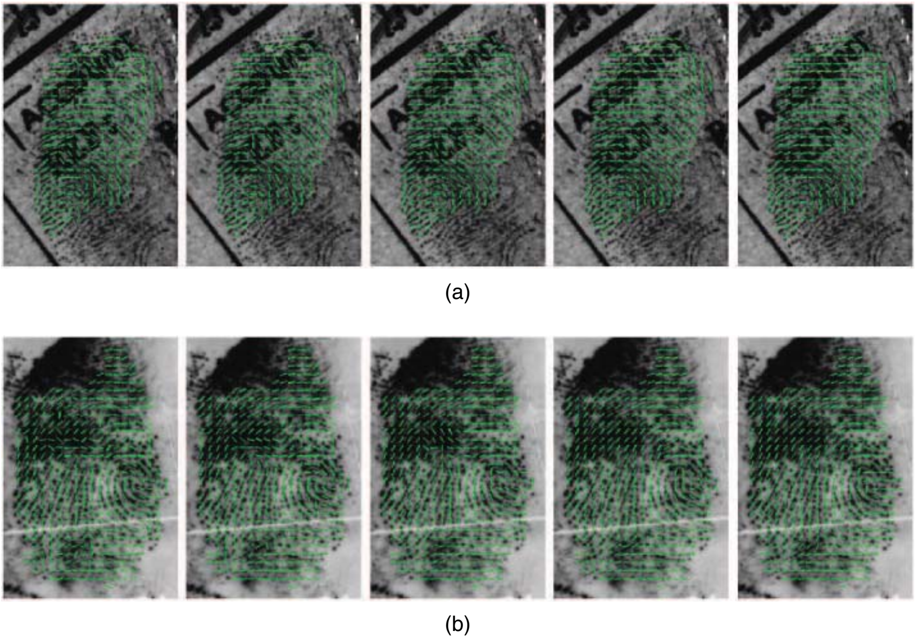


Figure 15.5 Examples of orientation patches in the dictionary; an orientation patch contains 10×10 orientation elements, and each orientation element corresponds to a block of 16×16 pixels.

The number of reference orientation patches in the dictionary depends on the number of reference fingerprints and the patch size. When the patch size is 10×10 blocks and 50 reference fingerprints are used, the number of reference orientation patches in the dictionary is around 23,000. While a larger size of the orientation patch is better for correcting errors in the initial orientation field, it would require a larger dictionary, which takes more time to search. To further demonstrate the impact of patch size, orientation fields corrected using different patch sizes are compared in Figure 15.6, where an initial orientation patch is directly replaced by its closest orientation patch in the dictionary without considering compatibility between neighboring patches. The performance of the dictionary-based approach improves with an increase in patch size. The estimation errors close to the fingerprint boundary are due to border effect (these patches contain very few foreground blocks with a friction ridge pattern).

15.2.2 Ridge structure dictionary construction

The orientation patch dictionary characterizes ridge flow patterns of fingerprints, which can be used for correcting an initial orientation field. However, the structure of ridges and valleys is ignored in the orientation patch dictionary. To remedy this, a ridge structure dictionary, which is learned directly from the fingerprint image patches, is introduced in this section. A large size of image patches will result in high dimensionality (a 4096-dimensional vector for a 64×64 image patch, compared to 100-dimensional vector for a 10×10 orientation patch which covers a 160×160 image patch) and hence, a large dictionary. On the other hand, image patches of small size are not robust to structured background noise in the latent images. Hence, two levels of ridge structure dictionaries are constructed: (i) a coarse-level dictionary with large patch size of 64×64 pixels, and (ii) 16 fine-level (orientation-specific) dictionaries with small patch size of 32×32 pixels. The patch size for fine-level dictionaries is 32×32 pixels, which covers about two ridges and valleys for 500 ppi fingerprints, and



(a)

(b)

Figure 15.6 Orientation fields extracted from two latent fingerprint impressions ((a) and (b)) estimated using different patch sizes (increasing from left to right: 3×3, 5×5, 7×7, 9×9 and 11×11). Only the nearest dictionary element of each initial orientation patch is considered here (*Source:* Feng et al. (2013), Figure 8, p. 931. Reproduced by permission of IEEE).

is robust to structured noise. The patch size for the coarse-level dictionary is twice the size of the fine-level dictionary to cover additional ridge structures.

15.2.2.1 Training set selection

A large number of high-quality fingerprint patches from rolled fingerprints in NIST SD4 (NIST Special Database 4 2014) are selected for the dictionary construction as follows:

1. High-quality fingerprint selection: NIST Fingerprint Image Quality (NFIQ) (Tabassi et al. 2004) is used to select 500 fingerprints of high quality[3] (i.e., NFIQ < 3) in NIST SD4.
2. High-quality patch selection: The block-wise orientation field and ridge quality map of the selected fingerprints are computed by MINDTCT (Garris et al. 2004). For the coarse-level dictionary, an image patch of size 64×64 pixels is included into the training set P^c if the average quality value of the image patch is larger than a predefined threshold T (T is set to 3.75, where the block ridge quality in MINDTCT ranges from 0 (the lowest quality) to 4 (the highest quality)). For the fine-level dictionary with the ith orientation ($i = 1, \ldots, 16$) orientation-specific dictionary, an image

[3] NFIQ ranges from 1 to 5, with 1 indicating the highest quality and 5 indicating the lowest quality fingerprint.

patch is included in the training set P_i^f if it satisfies the following two conditions: (i) average quality value of the patch is larger than T and (ii) average ridge orientation of the patch is within the range $\left[(i-1) \times \frac{\pi}{16}, i \times \frac{\pi}{16}\right)$.

3. *Vector normalization:* Each patch p in the training sets is converted into a vector by concatenating the rows and normalized with a mean of zero and standard deviation of one.

Let $P^c = \{p_j^c\}_{j=1}^{N^c}$ be the training set with N^c training patches for the coarse-level dictionary, and $P_i^f = \{p_{i,j}^f\}_{j=1}^{N_i^f}$, $i = 1, \ldots, 16$, be the training set for the ith fine-level dictionary, where N_i^f denotes the number of training patches for the ith fine-level dictionary specified by ridge orientation. We then randomly select 80,000 image patches from P^c and 10,000 image patches from each P_i^f for dictionary learning.

15.2.2.2 Dictionary learning

The goal of dictionary learning is to construct a dictionary D of size $N_P \times N_D$ that provides the best sparse representation for each patch in $P = \{p_j\}_{j=1}^{N}$, where N_P is the dimensionality of the patches in P and N_D is the number of elements in the dictionary D. After the ridge dictionaries are constructed by K-SVD (Aharon et al. 2006), each dictionary element is normalized with a mean of zero and standard deviation of one.

A total of 17 different dictionaries are constructed by taking subsets selected from P^c and P_i^f, $i = 1, \ldots, 16$ as the training sets. The number of elements N_D^c in the coarse-level dictionary is set to 1,024, and the total number of elements N_D^f in each fine-level dictionary is set to 64. Figure 15.7(a) shows a subset of dictionary elements in the coarse-level dictionary D^c and Figure 15.7(b) shows a subset of dictionary elements in the 16 fine-level dictionaries D_i^f.

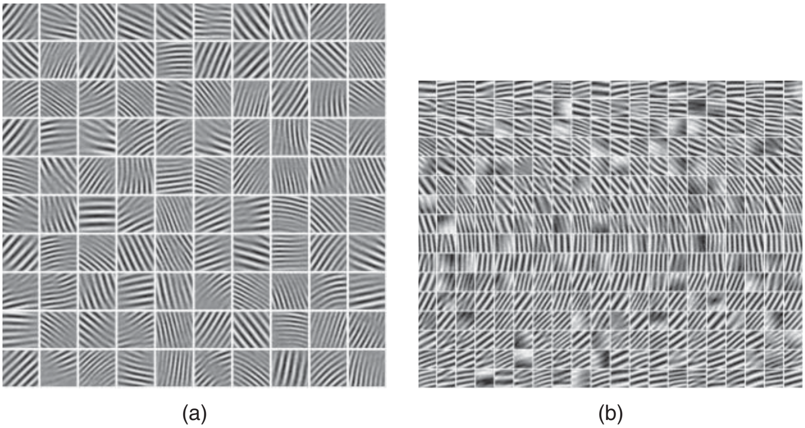


Figure 15.7 Examples of coarse and fine-level dictionaries. (a) A subset of elements in the coarse-level dictionary, and (b) a subset of elements in the 16 orientation-specific dictionaries. The ith row in (b) corresponds to the orientation range $\left[(i-1) \times \frac{\pi}{16}, i \times \frac{\pi}{16}\right)$, $i = 1, \ldots, 16$.

15.3 Orientation field estimation using orientation patch dictionary

Given the orientation patch dictionary learned in Section 15.2.1, the orientation field in the latent foreground is estimated using the following three steps (see Figure 15.8):

1. *Initial estimation:* The initial orientation field is obtained using a local orientation estimation method, such as local Fourier analysis (Jain and Feng 2009).
2. *Dictionary lookup*: The initial orientation field is divided into overlapping patches. For each initial orientation patch, its six nearest orientation patches in the dictionary are selected as candidates for replacing the noisy initial orientation patch.
3. *Context-based correction*: The optimal combination of candidate orientation patches is found by minimizing the energy function, which includes the similarity between an initial orientation patch and its selected reference orientation patch in the dictionary and the compatibility between neighboring reference orientation patches.

Details of the orientation field estimation algorithm are presented in the following sections.

15.3.1 Initial orientation field estimation

The initial orientation field (16 × 16 pixel block size) is obtained by detecting a peak in the magnitude spectrum of the local image (Jain and Feng 2009). Other local estimation approaches should also suffice for this initial estimation. Although the initial orientation field is typically very noisy due to the poor quality of latents, it should not be smoothed at this stage since the correct orientation elements may be degraded by noise in the neighboring

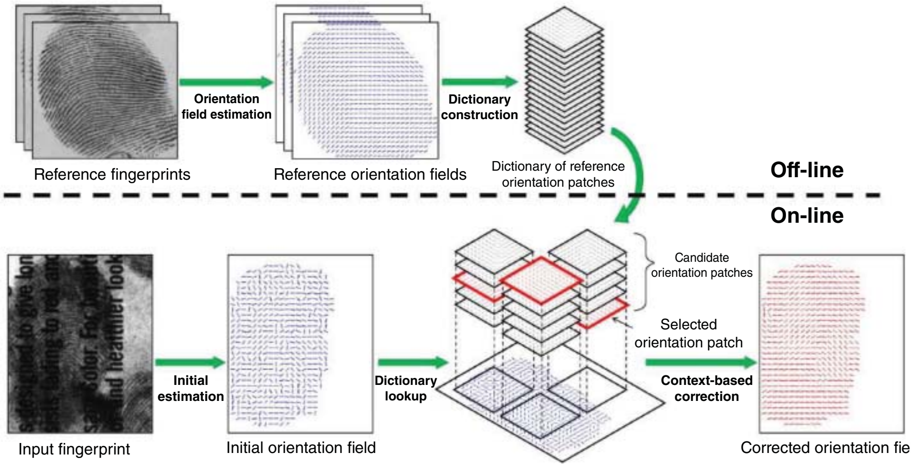

Orientation field estimation
Dictionary construction
Reference fingerprints
Reference orientation fields
Dictionary of reference orientation patches
Off-line
On-line
Candidate orientation patches
Selected orientation patch
Initial estimation
Dictionary lookup
Context-based correction
Input fingerprint
Initial orientation field
Corrected orientation field

Figure 15.8 Flowchart of the orientation field estimation algorithm, which consists of an off-line dictionary construction stage and an online orientation field estimation stage (*Source:* Feng et al. (2013), Figure 5, p. 929. Reproduced by permission of IEEE).

regions. The initial orientation field is updated in the later stages by utilizing prior knowledge of fingerprints contained in the orientation patch dictionary.

15.3.2 Dictionary lookup

Given an initial orientation patch that contains at least one foreground block, a number of candidate reference orientation patches from the dictionary are retrieved based on their similarity with the initial orientation patch. The similarity $S(\Theta, \Phi)$ between an initial orientation patch Θ and a reference orientation patch Φ is defined as

$$S(\Theta, \Phi) = n_s/n_f, \tag{15.1}$$

where n_f is the number of orientation elements in the initial orientation patch and n_s is the number of orientation elements whose differences are less than a predefined threshold (empirically set as $\pi/12$). However, for many initial orientation patches, the top candidate orientation patches of an initial orientation patch are quite similar to each other. In order to increase the probability of including the correct reference orientation patches in a short list, it is better to select a set of diverse candidates. A diverse set of n_c (empirically set as 6) candidates is selected from the top $10n_c$ initial candidates using the following greedy strategy:

1. Choose the first initial candidate orientation patch.
2. The next candidate patch is compared to each of the chosen candidates. If its similarity to all the chosen candidates is less than a predefined threshold (empirically set as 0.8), it is included in the list. Note that similarity is computed using only the foreground blocks in the initial orientation patch.
3. Repeat step 2 for all the initial candidates until n_c candidates have been chosen or all initial candidates have been checked.

As a result of the diversity heuristic, the correct orientation patch is more likely to appear in the candidate list, even if the initial orientation patch is very noisy or incomplete.

15.3.3 Context-based orientation field correction

After dictionary lookup, we obtain a list of c_i ($1 \leq c_i \leq n_c$) candidate orientation patches, $\Phi_i = \{\Phi_{i,1}, \Phi_{i,2}, \ldots, \Phi_{i,c_i}\}$, for an initial orientation patch Θ_i. Let r_i denote the index of the selected candidate for the patch i, and $\mathbf{r} = \{r_1, r_2, \ldots, r_{n_p}\}$ be the vector of the indices of the selected candidates for all n_p foreground patches. Any combination of candidate indices could be a solution for the orientation field estimation. But this leads to a large solution space, so we utilize contextual information to reduce the search space.

We address this problem by searching for optimal indices vector, $\mathbf{r}^*$, which minimizes an energy function $E(\mathbf{r})$. Choice of a proper energy function is crucial for the success of this method. We consider two factors in designing the energy function: (i) the similarity between the reference orientation patches and the corresponding initial orientation patches and (ii) the compatibility between neighboring reference orientation patches. The proposed energy function $E(\mathbf{r})$ is defined as

$$E(\mathbf{r}) = E_s(\mathbf{r}) + w_c E_c(\mathbf{r}), \tag{15.2}$$

where $E_s(\mathbf{r})$ and $E_c(\mathbf{r})$ denote the similarity term and compatibility term, respectively, and w_c (empirically set to 1 by the authors in Feng et al. 2013) is the weight of compatibility term. The similarity term is defined as

$$E_s(\mathbf{r}) = \sum_{i \in \mathcal{V}} (1 - S(\Theta_i, \Phi_{i,r_i})), \tag{15.3}$$

where $\mathcal{V}$ denotes the set of foreground patches and $S(\cdot)$ is defined in Equation (15.1). The compatibility term is defined as

$$E_c(\mathbf{r}) = \sum_{(i,j) \in \mathcal{N}} (1 - C(\Phi_{i,r_i}, \Phi_{j,r_j})), \tag{15.4}$$

where $\mathcal{N}$ denotes the set of adjacent foreground patches, which are four-connected neighbors.

The compatibility between two neighboring orientation patches Φ_{i,r_i} and Φ_{j,r_j} is measured by the similarity of orientations in the overlapping blocks. Let $\{\alpha_n\}_{n=1}^{N_o}$ and $\{\beta_n\}_{n=1}^{N_o}$ be the set of orientations in the N_o overlapping blocks of two orientation patches. The compatibility is computed as

$$C(\Phi_{i,r_i}, \Phi_{j,r_j}) = \frac{1}{N_o} \sum_{n=1}^{N_o} |\cos(\alpha_n - \beta_n)|. \tag{15.5}$$

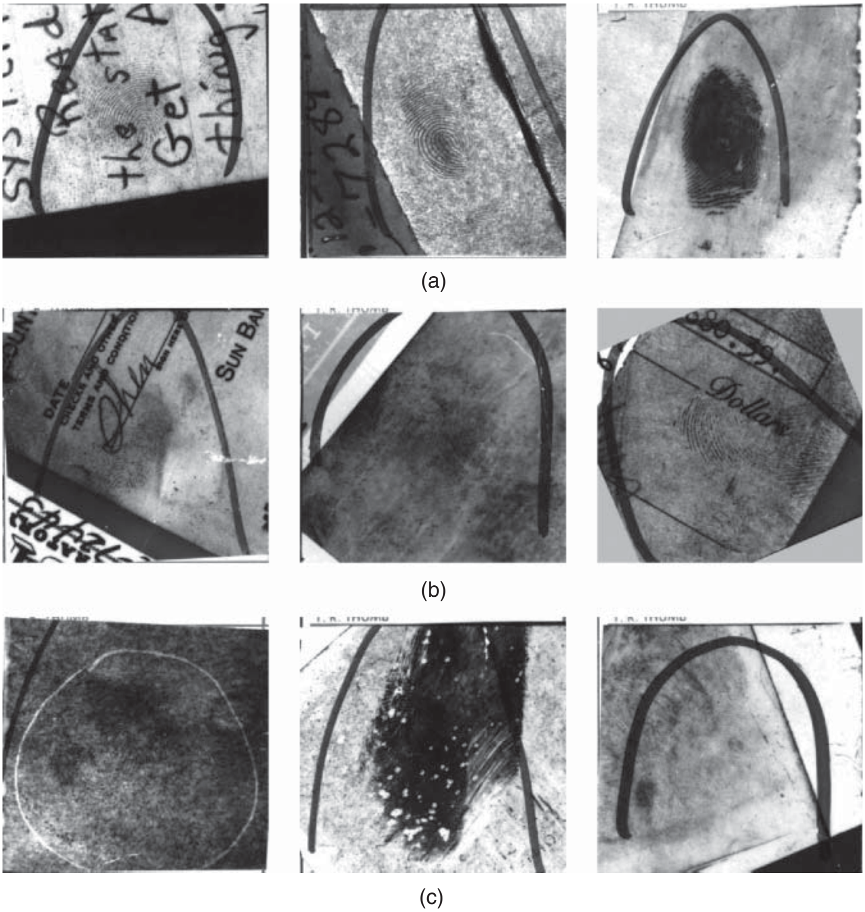
To minimize the energy function in Equation (15.2), the well-known loopy belief propagation algorithm (Blake et al. 2011) is used for optimization.

15.3.4 Experiments

The goal of an orientation field estimation algorithm is to obtain an accurate estimation of fingerprint orientation field for fingerprint enhancement and feature extraction and then to improve the fingerprint matching accuracy. The dictionary-based algorithm is, therefore, evaluated in terms of the accuracy of orientation field estimation and the accuracy of fingerprint matching, respectively. The latent orientation field estimation and subsequent matching experiments are conducted on NIST SD27 (NIST Special Database 27 2014), which contains 258 latent fingerprint images (500 ppi). These latents have been classified into three different qualities, namely, “Good,” “Bad,” and “Ugly” (very bad), and their corresponding mated rolled fingerprints. The numbers of “Good,” “Bad,” and “Ugly” latents are 88, 85, and 85, respectively. Figure 15.9 displays examples of latent with these three qualities. To make the latent matching problem more realistic and challenging, 27,000 rolled fingerprints (file fingerprints) in the NIST SD14 database were also included in the background database.

In addition to the proposed orientation field estimation algorithm, two other approaches were included for comparison:

1. *FOMFE:* Combination of gradient-based local estimation and FOMFE-based global model (Wang et al. 2007).
2. *STFT:* Combination of STFT-based local estimation and low-pass filtering (Chikkerur et al. 2007).



(a)

(b)

(c)

Figure 15.9 Examples of latents of different qualities. (a) Good, (b) bad, and (c) ugly.

The accuracy of orientation field estimation algorithms is measured in terms of the average Root Mean Square Deviation (RMSD) (Turroni et al. 2011). The ground truth orientation fields were manually marked by one of the authors in Feng et al. (2013). Average RMSD of the dictionary-based algorithm, FOMFE, and STFT are computed on all the 258 latents in the NIST SD27 database and also on the three quality level subsets (Good, Bad, and Ugly). Table 15.1 shows that the dictionary-based algorithm outperforms FOMFE and STFT for all three subsets of latent fingerprints in NIST SD27.

In order to evaluate the matching performance, latent fingerprints are enhanced using a Gabor filter (Hong et al. 1998) whose frequency parameter is fixed at $1/9$ cycles per pixel, orientation parameter is tuned to the estimated orientation field, and standard deviations of the Gaussian envelope in x and y directions are fixed at 4. VeriFinger SDK 6.2 (Neurotechnology Inc. 2012) is used for feature extraction and matching.

Figure 15.10 shows the Cumulative Match Characteristic (CMC) curves obtained from the three orientation field estimation algorithms and the manual markup (ground truth). The dictionary-based algorithm consistently outperforms the other two algorithms (FOMFE and

Table 15.1 Average estimation error (in degrees) of the orientation field estimation algorithm based on orientation patch dictionary and two competing algorithms on the latent fingerprints in the NIST SD27 Database.

Algorithm	All	Good	Bad	Ugly
Orientation patch dictionary (Feng et al. 2013)	18.44	14.40	19.18	21.88
FOMFE (Wang et al. 2007)	28.12	22.83	29.09	32.63
STFT (Chikkerur et al. 2007)	32.51	27.27	34.10	36.36

Source: Feng et al. (2013), Table 2, p. 932. Reproduced by permission of IEEE.

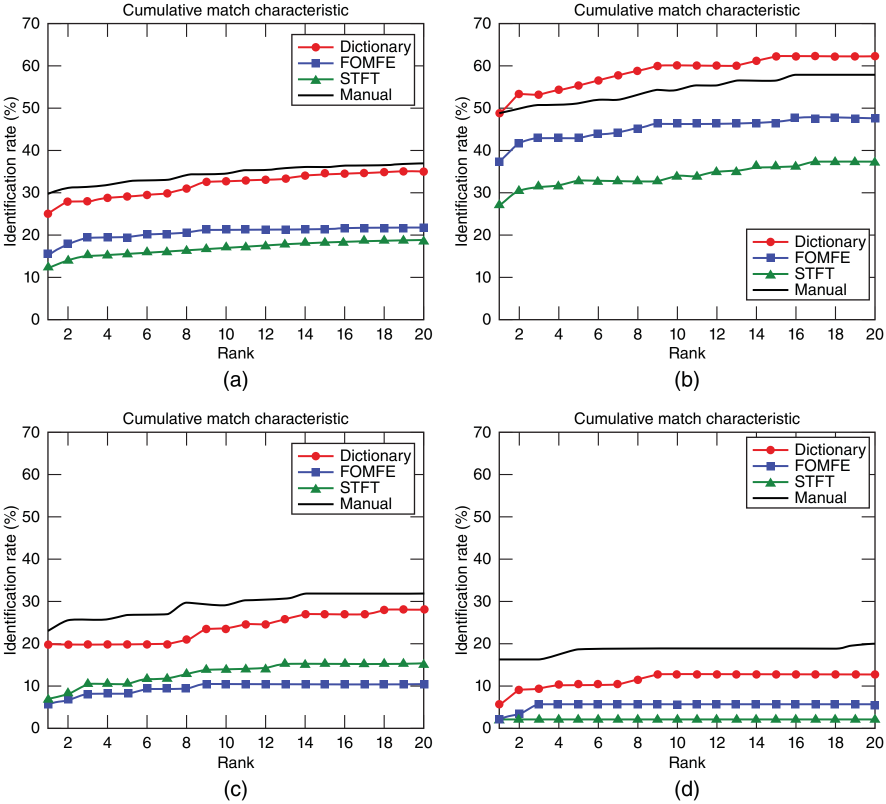


Figure 15.10 CMC curves of three orientation field estimation algorithms and the manual markup of orientation field on the NIST SD27 latent database: (a) all (258 latents), (b) good quality (88 latents), (c) bad quality (85 latents), and (d) ugly quality (85 latents) (*Source*: Feng et al. 2013, Figure 13, p. 933. Reproduced by permission of IEEE).

STFT) on latents of all three quality levels. Three examples given in Figure 15.11 compare the enhanced latent fingerprints using the orientation fields obtained by the dictionary-based algorithm, FOMFE, and STFT.

For many latents of good quality, the dictionary-based algorithm outperforms the manual markup (see Figure 15.10(b)). Our analysis of these examples demonstrates that the

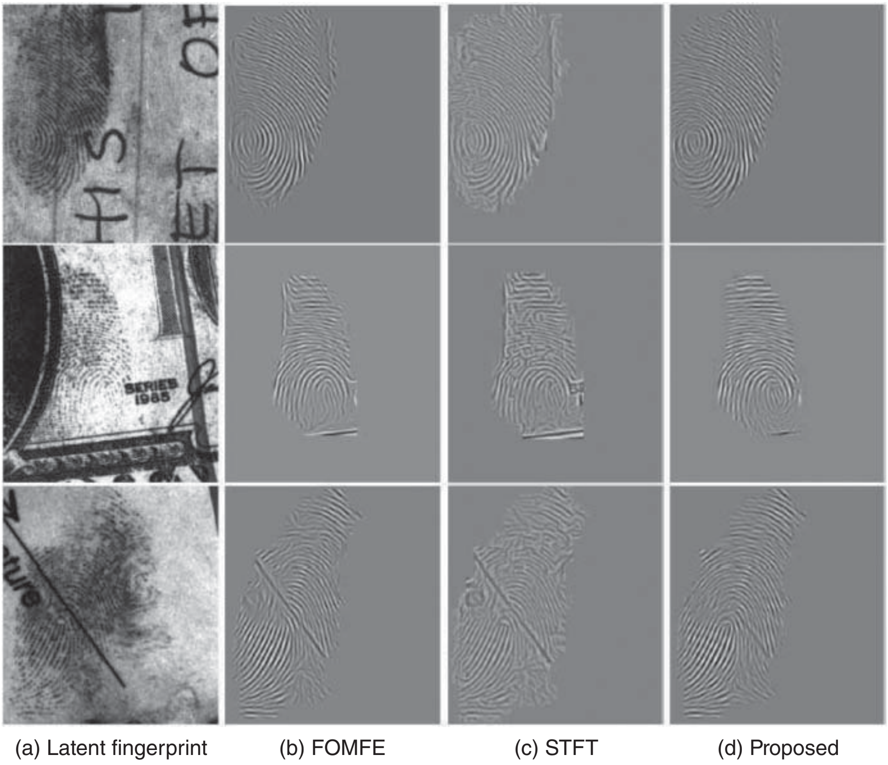


Figure 15.11 Enhanced images of three latent fingerprints in (a) using orientation fields estimated by (a) FOMFE, (b) STFT, and (c) the orientation patch dictionary-based algorithm (*Source:* Feng et al. (2013, Figure 14, p. 934. Reproduced by permission of IEEE).

dictionary-based algorithm has smaller deviation from true ridge orientation for good quality latents because it is difficult and time consuming for a fingerprint expert to accurately mark the complete orientation field in a latent. Manual markup still performs better on bad and ugly quality latents.

15.4 Latent segmentation and enhancement using ridge structure dictionary

Given the learned ridge structure dictionaries (Section 15.2.2), latent segmentation and enhancement consists of the following steps (see Figure 15.12):

1. *Decomposition*: Input latent is decomposed into cartoon and texture images using local total variations (LTVs) (Buades et al. 2010); the cartoon image, which primarily consists of structured noise, is discarded.
2. *Coarse-level estimation*: The coarse-level dictionary is used to estimate orientation and frequency fields on the texture image and assess coarse-level quality of the latent.

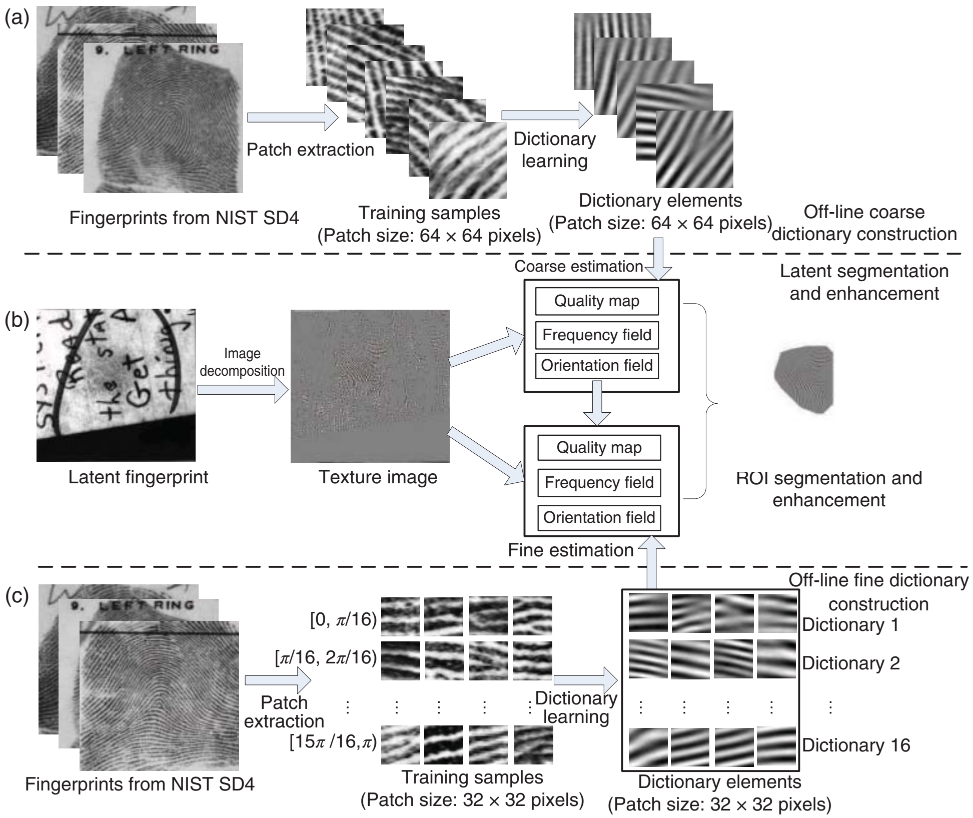

(a)
Patch extraction
Dictionary learning
Fingerprints from NIST SD4
Training samples (Patch size: 64 × 64 pixels)
Dictionary elements (Patch size: 64 × 64 pixels)
Off-line coarse dictionary construction
Coarse estimation
Latent segmentation and enhancement
(b)
Image decomposition
Quality map
Frequency field
Orientation field
Latent fingerprint
Texture image
Quality map
Frequency field
Orientation field
ROI segmentation and enhancement
Fine estimation
(c)
Off-line fine dictionary construction
[0, π/16)
[π/16, 2π/16)
Patch extraction
Dictionary learning
[15π /16,π)
Dictionary 1
Dictionary 2
Dictionary 16
Fingerprints from NIST SD4
Training samples (Patch size: 32 × 32 pixels)
Dictionary elements (Patch size: 32 × 32 pixels)

Figure 15.12 Overview of latent segmentation and enhancement algorithm based on ridge structure dictionary. The off-line dictionary learning (a and c) and online latent segmentation and enhancement stage (b) are shown (*Source:* Cao et al. (2014), Figure 3, p. 1850. Reproduced by permission of IEEE).

3. *Fine-level estimation*: Using coarse-level orientation field, select one fine-level dictionary out of the 16 fine-level dictionaries for each image patch in the texture image; this gives fine-level orientation and frequency fields and ridge quality map.

4. *Segmentation and enhancement*: The coarse-level quality map and fine-level quality map are fused for latent segmentation. In the foreground of texture image, a Gabor filter tuned to the orientation and frequency fields obtained in steps 2 and 3 is applied for latent enhancement.

15.4.1 Latent image decomposition

A latent fingerprint image, f, is decomposed into a cartoon (piece-wise smooth) image and a texture (oscillatory) image. The texture image primarily includes the ridge structure patterns, so it is kept for further latent segmentation and enhancement, while the cartoon image, viewed as structured noise, is discarded. We adopt the nonlinear decomposition method based on LTV proposed by Buades et al. (2010). Figure 15.14(b) shows the texture component of three different latent images shown in Figure 15.14(a); most of the structured noise in latents has been successfully removed by excluding the cartoon image and only the friction ridge pattern is retained in the texture image.

15.4.2 Coarse estimates of ridge quality, orientation, and frequency

15.4.2.1 Sparse coding and patch quality

The texture image is divided into overlapping patches of size 64×64 pixels (P_L^c). Each patch has 64×48 or 48×64 overlapping pixels, with each of its four-connected neighboring blocks. Each patch $p \in P_L^c$ is converted into a vector by row concatenation and normalized with a mean of zero and standard deviation of one. The sparse code α of p with respect to coarse-level dictionary D^c is obtained using orthogonal matching pursuit (Mallat and Zhang 1993). In general, the reconstructed patch $\hat{p}$ is close to p if p is a fingerprint patch. In order to measure the similarity between $\hat{p}$ and p, we have used the structural similarity index SSIM$(p, \hat{p})$ (Wang et al. 2004).

Figure 15.13 compares the reconstructed patches using different values of dictionary elements and SSIM indices for two fingerprint patches (the top and middle rows) and one non-fingerprint patch (the bottom row). We observe that the value of SSIM indicates the quality of a patch in terms of fingerprint ridges. The quality of patch p, Q_p, is therefore defined by

$$Q_p = \text{SSIM}(p, \hat{p}). \tag{15.6}$$

A single dictionary element is selected for each image patch for the reconstruction. This is because (i) with just one element, the sparse code and SSIM index are easy to compute,

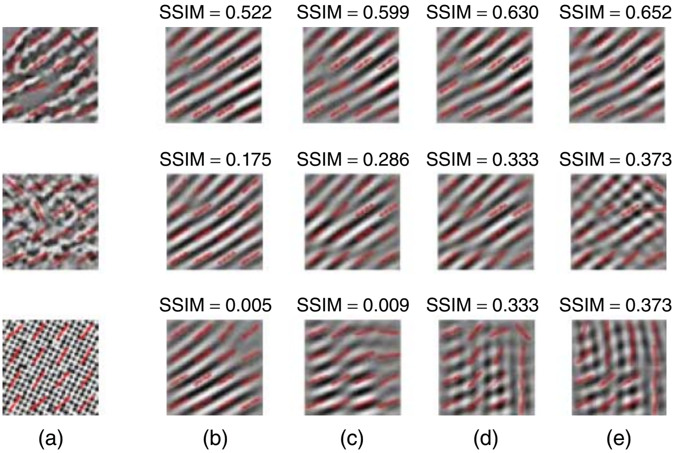


Figure 15.13 Patch reconstruction results (overlaid on orientation field) with different number of dictionary entries, T_1. (a) Texture component of a high-quality fingerprint patch (top), low-quality fingerprint patch (the middle), and non-fingerprint patch (the bottom), (b), (c), (d), and (e) are the reconstruction results when $T_1 = 1, 2, 3, 4$, respectively. The SSIM indices between the given patch (column (a)) and the reconstructed patch with different value of T_1 are shown above the reconstructed patches.

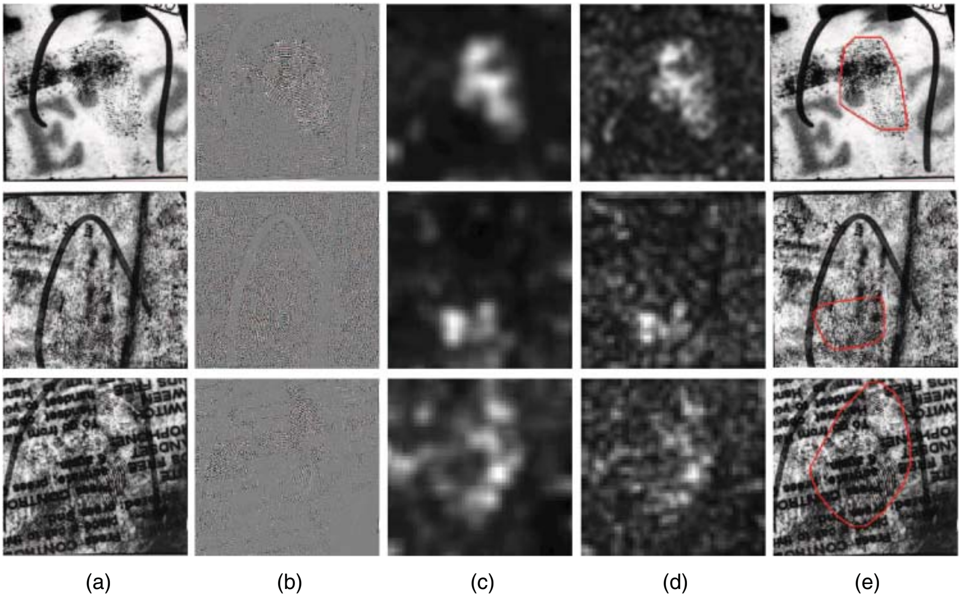


Figure 15.14 Illustration of latent fingerprint segmentation. (a) Gray-scale latent images, (b) texture component images, (c) coarse-level quality maps, (d) fine-level quality maps, and (e) segmentation results shown overlaid on the gray-scale latent images. The top, middle, and bottom latent fingerprints in column (a) are of good, bad, and ugly quality in NIST SD27. The contrast of the middle and bottom latent fingerprints has been adjusted for better visual quality.

and (ii) the orientation and frequency fields of $\hat{p}$ can be computed off-line since $\hat{p}$ is simply one of dictionary elements. Figure 15.14(c) shows some examples of coarse-level quality maps when a single dictionary element is retrieved for reconstruction.

15.4.2.2 Ridge quality map, and orientation and frequency fields estimation

As shown in Figure 15.13, the reconstructed patches have better ridge quality. The orientation field and frequency field of patch p in the latent image can be obtained from the reconstructed patch $\hat{p}$. For a block b (16×16 pixels) in the latent covered by multiple patches, let $\{q_i, \theta_i, f_i\}$ be the ridge quality, orientation, and frequency of the ith patch covering the block b. The coarse estimates of ridge quality Q_b^c, orientation θ_b^c, and frequency f_b^c for block b are computed as:

$$Q_b^c = \frac{1}{n_b} \sum_{i=1}^{n_b} q_i, \tag{15.7}$$

$$\theta_b^c = \frac{1}{2} \tan^{-1} \left(\sum_{i=1}^{n_b} q_i \sin 2\theta_i, \sum_{i=1}^{n_b} q_i \cos 2\theta_i \right), \tag{15.8}$$

$$f_b^c = \frac{1}{\sum_{i=1}^{n_b} q_i} \sum_{i=1}^{n_b} q_i f_i, \tag{15.9}$$

where n_b is the number of patches covering the block b. In this case, a higher weight (q_i) is assigned to the patch with better ridge and valley structures. Other elaborate weighting strategies may lead to better results.

15.4.3 Fine estimates of ridge quality, orientation, and frequency

While the coarse-level dictionary is robust to local noise, it cannot extract detailed ridge information. Instead, small patch size dictionaries can be used to compute the fine-level quality map and fine-level orientation and frequency fields. The texture image is divided into smaller overlapping patches of size 32×32 pixels (P_L^f). Each patch has 32×16 or 16×32 overlapping pixels with each of its four-connected neighboring blocks. All patches in P_L^f are normalized with a mean of zero and standard deviation of one. For each patch $p \in P_L^f$, a dominant orientation θ is used to select the corresponding orientation-specific dictionary D_k^f, where $k = \lceil \frac{16 \times \theta}{\pi} \rceil$ and $\lceil \cdot \rceil$ is the ceiling operator. The closest dictionary element $\hat{p}$ to p in D_k^f is selected, and the quality of the patch p is determined by the SSIM index between p and $\hat{p}$. For each 16×16 block b in the latent, the fine-level quality Q_b^f, orientation θ_b^f, and frequency f_b^f are obtained from the covering patches using Equations (15.7)–(15.9).

15.4.4 Segmentation and enhancement

The final quality map Q is computed as the average of the coarse-level quality map and fine-level quality map. Q is then normalized to the range [0,1] and a global threshold T_Q, determined by Otsu's method (Otsu 1979) is used to binarize the normalized quality map. The blocks with normalized quality less than T_Q are regarded as background, otherwise foreground. Morphological operations (dilation and opening) are then applied to remove small foreground blocks and fill holes inside the foreground. Finally, the convex hull of the set of foreground blocks determines the final segmentation result. Figure 15.14(e) shows the segmentation results for latent fingerprints in NIST SD27.

In the foreground region, the texture image of a latent obtained from the decomposition is enhanced by Gabor filtering (Hong et al. 1998), where the orientation and frequency parameters of the filter are tuned based on the fine-level orientation field (θ^f) and the average frequency of coarse-level frequency field and fine-level frequency field ($\frac{f^f + f^c}{2}$); the standard deviations of the Gaussian envelope in x and y directions are set to 4.

15.4.5 Experimental results

Two latent databases are used for performance evaluation: NIST SD27 and the West Virginia University latent database (WVU DB) (iPRoBe 2014). The NIST SD27 contains 258 latent fingerprints with their mated rolled fingerprints. The WVU DB contains 449 latent fingerprints with their mated rolled fingerprints and an additional 4,290 rolled fingerprints. All these latent fingerprint images are 500 ppi images. The algorithm was implemented in MATLAB and C/C++ and run on a dual-core 2.66 GHz, 4GB RAM machine running a Windows 7 operating system. The average computation time for segmentation and enhancement per latent is about 2.6 seconds for latents in NIST SD27 and 1.6 seconds for latents in WVU DB.

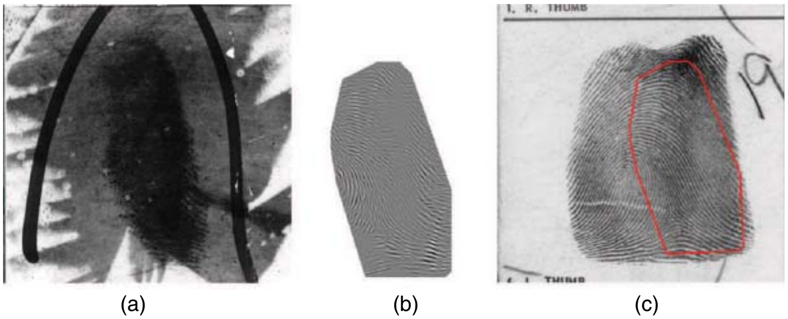


Figure 15.15 An example of latent segmentation and enhancement by the proposed algorithm. (a) A latent fingerprint (U286 from NIST SD27); (b) fully automatic segmentation of (a) by the proposed algorithm; (c) the true mate (rolled print) of (a) with the segmentation boundary in (c) outlined on the mate. By feeding the original latent in (a) and the segmented and enhanced latent in (b) into a commercial off-the-shelf (COTS) latent matcher (with a background database of 31,997 reference prints), the mated print is retrieved at ranks 4,152 and 2, respectively.

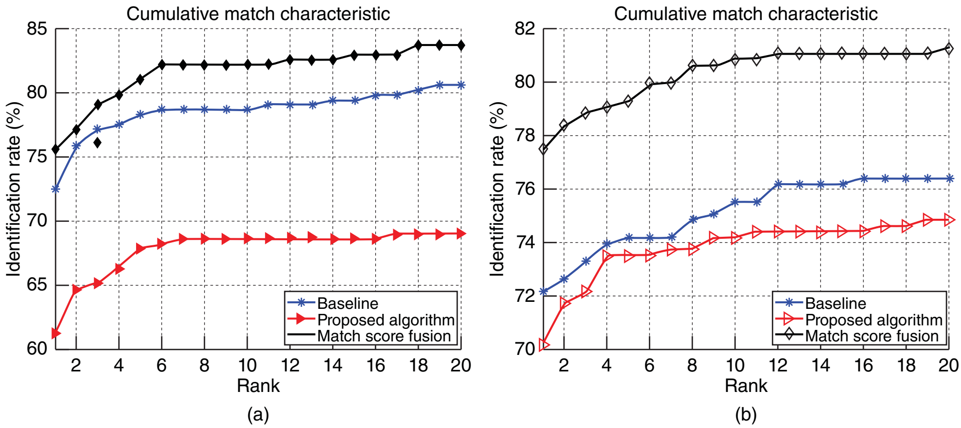


Figure 15.16 CMC curves of latent fingerprint identification with the COTS latent matcher on (a) NIST SD27 and (b) WVU DB (*Source:* adapted from Cao et al. (2014), Figure 12, p. 1857. Reproduced by permission of IEEE).

The ultimate goal of fingerprint segmentation and enhancement of latent images is to improve the latent matching performance. To make the latent matching problem more realistic and challenging, the background database size is expanded to 31,997 rolled fingerprints by including 27,000 rolled fingerprints in the NIST SD14, 258 rolled fingerprints in NIST SD27, and 4,739 rolled fingerprints in WVU DB.

The segmentation and enhancement algorithm based on ridge structure dictionary is evaluated by a state-of-the-art latent matcher (COTS) to determine whether the proposed algorithm is able to boost latent matching performance. The match scores from the two input images (original latent image in Figure 15.15(a) and segmented and enhanced image in Figure 15.15(b)) are fused by a weighted sum method; the weights for original latent image template and segmented and enhanced image templates are empirically set as 0.7 and 0.3, respectively. The resulting CMC curves for the COTS latent matcher on NIST SD27 and WVU DB are shown in Figure 15.16. The fusion of match scores from these two inputs improves the rank-1 identification rate of COTS latent matcher on both NIST SD27 and WVU DB. Fusing outputs of diverse search templates extracted from different segmentation and enhancement algorithms appears to be a good strategy to boost latent matching performance.

15.5 Conclusions and future work

Automatic Fingerprint Identification Systems (AFIS) have achieved extremely high matching accuracies in tenprint searches (rolled or plain fingerprints). For this reason, almost every law enforcement agency in the world relies on the use of AFIS to identify suspects and criminals. Further, there is a growing use of fingerprints to conduct background searches of applicants for visa and other government-issued secure documents. However, latent fingerprint search is still a challenging problem due to the presence of complex background noise and poor quality of friction ridge structure that is typical for latent fingerprint images found at crime scene investigations. For this reason, human intervention, such as manual markup of minutiae and singular points, is common practice for latent fingerprint identification.

A fully automatic latent identification (“lights-out” mode) is highly desired to alleviate the concerns about repeatability and reproducibility of latent examiners’ performance and increase the throughput of the latent matching process. Automatic feature extraction is one of the most crucial steps in “lights-out” latent identification. In this chapter, we have summarized the role of two types of dictionaries, orientation patch dictionary and ridge structure dictionary, as representations of prior knowledge about fingerprint patterns. We show how these dictionaries can be used in latent segmentation and enhancement. The orientation patch dictionary is used to update the initial orientation field in the input ROI. The ridge structure dictionary is used for ROI segmentation and enhancement. Experimental results on two different latent fingerprint databases demonstrate the advantages of our dictionary-based approach for fingerprint segmentation and enhancement.

In order to further improve the dictionary-based algorithms, we need to address the following issues:

1. Instead of simply using local orientation patch dictionary, global orientation field dictionary may further improve the accuracy of orientation field estimation.

2. A multiresolution approach should be considered to construct orientation patch dictionaries for both small and large friction ridge areas.

3. A robust ridge quality estimation for fingerprint images with low contrast (as in “dry” fingerprints) and background line structure noise is needed.

References

Aharon M, Elad M and Bruckstein A 2006 K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing* **54**(11), 4311–4322.

Apple, Inc. 2014 iPhone 5s: About Touch ID security https://support.apple.com/kb/HT5949.

Ashbaugh DR 1999 *Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology*. CRC Press.

Blake A, Kohli P and Rother C (eds) 2011 *Markov Random Fields for Vision and Image Processing*. MIT Press.

Buades A, Le T, Morel JM and Vese L 2010 Fast cartoon + texture image filters. *IEEE Transactions on Image Processing* **19**(8), 1978–1986.

Cao K, Liu E and Jain A 2014 Segmentation and enhancement of latent fingerprints: a coarse to fine ridgestructure dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(9), 1847–1859.

Chikkerur S, Cartwright AN and Govindaraju V 2007 Fingerprint enhancement using STFT analysis. *Pattern Recognition* **40**(1), 198–211.

Department of Homeland Security 2014 Office of biometric identify management http://bias.dhs.gov/obim.

Dror IE, Wertheim K, Fraser-Mackenzie P and Walajtys J 2011 The impact of human-technology cooperation and distributed cognition in forensic science: biasing effects of AFIS contextual information on human experts. *Journal of Forensic Sciences* **57**(2), 343–352.

Duta N, Jain AK and Mardia KV 2002 Matching of palmprints. *Pattern Recognition Letters* **23**(4), 477–485. In Memory of Professor E.S. Gelsema.

Elad M and Aharon M 2006 Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing* **15**(12), 3736–3745.

FBI- NGI 2014 FBI-Next Generation Identification (NGI) http://www.fbi.gov/about-us/cjis/fingerprints_biometrics/ngi.

Feng J, Zhou J and Jain AK 2013 Orientation field estimation for latent fingerprint enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **54**(4), 925–940.

FVC200 2002 FVC2002 is the Second International Competition for Fingerprint Verification Algorithms http://bias.csr.unibo.it/fvc2002/.

Garris MD, Tabassi E, Wilson CI, McCabe RM, Janet S and Watson CI 2004 NIST fingerprint image software 2.

Hawthorne M 2008 *Fingerprints: Analysis and Understanding*. CRC Press.

Hong L, Wan Y and Jain A 1998 Fingerprint image enhancement: algorithm and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8), 777–789.

Indovina MD, Dvornychenko V, Hicklin RA and Kiebuzinski GI 2012 Evaluation of latent fingerprint technologies: extended feature sets (evaluation 2). *Technical Report NISTIR 7859, NIST*.

Indovina MD, Dvornychenko VN, Tabassi E, Quinn GW, Grother PJ, Meagher S and Garris MD 2009 ELFT phase II - an evaluation of automated latent fingerprint identification technologies. *NISTIR 7577*.

iPRoBe 2014 http://www.cse.msu.edu/ rossarun/i-probe/.

Jain A and Feng J 2009 Latent palmprint matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(6), 1032–1047.

Jain A, Hong L and Bolle R 1997 On-line fingerprint verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**, 302–314.

Jiang X 2000 Fingerprint image ridge frequency estimation by higher order spectrum *IEEE International Conference on Image Processing*, Vol. 1, pp. 462–465.

Karimi-Ashtiani S and Kuo CC 2008 A robust technique for latent fingerprint image segmentation and enhancement *IEEE International Conference on Image Processing*, pp. 1492–1495.

Lian XC, Li Z, Wang C, Lu BL and Zhang L 2010 Probabilistic models for supervised dictionary learning *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2305–2312.

Liao S, Jain AK and Li SZ 2013 Partial face recognition: alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(5), 1193–1205.

Mairal J, Bach F, Ponce J, Sapiro G and Zisserman A 2008a Discriminative learned dictionaries for local image analysis *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.

Mairal J, Elad M and Sapiro G 2008b Sparse representation for color image restoration. *IEEE Transactions on Image Processing* **17**(1), 53–69.

Mallat S and Zhang Z 1993 Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing* **41**(12), 3397–3415.

Maltoni D, Maio D, Jain A and Prabhakar S 2009 *Handbook of Fingerprint Recognition* 2nd ed. Springer-Verlag.

Mardia K, Baczkowski A, Feng X and Hainsworth T 1997 Statistical methods for automatic interpretation of digitally scanned finger prints. *Pattern Recognition Letters* **18**(11–13), 1197–1203.

Neurotechnology Inc. 2012 Verifinger http://www.neurotechnology.com/verifinger.html.

NIST Special Database 27 2014 Fingerprint Minutiae from Latent and Matching Tenprint Images http://www.nist.gov/srd/nistsd27.cfm.

NIST Special Database 4 2014 NIST 8-Bit Gray Scale Images of Fingerprint Image Groups(FIGS) http://www.nist.gov/srd/nistsd4.cfm.

Otsu N 1979 A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**(1), 62–66.

Planning Commission, Goverment of India 2014 Unique Identification Authority of India http://uidai.gov.in/.

Short NJ, Hsiao MS, Abbott AL and Fox EA 2011 Latent fingerprint segmentation using ridge template correlation *4th International Conference on Imaging for Crime Detection and Prevention*, pp. 1–6.

Tabassi E, Wilson C and Watson C 2004 Fingerprint image quality. *NISTIR 7151*.

Trauring M 1963 Automatic comparison of finger-ridge patterns. *Nature* **197**, 938–940.

Turroni F, Maltoni D, Cappelli R and Maio D 2011 Improving fingerprint orientation extraction. *IEEE Transactions on Information Forensics and Security* **6**(3), 1002–1013.

Ulery BT, Hicklin RA, Buscaglia J and Roberts MA 2011 Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences of the United States of America* **108**(19), 7733–7738.

Ulery BT, Hicklin RA, Buscaglia J and Roberts MA 2012 Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS One* **7**(3), e32800.

Wang Z, Bovik A, Sheikh H and Simoncelli E 2004 Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600 –612.

Wang Y, Hu J and Phillips D 2007 A fingerprint orientation model based on 2d Fourier expansion (FOMFE) and its application to singular-point detection and fingerprint indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(4), 573–585.

Wilson C, Hicklin RA, Korves H, Ulery B, Zoepfl M, Bone M, Grother P, Micheals R, Otto S and Watson C 2004 Fingerprint vendor technology evaluation 2003: Summary of results and analysis report. *NISTIR 7123*.

Wright J, Yang A, Ganesh A, Sastry S and Ma Y 2009 Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(2), 210–227.

Yoon S, Feng J and Jain A 2011 Latent fingerprint enhancement via robust orientation field estimation *2011 International Joint Conference on Biometrics (IJCB)*, pp. 1–8.

Zhang J, Lai R and Kuo CC 2013 Adaptive directional total-variation model for latent fingerprint segmentation. *IEEE Transactions on Information Forensics and Security* **8**(8), 1261–1273.

Part V

BIOINFORMATICS

16

Do protein structures evolve around 'anchor' residues?

Colleen Nooney, Arief Gusnanto, Walter R. Gilks and Stuart Barber
Department of Statistics, School of Mathematics, University of Leeds, Leeds, UK

16.1 Introduction

16.1.1 Overview

The structure of a protein is constrained by its function. Sequence alignments from homologous proteins that are from a range of species provide information on these evolutionary constraints. The analysis of correlated mutations within multiple-sequence alignments can be used to predict residues that are in proximity in three-dimensional space. We study the co-evolution of protein sequences and structure to distinguish between the residue correlations that correspond to structural proximity and potential confounding residue correlations. Confounding residue correlations can occur as a result of noise or other biological evolutionary constraints (Marks et al. 2011).

The exploratory data analysis reported here focusses on the trypsin protein family. The structures were aligned using a multiple structural alignment algorithm, MUSTANG (Konagurthu et al. 2006), to determine how the structure of the family has evolved. Calculating basic summary statistics on the resulting aligned distance matrices revealed an interesting result. We discovered a set of residues where the distance between these specific residues and every other in the structure is highly conserved across all of the

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

structures in the protein family. These residues appear to hold the structure of the trypsin protein family in place like anchors.

We conduct a series of tests to determine the validity and origin of the intriguing concept of ‘anchor’ residues and the resulting conclusions drawn about the trypsin protein family following their discovery. However, many of these tests proved inconclusive or provided conflicting evidence. Therefore, the question is still open: are the anchor residues artefacts?

16.1.2 Protein sequences and structures

Proteins are biological macromolecules comprised of polypeptide chains; these in turn are made up of amino-acid residues. Figure 16.1 displays the chemical structure common to all amino acids, where R represents the unique side chain of the 20 standard amino acids. The R group is connected to the alpha carbon, or C_α atom. To form the polypeptide chain,

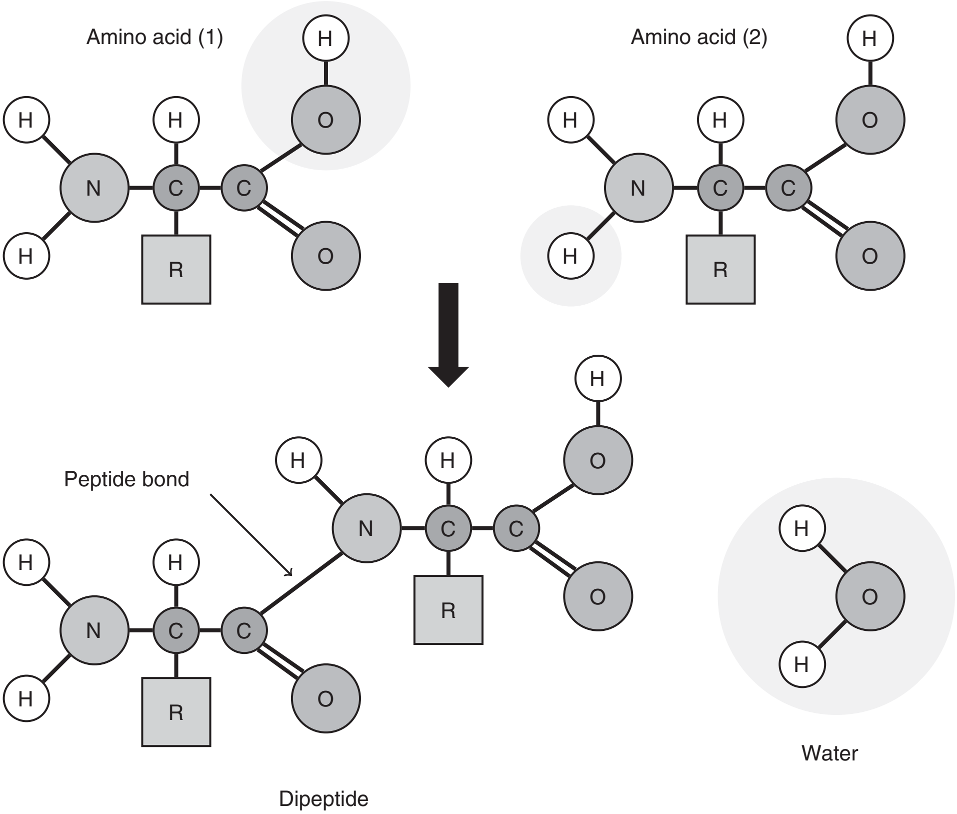


Figure 16.1 A two-dimensional ball-and-stick model of peptide bond formation between two amino acids. Atoms are represented by circles and bonds are lines between them, where double bonds are indicated by two parallel lines. Nitrogen, Carbon, Oxygen and Hydrogen are represented by ‘N’, ‘C’, ‘O’ and ‘H’, respectively. The unique side chains or ‘R’ groups of the two amino acids are represented by a square. Peptide bonds are formed when the carboxyl group of one amino acid reacts with the amino group of another resulting in the loss of a water molecule, as shown in the lower panel.

the amino-acid residues are combined by peptide bonds, resulting in the loss of a water molecule for each link.

The complex structure of a protein is determined by four different levels of folding, known as the primary, secondary, tertiary and quaternary structures. The primary structure is the sequence of amino-acid residues of each polypeptide chain. Each of the 20 amino acids is represented by a distinct single-letter code.

The secondary structures of a protein are the regions of the polypeptide chain that are organised into regular structures identified as alpha helices and beta-pleated sheets. Alpha helices are the most common type of secondary structure. The protein chain twists into a coil held together by hydrogen bonds where the side chains of the amino acids point outwards. The helix is orientated in an anti-clockwise direction, with approximately 3.6 amino-acid residues per turn. Beta sheets are rigid planar surfaces formed when two or more strands of the protein chain lie side by side. This structure is also held together by hydrogen bonds. The side chains lie alternately above and below the plane of the surface of the beta sheet. Between the organised secondary structure regions are less structured loops and turns, which are less rigid and freer to move.

The tertiary structure of a protein describes the folding of the polypeptide chain to form its final three-dimensional shape. Interactions between the side chains of amino-acid residues hold this structure in place. Disulphide bonds or sulphur bridges occur when cysteine side chains align as a result of higher order folding.

The quaternary structure of a protein is the combination of more than one polypeptide chain. For example, dimers are proteins comprised of two polypeptide chains. The quaternary structure is held together by the same interactions as the tertiary structure. Not all proteins have a quaternary structure. Those that do not consist of one polypeptide chain are known as monomers (Branden et al. 1991).

16.2 Exploratory data analysis

16.2.1 Trypsin protein family

Trypsin is a protein of the serine protease family involved in the digestive processes of most vertebrates. It is produced in the pancreas and breaks proteins down into smaller proteins to be absorbed through the lining of the small intestine. Trypsin has many applications; it is used in many biotechnological processes, the food industry, biological research, as a treatment for inflammation and in microbial form to dissolve blood clots (Bateman et al. 2004). Due to its multiple varied uses, over 2000 trypsin structures have been experimentally determined over a wide variety of species. A typical trypsin structure is displayed in Figure 16.2, displayed using the molecular visualisation software, Jmol (Jmol: an open source java viewer for chemical structures in 3D. http://www.jmol.org/). Trypsin is in the all-beta class of proteins because it consists entirely of beta sheets, with the exception of two alpha helices that are isolated on the outside of the structure. Trypsin contains two beta barrels that lie perpendicular to each other in the structure. The beta barrels are a closed structure formed when the beta sheets twist such that the first strand is hydrogen bonded to the last.

Trypsin structures were downloaded from the Protein Data Bank (PDB) (Berman et al. 2000). After filtering out inappropriate structures, such as low-resolution structures and corrupt PDB files, our final data consists of 83 trypsin chains, originating from a variety of species. The protein chains were aligned in order to identify regions of similarity throughout

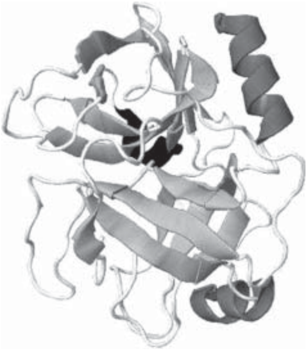


Figure 16.2 Ribbon representation of a trypsin molecule (Protein Data Bank (PDB) accession code: 1S5S) displayed with the molecular visualisation software, Jmol. The secondary structures are coloured; dark grey indicates an alpha helix, light grey indicates a beta sheet and the black helix is a 3_{10} helix; a helix with three residues per turn rather than 3.6.

evolution. There are two types of alignment: sequence and structural. Sequence alignments are constructed based on the similarity between amino-acid residues and their physiochemical properties, while structural alignments use shape and three-dimensional conformation to align the atomic coordinates of the structures. Structure alignments are of interest because the structure of a protein family evolves more gradually than the amino-acid sequence and is, therefore, more conserved. Due to the size of the sample, the MUSTANG multiple structural alignment algorithm (Konagurthu et al. 2006) was used as it is one of the few structural alignment algorithms capable of operating with a large number of structures.

16.2.2 Multiple structure alignment

An overview of the MUSTANG procedure is given in Figure 16.3. The main steps in the procedure are as follows. The MUSTANG method first tries to find structural similarity in pairwise fragments of the structures before building the multiple structure alignment . Each pair of structures is initially scored using root-mean-square deviation (RMSD) in order to find similar substructures. The RMSD is a measure of the average distance between the atoms of superimposed structures. The individual residue alignments are then scored using a similarity measure that is closely based on the elastic similarity function proposed by Holm and Sander (1993). These scores are used to align each pair of structures by a dynamic programming algorithm. The pairwise alignment scores are then recalculated in the context of all of the structures. This is achieved by taking every structure as an intermediate for each pairwise alignment. The more intermediate structures that support the alignment of a pair of residues, the higher the score assigned to them. The multiple structure alignment is finally obtained following a binary guide tree constructed using the neighbor-joining method (Saitou and Nei 1987) applied to the similarity scores.

Berbalk et al. (2009) and Konagurthu et al. (2006) compare MUSTANG with other multiple structure alignment algorithms; POSA, CE-MC, MALECON and MultiProt. According to Konagurthu et al. (2006), MUSTANG performs as well as the other alignment tools for closely related proteins and outperforms them for more distantly related proteins or proteins that exhibit conformational changes. Berbalk et al. (2009) supports the conclusion that

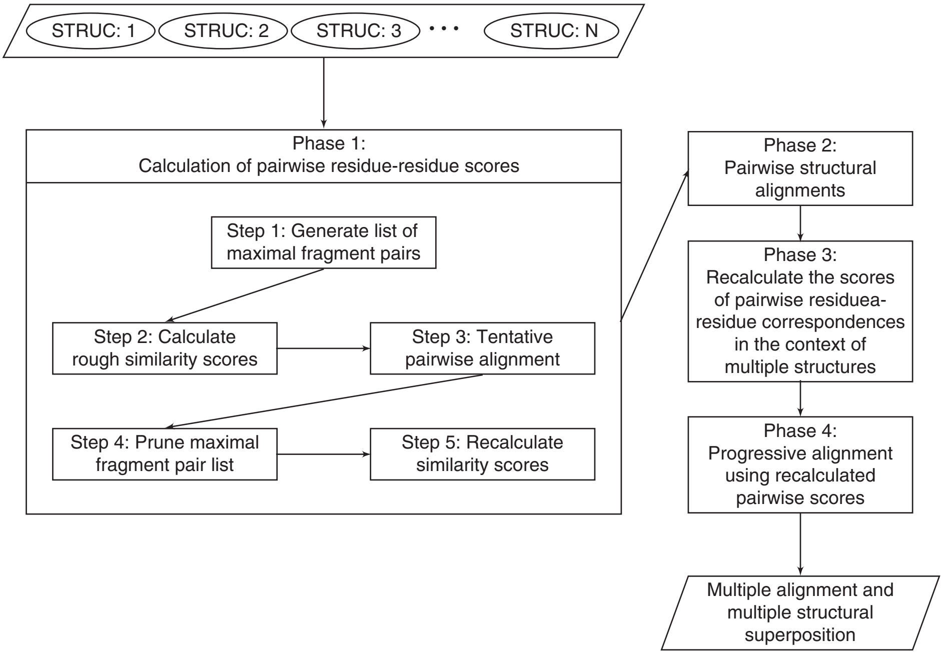

STRUC: 1
STRUC: 2
STRUC: 3
• • •
STRUC: N
Phase 1:
Calculation of pairwise residue-residue scores
Step 1: Generate list of maximal fragment pairs
Step 2: Calculate rough similarity scores
Step 3: Tentative pairwise alignment
Step 4: Prune maximal fragment pair list
Step 5: Recalculate similarity scores
Phase 2:
Pairwise structural alignments
Phase 3:
Recalculate the scores of pairwise residuea-residue correspondences in the context of multiple structures
Phase 4:
Progressive alignment using recalculated pairwise scores
Multiple alignment and multiple structural superposition

Figure 16.3 An overview of the MUSTANG algorithm (*Source:* adapted from Konagurthu et al. 2006, Figure 2, p. 562. Reproduced with permission of John Wiley and Sons).

MUSTANG performs as well as other alignment tools when the structures have high structural similarity but suggests that there is room for improvement when structures are more distantly related.

MUSTANG has several disadvantages; it can be very temperamental in what can be aligned and also only uses the information in the C$_\alpha$ coordinates of the structures and the distances between them, the information contained in the amino-acid sequence is ignored completely.

The output of the alignment is a multiple-sequence alignment constructed using the structural alignment of the chains. We prepared the alignment for subsequent analysis by removing all positions in the alignment where more than 20% of the entries consist of gaps. (Gaps are introduced in alignments where insertions or deletions are predicted to have occurred throughout evolution.) For smaller samples, MUSTANG produces a PDB file containing the coordinates for the superimposed structures; this can be visualised using Jmol (Jmol: an open source java viewer for chemical structures in 3D. http://www.jmol.org/). Visual analysis is impractical with such a large number of structures; instead, we considered the distances between the residues in the superimposed structures.

16.2.3 Aligned distance matrix analysis

The three-dimensional shape of a protein can be summarised by its residue–residue distances. A distance matrix for a protein structure, k, contains the Euclidean distance, $d_{i,j}^{(k)}$, between the C$_\alpha$ atoms of each amino-acid residue pair, i and j. The positions in the distance matrices can be aligned, or superimposed, using the MUSTANG alignment to analyse corresponding distances across the structures. The alignment produced by MUSTANG respects the sequence order of the amino acids.

There are 219 alignment positions in the MUSTANG alignment of the 83 trypsin structures downloaded from the PDB, resulting in a $219 \times 219 \times 83$ data array. This large data structure can be summarised by calculating a measure of location and divergence for every distance across the aligned structures. We achieved this by calculating a weighted median and a weighted interquartile range, where the weights are calculated using the method of Henikoff and Henikoff (1994) as follows:

- For each position in the alignment, divide a total weight of one evenly between the unique letter types in that position.
- Divide the weight that has been assigned to each letter type between the number of that letter type in that position.
- For each sequence, sum the weights that have been assigned at each position.
- Normalise the sequence weights to sum to one.

Sequences from the same species are likely to be very similar, whereas sequences from more diverged species differ more. If all of the sequences are weighted equally, then information may be lost when there are many similar sequences due to independent information from the more diverged sequences being diluted. The sequences are weighted so that very similar sequences are down-weighted and unusual sequences are up-weighted. We constructed a median matrix, $\tilde{d}$, and divergence matrix, div, using the aligned distance matrices; the (i, j)th element of each of these matrices is given by

$$\begin{aligned}\tilde{d}_{i,j} &= \text{MED}(d_{i,j}^{(k)}),\\ \text{div}_{i,j} &= \text{IQR}(d_{i,j}^{(k)}),\end{aligned}$$

for $i, j = 1, \ldots, 219$ and $k = 1, \ldots, 83$, where IQR and MED are the weighted interquartile range and weighted median.

To assess the relationship between the median and divergence matrices, they are plotted against each other in Figure 16.4. There are a vast number of data points as a result of the size of the matrices, $219^2 = 47\,691$ data points; however, there does not appear to be an obvious relationship between the divergence and the median. Intuitively it might be assumed that a larger median would correspond to a larger divergence, since the distance between the residues is larger. However, only a handful of points exhibit this property, suggesting that for the majority of the sample the overall framework of the structures is very similar. Interestingly, there are a collection of points where the divergence is high while the median is very low. This pattern corresponds to the scenario where the distance between the two residues are small, yet there is a lot of variation in the corresponding distances across the structures, suggesting a different local structure for some of the sample.

Each row (and column) of the median and divergence matrices corresponds to a position in the structural alignment. This is plotted in Figure 16.5. The bars appear as a result of many points plotted close together. The plot of divergence against position in Figure 16.5(b) shows that there are positions in the alignment where the range of divergences is low as indicated by distinct troughs between the peaks. This suggests that there are residues where

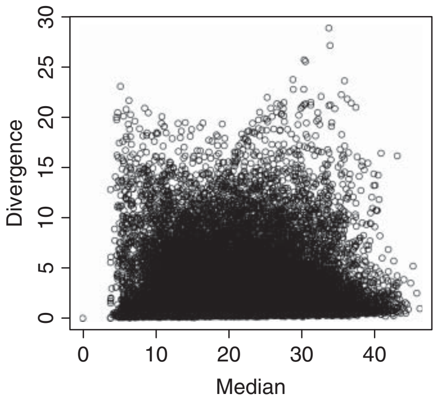


Figure 16.4 Plot of median aligned residue–residue distance against the divergence between the distances for each pair of residues, for the MUSTANG structural alignment of the trypsin sample.

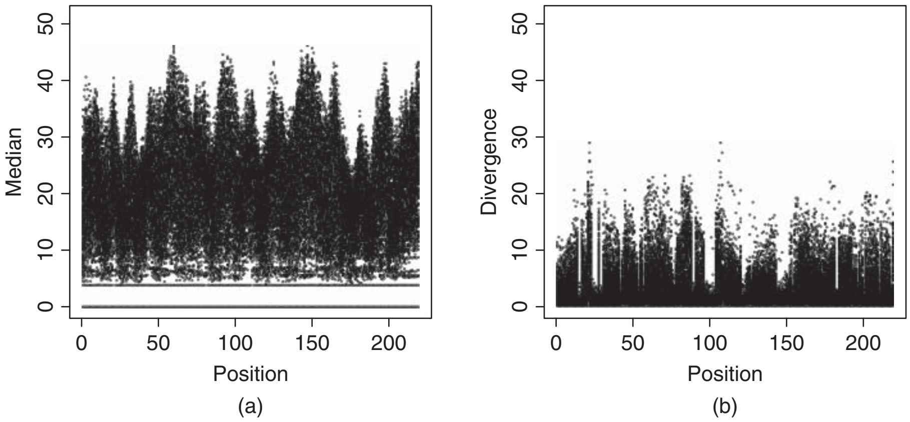


Figure 16.5 Plots of the rows of the median and divergence matrices calculated from structurally aligned distance matrices of the trypsin sample. The bars appear as a result of many points plotted close together. (a) Median, $\tilde{d}_{i,j}$, of the structurally aligned distances plotted against position, i, in the alignment. (b) Divergence, $\text{div}_{i,j}$, of the structurally aligned distances plotted against position, i, in the alignment.

the distance between that residue and every other residue in the structure is conserved, across all of the structures. If this result is genuine, these residues could be used to predict the structure of proteins in the trypsin family and might also provide a basis for predicting structure from multiple-sequence alignments of other protein families.

16.2.4 Median distance matrix analysis

The median matrix is plotted as a heatmap in Figure 16.6. The heatmap is interpreted identically to a typical heatmap for a structure; small distances are represented by white, while

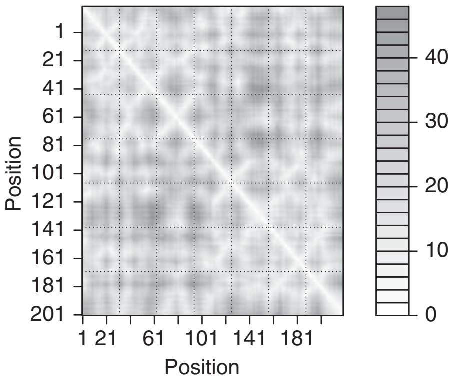


Figure 16.6 Median matrix heatmap. The median residue–residue distances are plotted in greyscale; small distances are white and large distances are dark grey.

large distances are given in dark grey. As a result, the heatmap is not dissimilar to a typical distance-matrix heatmap produced by any of the structures. This is unsurprising given that the median matrix is an average of the aligned distance matrices. This suggests that MUSTANG has produced a reasonable structure alignment and the median distance matrix is a suitable measure to be used to construct a consensus structure to represent the sample, that is, the average structure of the sample. Multidimensional scaling is a technique used to construct a configuration of data points in the Euclidean space using the distances, similarities or dissimilarities between them. The data points are assigned coordinates in n dimensions that aim to preserve the distances between them (Mardia et al. 1979). Metric multidimensional scaling can be applied to the weighted median distance matrix in order to obtain a consensus structure. We could also perform multidimensional scaling on the divergence matrix, which would allow us to see where the differences from the median structure are.

The R (R Core Team 2013) function cmdscale was used to perform metric multidimensional scaling on the median distance matrix. There are three eigenvalues that are much larger than the remaining eigenvalues. These normalised squared eigenvalues are 0.61, 0.28, 0.10, while the remaining values are close to zero, suggesting that the first three coordinates are sufficient to reproduce the median distance matrix. This is unsurprising given that we know that the distances are obtained from three-dimensional objects. The resulting coordinates are used to produce a PDB file that can be viewed in Jmol. The consensus structure is displayed superimposed over the trypsin structure 1S5S in Figure 16.7.

The consensus structure is comprised only of C_α atoms since the distance matrices used to construct it contain the distances between the C_α atoms of each residue. Despite this, Figure 16.7 shows that the configuration produced using multidimensional scaling is a good approximation of the trypsin structure 1S5S.

16.2.5 Divergence distance matrix analysis

The divergence matrix is plotted as a heatmap in Figure 16.8(a). In this case, dark grey indicates large divergences implying distances that are less conserved while white regions represent small divergences or distances that are more conserved. The scale in Figure 16.8(a) is inflated by a small area of high divergence. The low-range divergences identified in

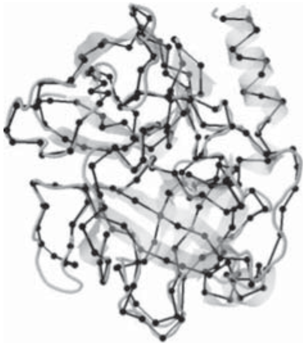


Figure 16.7 Multidimensional scaling structure of the median distance matrix, displayed in black. The C_α atoms of each position in the alignment are given by a black circle. C_α atoms corresponding to adjacent alignment positions are connected by black lines to represent the backbone of the median structure. The trypsin structure in Figure 16.2 is superimposed with the consensus structure and displayed in grey. The structures were superimposed using TM-align pairwise structural alignment algorithm (*Source:* Zhang and Skolnick 2005).

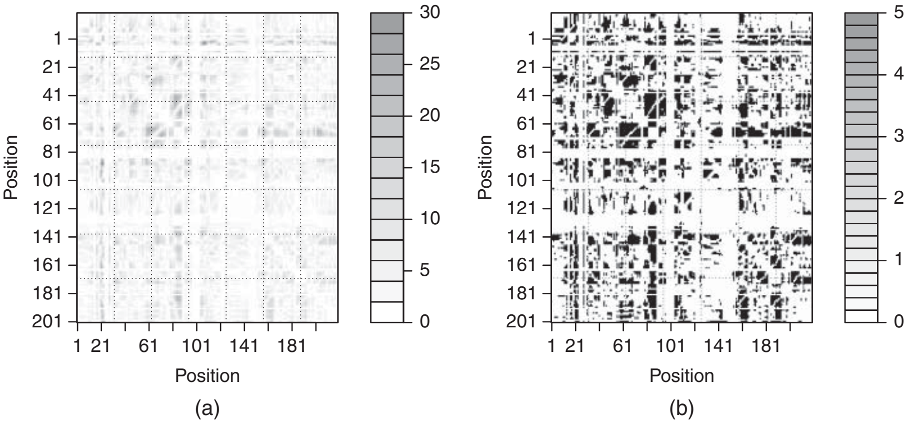


Figure 16.8 Divergence matrix heatmaps for different colour scales. The divergence between the residue–residue distances are plotted in greyscale; small divergences are white and large divergences are dark grey. (a) Divergence matrix heatmap based on the original scale. The information in white is diluted by a small amount of grey that is pulling up the scale. (b) Divergence matrix heatmap recalculated for all of the divergences that are less than 5 A°, larger divergences are blacked out.

Figure 16.5 are approximately 5 angströms (5 A°); to analyse alignment positions at this end of the scale, all divergences greater than 5 A° are coloured black and the heatmap recalculated based on the scale 0–5, as displayed in Figure 16.8(b).

The pattern of divergence at the lower end of the scale can now be visualised more clearly. There is a clear pattern of horizontal and vertical white lines running across the

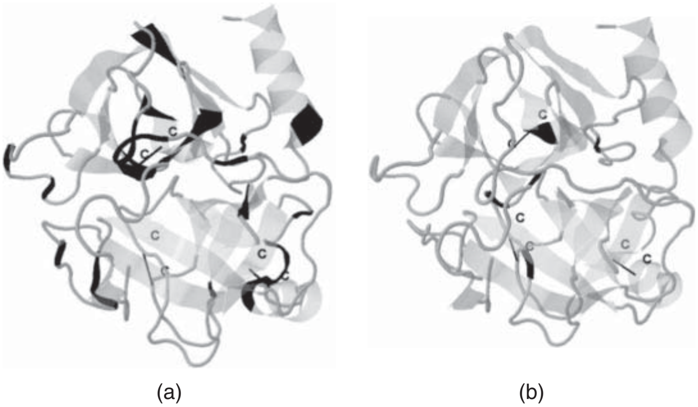


Figure 16.9 (a) Ribbon representation of a trypsin structure (PDB ID: 1JIR) identifying the location of the anchor residues, displayed in blocks of black, and the three disulphide bonds, indicated by black lines and labelled cysteine (C) residues. (b) The same structure identifying the location of functional residues, including the catalytic triad of residues and the oxyanion hole, displayed in blocks of black, and the three disulphide bonds, indicated by black lines and labelled cysteine (C) residues.

heatmap. These lines represent where in every structure the distance between one residue and every other residue is highly conserved, in agreement with the conclusions drawn from Figure 16.5. Four distinct groups of alignment positions can be identified as having a low range of divergences. These residues are of interest as they appear to be anchors for each of the structures, conserving their distances and holding them in place.

To accurately determine the positions in the multiple structure alignment corresponding to the low-range divergences, the maximum divergence in each position was analysed and a natural divide was found around $7\ \mathrm{A}^{\circ}$. It remains to identify which positions have a maximum divergence of less than $7\ \mathrm{A}^{\circ}$ and determine where these lie on each of the structures. We define an anchor residue to be any residue, i, with $\max_j \operatorname{div}_{i,j} < 7\ \mathrm{A}^{\circ}$. Figure 16.9(a) displays the structure of a representative sample structure (PDB identifier 1JIR), in a grey ribbon representation with the anchor residues identified in blocks of black. Consecutive anchor residues are coloured the same, resulting in longer bands of black where anchor residues lie next to each other in sequence. In fact 70 of the structures in the sample exhibit identical colourings to 1JIR.

The anchor residues are predominantly located on the outside of the protein and in loop regions. One of the beta barrels is the only region that appears to be completely devoid of colour. The beta sheets found on the section of the beta barrels that faces into the centre of the structure form the hydrophobic core that is important in attracting the specific residues that trypsin cleaves.

Protein structure is closely related to its function. The enzymatic mechanism of trypsin involves a catalytic triad of residues: the amino-acids histidine-57, aspartic acid-102 and serine-195, where the numbers after the hyphen indicate the sequence position. These three residues form a charge relay that causes the active site serine residue to become nucleophilic by modifying its electrostatic environment (Bateman et al. 2004). Trypsin also contains an ‘oxyanion hole’ formed by the backbone amide hydrogen atoms of glycine-193 and serine-195. This hole stabilises the developing negative charge on the carboxyl oxygen atom of the cleaved amides. Another important functional residue is aspartic acid-189 located in the catalytic pocket of trypsin. This residue is responsible for attracting and stabilising positively charged lysine and arginine residues (Bateman et al. 2004).

In order to determine whether these functional residues coincide with the anchor residues, Figure 16.9(b) displays the location of the functional residues, coloured in black. The functional residues are generally in the centre of the protein, in contrast to the location of the anchor residues. It can easily be seen that the functional residues and anchor residues do not overlap; that is, none of the anchor residues correspond to a functional residue.

Trypsin has a number of disulphide bonds stabilising its structure. Stroud (1974) claims that trypsin has six disulphide bonds; however, only 36 of the structures have the required number of cysteine residues, 12. According to Várallyay et al. (1997), there are three conserved disulphide bonds: C42–C58, C168–C182 and C191–C220. It was found that 80 of the 83 structures have enough cysteine residues to form at least three disulphide bonds. Figure 16.9 indicates by black lines connecting the ribbons in the structure where these three disulphide bonds are found in relation to the anchor residues and functional residues. The bonds appear to be positioned around the substrate-binding pocket; this is unsurprising given that this is the part of the structure vital to the protein's function . Only one of the bonds involves an anchor residue.

It is important to check that the positions in the structural alignment that correspond to the anchor residues are not predominantly comprised of gaps. If most of the sequences correspond to gaps in the anchor positions, then the structural conservation in these positions would be the result of a small number of structures in the sample. The median percentage of gaps in the anchor columns is 12.05 compared to a median percentage of gaps of 4.22 in the other columns in the alignment. However, because there are fewer anchor columns, the percentage of gaps in the anchor columns is much less variable, with a standard deviation of 1.89 compared to a standard deviation of 37.09 for the percentage of gaps in the other columns. Overall, the anchor columns of the alignment are not excessively gapped compared to the other columns in the alignment; however, the median number of gaps in the anchor columns is larger than that of the other alignment columns. Given that the anchor columns are not disproportionately gapped, it remains to determine which residue types are found in each anchor position and how conserved these residues are. Table 16.1 contains the percentage of each residue type in each of the anchor columns. Some of the anchor columns appear to be conserved in sequence; however, overall they do not appear to be more conserved than every other column in the alignment.

Rypniewski et al. (1994) propose several conserved residues, in both sequence and structure. Comparing the anchor residues in Table 16.1 to those proposed by Rypniewski et al. (1994) results in an overlap for some of the residues; there are 7 anchor columns that correspond to the conserved residues identified in the paper. The residues 42, 43 and 44 correspond to anchor columns 3, 4 and 5 in Table 16.1. These three residues are strongly conserved in the aligned sequences and they are identified as conserved in the paper. These three residues are found close to the active site; glycine-43 forms a hydrogen bond with the carbonyl oxygen of serine-195, one of the catalytic triad residues, and cysteine-42 forms a disulphide bond, as displayed in Figure 16.9. Anchor column 11 corresponds to residue 94, which lies in the exposed side of the loop that contains the active site residue aspartic acid-102 and is important in maintaining structure; its side chain is in contact with two residues of the catalytic triad: aspartic acid-102 and histidine-57. In the paper, residue 94 is

Table 16.1 Conservation in percentage of each amino-acid residue and gaps in the positions of the structural alignment corresponding to the anchor residues.

Amino	Anchor column																														
acid	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
L	1					63	23	24					86					1	18		1										23
Q	84									41												18				6			5		
A		23	2		1										1			23					29		6				18		1
I		18				23	54						2				2	29		1						1					18
V		45					8	18		43				2			83	14				1									34
C			86																				6								
G				86	87													20					1		18	42		44		5	
S				2		2			2					2	53				1			39	37	23	23	18					
T							2				1	2			8					83	1	1		1			72				1
D								1							1							8									
N								1	83					60		23			43			1									8
Y								42			39					28	2						18	43				41	1	24	
R									2	1		41		23		36			23	1	20	23		6	1						
P										2															43					61	
F											28															23			23		
K											1	45			1				1		33						18		29		1
W											18													18				1			
M															23					1	8										
H																1					23										
E																										1			10		
–	14	14	12	12	12	12	12	13	12	12	13	12	12	12	12	12	12	12	13	13	13	8	8	8	8	8	10	14	14	10	13

tyrosine; however, in Table 16.1 the corresponding column shows that the residue is tyrosine in only 39% of the structures. This could be due to the fact that the amino acid at residue 91 that forms a hydrogen bond with residue 94 is variable, and thus residue 94 varies to accommodate this. Conserved residues 171 and 172 are important in the specificity function of trypsin. In particular, residue 172 forms a hydrogen bond with a residue at the bottom of the specificity pocket. These residues correspond to anchor columns 23 and 24. Rypniewski et al. (1994) identify residue 172 as tyrosine, but also state that it is substituted in many sequences, explaining why it is not very conserved in Table 16.1. The final residue that is identified as conserved in the paper and is also an anchor residue is residue 225, or anchor column 30. This residue is a conserved proline residue in Table 16.1, and its role is linked to residues 171 and 172. A number of the anchor columns are found next to the residues identified as conserved by Rypniewski et al. (1994). Overall, this identifies that some of the anchor columns correspond to known conserved residues, suggesting that MUSTANG has managed to align some of the key conserved residues well.

16.3 Are the anchor residues artefacts?

The anchor residues identified by analysing the structure alignment produced by MUSTANG are intriguing. It is necessary to test that these residues are not simply an artefact produced by MUSTANG. There is no common standard for assessing the quality of a structural alignment (Liu et al. 2011); therefore, we propose the following tests.

16.3.1 Aligning another protein family

One way to identify whether the anchor residues are an artefact of MUSTANG is to align another protein family and determine whether low-range divergences are apparent. If MUSTANG is reliable, we expect the anchor residues not to be present because it is unlikely that this feature would be observed in every protein family. However, if the anchor residues are a feature of protein evolution, we would expect to see them in another protein family.

A search of Pfam (Bateman et al. 2004) produced a suitable family from a diverse range of species, short-chain dehydrogenase. A sample of 49 structures were aligned and divergence and median matrices calculated for the aligned distance matrices. Figure 16.10 displays the divergences and medians in each position of the alignment. The plot of divergences in Figure 16.10(b) does not exhibit the distinct troughs that were seen for trypsin; however, the majority of the divergences are low at less than 5 A°. The distances between the residues in the structures of this protein family are more similar than those in the trypsin family, suggesting that the short-chain dehydrogenase family of proteins is more highly conserved in structure than the trypsin protein family. Therefore, aligning short-chain dehydrogenase does not conclusively determine whether MUSTANG introduces bias. However, it does cast doubt on the significance of the anchor residues, suggesting that they are merely well-aligned regions of the trypsin protein family.

16.3.2 Aligning an artificial sample of trypsin structures

The following method generates a sample of 83 artificial proteins consisting only of C_α atoms by resampling the C_α-atom coordinates of residues from one structure. We expect the

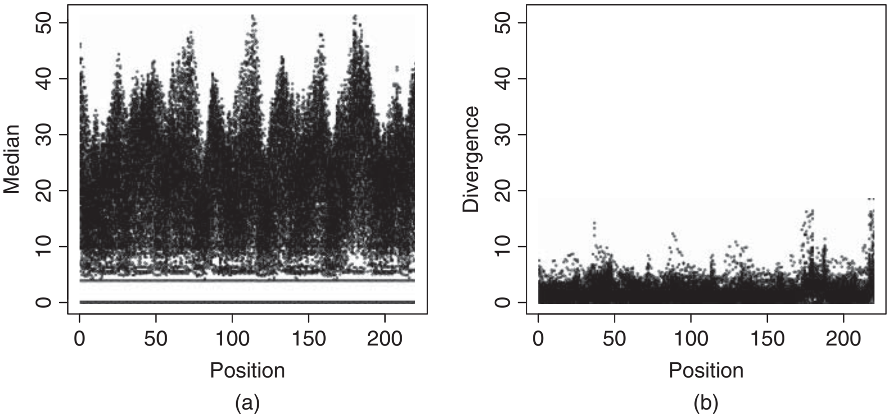


Figure 16.10 Plots of the rows of the median and divergence matrices calculated from structurally aligned distance matrices of the short-chain dehydrogenase sample. The bars appear as a result of many points plotted close together. (a) Median, $\tilde{d}_{i,j}$, of the structurally aligned distances plotted against position, i, in the alignment. (b) Divergence, $\text{div}_{i,j}$, of the structurally aligned distances plotted against position, i, in the alignment.

anchor residues to be present in the artificial sample if they truly exist, since the structures are created from one structure that exhibits the anchor residue property.

The trypsin structure 1LVY was chosen for the resampling procedure because it is relatively long and exhibits the conserved anchor residue pattern displayed in Figure 16.9. The resampled structures are generated by selecting residues to remove from 1LVY at random and then closing the resulting gaps in three-dimensional space. The gaps are closed using the following method.

When a gap is produced, the adjacent residues are linearly translated such that the Euclidean distance between their C_α atoms is equal to the standard bond length between these atoms in a typical structure.

Consider the example structure displayed in Figure 16.11. The nodes represent the C_α atoms. Let $\mathbf{x_0}$ be the vector of (x, y, z) coordinates of the C_α atom to be removed.

Once the C_α atom corresponding to $\mathbf{x_0}$ is removed, the coordinates of the adjacent C_α atoms, $\mathbf{x_{-1}}$ and $\mathbf{x_1}$, are translated using the following equation

$$\begin{aligned} \mathbf{x_{-1}}\prime &= \mathbf{x_0} + \lambda_0(\mathbf{x_{-1}} - \mathbf{x_0}), \\ \mathbf{x_1}\prime &= \mathbf{x_0} + \lambda_0(\mathbf{x_1} - \mathbf{x_0}), \end{aligned} \tag{16.1}$$

where $\mathbf{x_{-1}}\prime$ and $\mathbf{x_1}\prime$ are the new coordinates of the adjacent C_α atoms and where $\lambda_0 \in [0, 1]$. When $\lambda_0 = 0$ the new coordinates are $\mathbf{x_{-1}}\prime = \mathbf{x_0}$ and $\mathbf{x_1}\prime = \mathbf{x_0}$. At the other extreme, when $\lambda_0 = 1$, the new coordinates are $\mathbf{x_{-1}}\prime = \mathbf{x_{-1}}$ and $\mathbf{x_1}\prime = \mathbf{x_1}$. We want to choose λ_0 such that $\mathbf{x_{-1}}\prime$ and $\mathbf{x_1}\prime$ lie between these extremes, specifically at a distance of d_α apart, where d_α is defined to be the standard distance between C_α atoms:

$$d_\alpha^2 = (\mathbf{x_1}\prime - \mathbf{x_{-1}}\prime)^T(\mathbf{x_1}\prime - \mathbf{x_{-1}}\prime). \tag{16.2}$$

The average C_α atom to C_α atom bond distance in the structure 1LVY is calculated to be 3.81 A°; therefore, d_α is taken to be 3.81 A°.

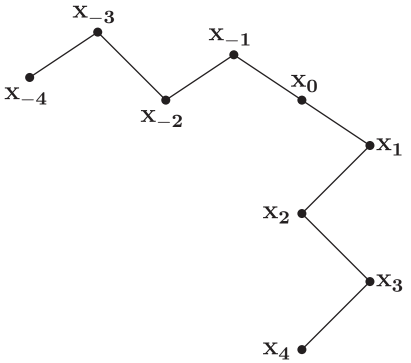


Figure 16.11 Example structure consisting only of C$_\alpha$ atoms, represented by dots; adjacent residues are connected by lines to form the backbone of the structure. The C$_\alpha$ atoms are labelled in accordance with the method for closing gaps in structure; $\mathbf{x}_0$ is a vector containing the (x, y, z)-coordinates of the residue that will be removed to form the gap, and $\mathbf{x}_1, \ldots, \mathbf{x}_4$ and $\mathbf{x}_{-1}, \ldots, \mathbf{x}_{-4}$ are the coordinates of the sequence of residues reading away from the gap on either side. See text for further explanation.

Substituting $\mathbf{x}_{-1}\prime$ and $\mathbf{x}_1\prime$ from Equation (16.1) into Equation (16.2) and rearranging in terms of λ_0 gives

$$\lambda_0 = \frac{d_\alpha}{\sqrt{(\mathbf{x}_1 - \mathbf{x}_{-1})^T(\mathbf{x}_1 - \mathbf{x}_{-1})}}.$$

Next we translate the residues adjacent to those either side of the gap. In this case, only one residue is moved in order to preserve the distance between the residues that were translated in the previous step; $\mathbf{x}_{-2}$ is translated to correct for the distance between $\mathbf{x}_{-2}$ and $\mathbf{x}'_{-1}$ as follows:

$$\mathbf{x}_{-2}\prime = \mathbf{x}_{-1}\prime + \lambda_{-1}(\mathbf{x}_{-2} - \mathbf{x}_{-1}\prime), \tag{16.3}$$

where the scale constant λ_{-1} is calculated similarly to λ_0 and is thus given by

$$\lambda_{-1} = \frac{d_\alpha}{\sqrt{(\mathbf{x}_{-2} - \mathbf{x}_{-1}\prime)^T(\mathbf{x}_{-2} - \mathbf{x}_{-1}\prime)}}.$$

Equation (16.3) is applied successively to each C$_\alpha$ atom in the structure, substituting for the appropriate coordinates at each iteration.

The number of residues removed was calibrated such that the number of gaps produced by the alignment was close to the average number of gaps in the original sample alignment. The average number of gaps per row in the original alignment was 71.99, and an alignment with 66 gaps per row was produced for the resampled structures when 28 residues are removed at random from 1LVY and aligned using MUSTANG. Removing 30 residues produced too many gaps.

To complete the simulation of artificial proteins, it is necessary to add noise to the C$_\alpha$ atom coordinates since all of the resampled structures originate from the same structure and because not all C$_\alpha$ atom to C$_\alpha$ atom bond lengths are precisely 3.81 A°.

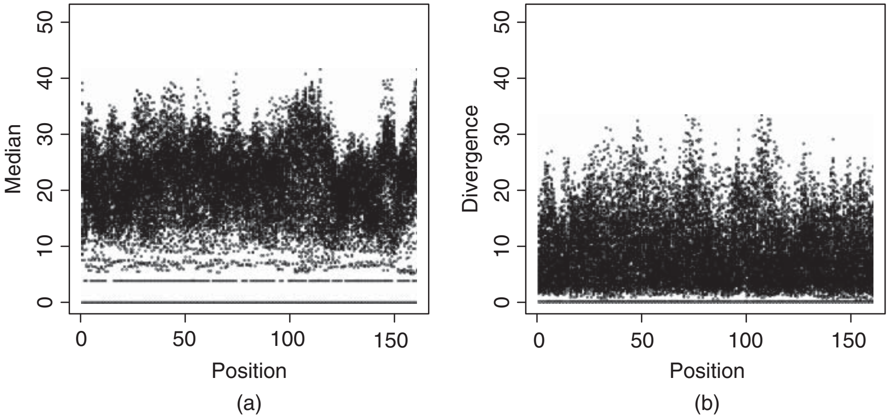


Figure 16.12 Plots of the rows of the median and divergence matrices calculated from structurally aligned distance matrices of the artificial trypsin sample. The bars appear as a result of many points plotted close together. (a) Median, $\tilde{d}_{i,j}$, of the structurally aligned distances plotted against position, i, in the alignment. (b) Divergence, $\text{div}_{i,j}$, of the structurally aligned distances plotted against position, i, in the alignment.

The previous analysis was carried out on the aligned sample of artificial structures to produce the divergence and median plots displayed in Figure 16.12. The plot of divergence against position in Figure 16.12(b) shows that the range of divergences is very high, certainly none are below 5 A°. There is no evidence to suggest the existence of anchor residues. It might be expected that the structures are very similar and would thus align well, producing low divergences; however, the range of divergences is high suggesting that the distances are less conserved than in the trypsin sample. There is certainly no evidence of the previously observed anchor residues.

When compared to Figure 16.5(a), the plot of median against position in Figure 16.12(a) for the artificial structures does not exhibit similarities with the plot for the real trypsin sample. This difference in median distances suggests that the artificial structures have a different structure to the trypsin sample structures. This is not unusual since the artificial structures are all variations of one structure, 1LVY. However, it is necessary to understand the effect that the gap-closing method has on the shape of a structure; this is explored in Section 16.4.

Similarly to Figure 16.8(b), the divergence matrix can be displayed as a heatmap for the artificial sample, given in Figure 16.13. In this case, it is the high divergences that bring up the scale; as a result divergences greater than 10 A° have been coloured black. There is no longer the pattern of horizontal lines that could be observed in Figure 16.8(b), confirming that there are no anchor residues. In fact, there are very few areas on the off diagonal that have low divergences at all.

The number of gaps removed was also varied for each resampled structure in the sample; however, the same results were obtained concerning the low-range divergences or anchor residues.

There are a number of ways in which this methodology for producing artificial structures could be improved. In order to reflect true evolutionary processes, insertions and

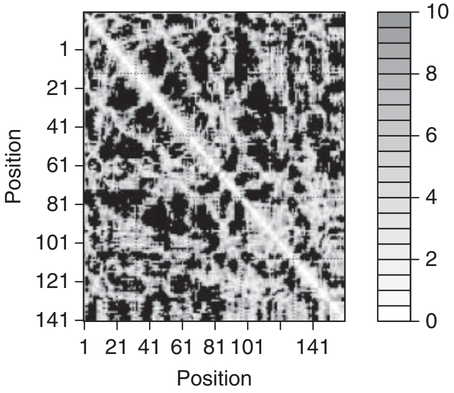


Figure 16.13 Divergence matrix heatmap for the artificial trypsin sample, recalculated for all of the divergences that are less than 10 A°. Larger divergences are blacked out. The divergence between the residue–residue distances is plotted in greyscale; small distances are white and large distances are dark grey.

substitutions could be incorporated as well as deletions. The method could also be extended to include all of the atoms in the starting structure, not just the C_α atoms.

Therefore, this method provides evidence against MUSTANG; we would expect anchor residues to be apparent in 1LVY if they truly exist. However, they are not apparent in artificial structures, suggesting that the phenomenon is an artefact of MUSTANG.

16.3.3 Aligning C_α atoms of the real trypsin sample

Since the method in the previous section uses only the C_α atom coordinates, it is necessary to compare the structural alignment of the trypsin sample with the alignment produced when only the C_α atoms of their residues are structurally aligned. MUSTANG appears to use only the C_α atoms of structures when producing an alignment. Therefore, we expect the full-atom trypsin alignment and the C_α only trypsin alignment to be similar.

In this case, the plots of divergence and median against position displayed in Figure 16.14 are produced and compared to the full-atom structural alignment of the trypsin sample in Figure 16.5. The distinct troughs in the divergences in Figure 16.5(b) are not apparent when only the C_α atoms of trypsin are aligned; however, there are a lower range of divergences compared to the artificial structures. There appears to be some correspondence between the peaks of the median distances in Figure 16.5 and Figure 16.14, suggesting the overall shape of the structures is not too different, and thus the two alignments are reasonably similar. However, it also suggests that using only C_α atoms is not representative of the full sample.

To understand more about how the full-atom trypsin structural alignment and the corresponding C_α-atom-only structural alignment differ, their gaps are analysed. In this case, a gap is defined to be a consecutive run of insertions where the length of the gap is the number of insertions. The median number of gaps in the C_α-atom alignment is much larger at 41.00, compared to a median number of gaps of 24.00 in the full-atom alignment. The number of gaps is also much more variable in the C_α-atom alignment with a standard deviation of

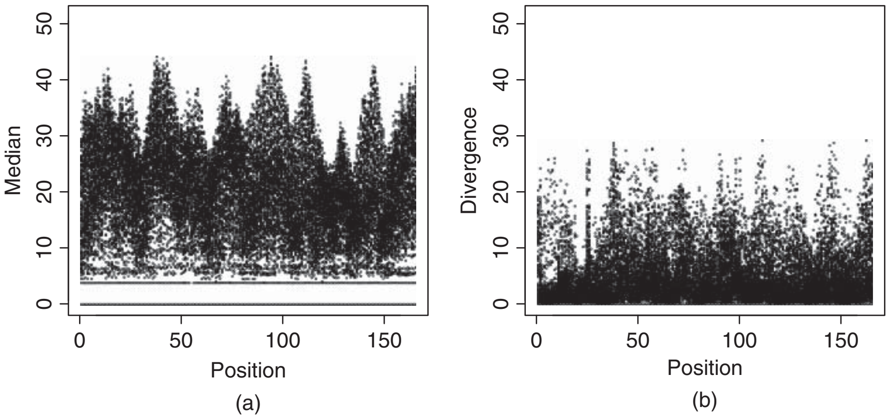


Figure 16.14 Plots of the rows of the median and divergence matrices calculated from structurally aligned distance matrices of the trypsin sample with only C_α atoms. The bars appear as a result of many points plotted close together. (a) Median, $\tilde{d}_{i,j}$, of the structurally aligned distances plotted against position, i, in the alignment. (b) Divergence, $\mathrm{div}_{i,j}$, of the structurally aligned distances plotted against position, i, in the alignment.

13.66 compared to a standard deviation of 1.92 in the full-atom alignment. However, are these gaps shorter than those in the original alignment?

The median length of the gaps in the two alignments is the same at 2.00; however, the range of values is very different. The largest gap in the full-atom alignment is 21.00, compared to an incredibly long gap of 119.00 in the C_α-atom alignment. Unsurprisingly, the standard deviation for the length of the gaps in the C_α-atom alignment is larger at 7.80, compared to a standard deviation of 2.703 for the length of the gaps in the full-atom alignment. Therefore, not only does the C_α-atom alignment appear to have more gaps for most sequences, some of the gaps are also significantly longer compared to the original alignment.

Clearly, the full-atom alignment and the C_α atom alignment are quite different; therefore, the methods for testing bias may not be entirely representative of the full-atom case. This is an interesting result since MUSTANG aligns structures by using only the information from the C_α atoms and the distances between them; therefore, the alignments should be similar.

16.3.4 Aligning the real trypsin sample with anchor residues removed

The following further test was conducted. The anchor residues were removed from the structures in the sample and the resulting structures aligned; if the alignment results in more anchor residues, then MUSTANG is unreliable. The divergence and median were again plotted against position and are displayed in Figure 16.15.

The peaks of the median distance plots in Figures 16.5(a) and 16.15 are very similar, suggesting that the alignments are similar. However, there is no longer evidence of low-range divergences or anchor residues as the distinct troughs in the divergences in Figure 16.5(b) are no longer apparent; the divergence between the distances appears to be higher overall.

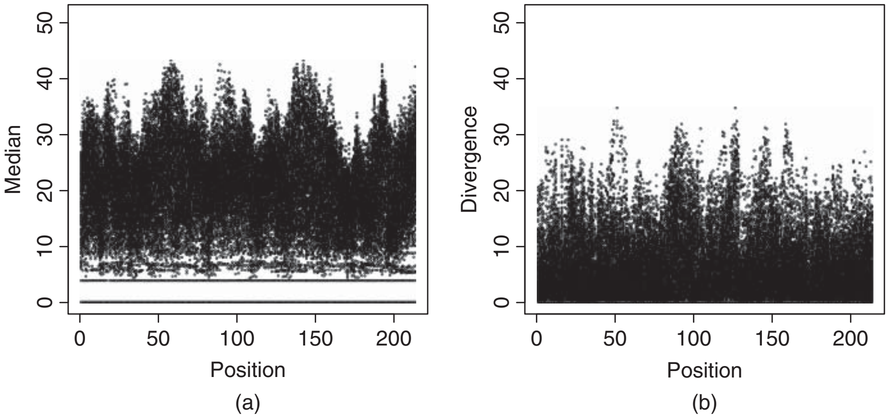


Figure 16.15 Plots of the rows of the median and divergence matrices calculated from structurally aligned distance matrices of the trypsin sample with the anchor residues removed. The bars appear as a result of many points plotted close together. (a) Median, $\tilde{d}_{i,j}$, of the structurally aligned distances plotted against position, i, in the alignment. (b) Divergence, $\text{div}_{i,j}$, of the structurally aligned distances plotted against position, i, in the alignment.

Therefore, removing the anchor residues produces results in favour of MUSTANG. This suggests that more tests are necessary in order to definitively determine whether the anchor residues are artefacts of MUSTANG.

16.4 Effect of gap-closing method on structure shape

In order to explore the effect of the gap-closing method in Section 16.3.2 on the shape of a structure, we applied it to a selection of shapes typically found in protein secondary structures. The shapes investigated include a zigzag and an idealised helix.

16.4.1 Zig-zag

The structure of trypsin has many beta sheets, where the C_α-atoms of residues lie alternately above and below the plane of the beta sheet, not dissimilar to a zigzag. A zigzag structure was generated such that the residues were d_α apart, and such that each set of three consecutive residues formed an equilateral triangle with sides of length d_α. Figure 16.16(b) shows how the zigzag structure is affected when a gap is closed. The same pattern is observed wherever the gap is placed. However, Figure 16.16(c) shows the effect on the structure when a gap of size 16 is closed. Clearly, closing large gaps disrupts the structure around the gap significantly.

16.4.2 Idealised helix

The structure of trypsin has two small helices; therefore, it is of interest to analyse how the structure of a helix changes when residues are removed and the gap closed. An idealised helix with 50 residues was generated such that the residues are d_α apart and the helix has 3.6

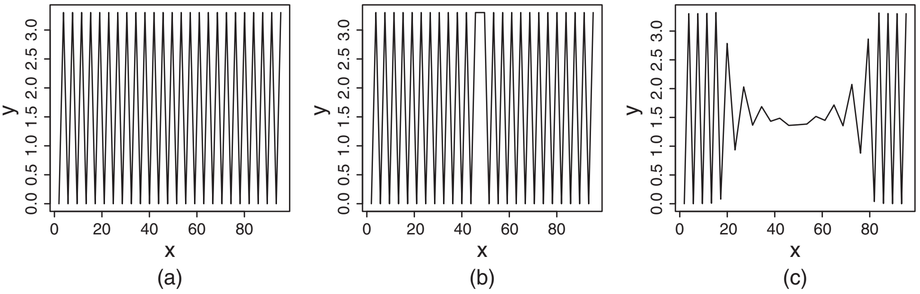


Figure 16.16 Plots displaying the effect of the gap-closing method on a zigzag structure. (a) Zigzag structure before a gap is closed. (b) Zigzag structure after closing a gap of size one that is introduced in the middle of the structure. (c) Zigzag structure after closing a gap of size 16 that is introduced in the middle of the structure.

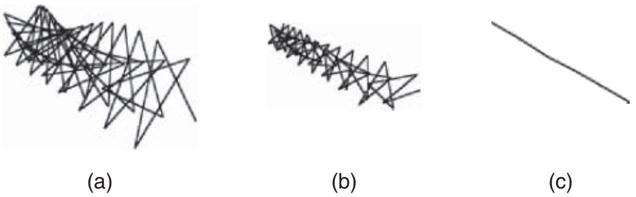


Figure 16.17 Plots displaying the effect of the gap-closing method on a helix structure. (a) Helix structure before a gap is closed. (b) Helix structure after closing a gap of size one that is introduced in the middle of the structure. (c) Helix structure after closing a gap of size 16 that is introduced in the middle of the structure.

residues per turn. Figure 16.17 displays the effect of the gap-closing method on the helical structure. Figure 16.17(b) displays the helix structure after one residue is removed. It is difficult to spot, but there is an irregular kink at the end of the helix. This kink occurs regardless of the position of the residue being removed. However, when more residues are removed, the gap is far less subtle. Figure 16.17(c) displays the result of removing 16 residues and closing the gap; the helical structure is barely recognisable. In fact, the helix structure is almost completely destroyed after only five residues are removed.

16.5 Alternative to multiple structure alignment

One way to be sure that MUSTANG introduces no structural bias is to conduct the analysis using a multiple-sequence alignment of the structures where only sequence and no structural information is used. Distance matrices can be obtained based on the sequence alignment and divergence and median matrices calculated as before. The sequences are aligned using Clustal W (Thompson et al. 1994), and the divergences and medians plotted against position in Figure 16.18.

Compared to Figure 16.5(b), the divergences in Figure 16.18(b) are similar in range; however, the divergences in the anchor positions are not small or distinct. The median plots

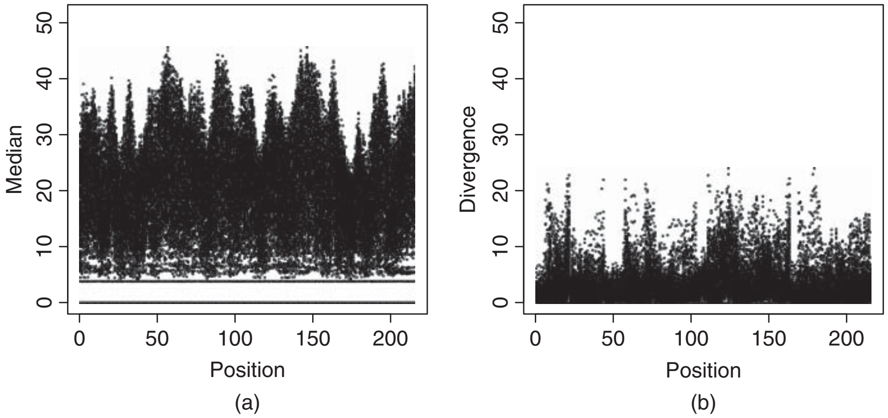


Figure 16.18 Plots of the rows of the median and divergence matrices calculated from the aligned distance matrices of the Clustal W multiple-sequence alignment of the trypsin sample. The bars appear as a result of many points plotted close together. (a) Median, $\tilde{d}_{i,j}$, of the aligned distances plotted against position, i, in the alignment. (b) Divergence, $\text{div}_{i,j}$, of the aligned distances plotted against position, i, in the alignment.

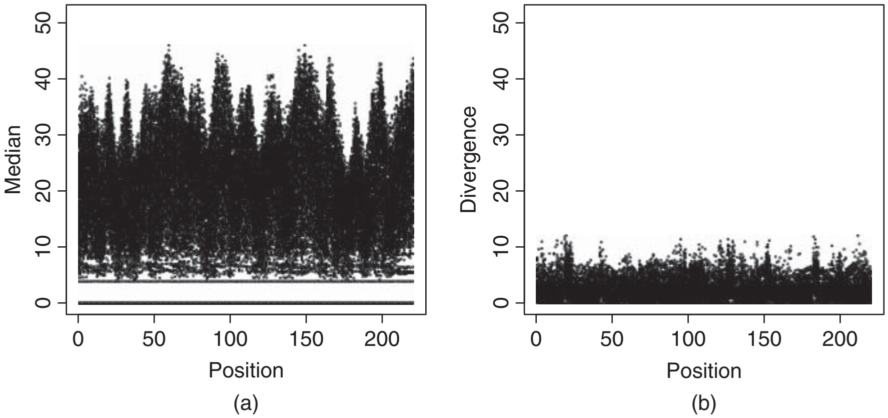


Figure 16.19 Plots of the rows of the median and divergence matrices calculated from the aligned distance matrices of the MUSCLE multiple-sequence alignment of the trypsin sample. The bars appear as a result of many points plotted close together. (a) $\tilde{d}_{i,j}$, of the aligned distances plotted against position, i, in the alignment. (b) Divergence, $\text{div}_{i,j}$ of the aligned distances plotted against position, i, in the alignment.

in Figures 16.5(a) and 16.18(a) have a very similar pattern of peaks, further suggesting that the structure alignment is similar to the sequence alignment.

For comparison, a second multiple-sequence alignment algorithm is used, MUSCLE (Edgar 2004). The same plots for this alignment are displayed in Figure 16.19. Compared to Figure 16.18(b), the divergences in Figure 16.19(b) are much smaller overall and there are fewer large divergences. Most of the positions contain divergences small enough to be considered as the anchor residues that were identified previously; however, the divergences are not as low as the troughs in Figure 16.5(b). This suggests that the MUSCLE sequence alignment results in more conserved aligned distances compared to the MUSTANG structure alignment, and even the Clustal sequence alignment. However, despite producing a better structural alignment than MUSTANG overall, the anchor positions do not appear to be aligned as well. Similarly to Figure 16.18(a), the median distances exhibit almost identical peak patterns to Figure 16.5(a).

16.6 Discussion

We have presented an investigation into the possibility that the trypsin protein family contains ‘anchor’ residues. That is, residues where the distance between these residues and every other in the structure is highly conserved across all of the structures in the protein family, compared to the other distances in the structure. These anchor residues were identified from the aligned distance matrices from the structural alignment produced by MUSTANG. We conducted several tests to determine the validity and origin of these anchor residues.

Investigation into the origin of the putative anchor residues did not result in a definitive explanation; while some of the anchor residues appeared to correspond to important conserved residues identified by Rypniewski et al. (1994), the evidence was not overwhelming. The anchor residues were not more conserved in sequence compared to the rest of the columns in the structural alignment.

The artefact testing method proposed in Section 16.3.2 proved inconclusive; we would expect anchor residues to be apparent in 1LVY if they truly exist. However, they were not apparent in the artificial structures suggesting that the phenomenon is an artefact of MUSTANG. When the artefact testing method was investigated in Section 16.4, it became clear that the gap-closing method distorts the structures significantly and as a result the distances are also distorted. This method used only the information contained in the C_α atoms of the structures. This was considered reasonable because MUSTANG appears to use this information only. Despite this, aligning only the C_α atoms of the trypsin sample produced a different alignment compared to the trypsin sample; the alignment has more insertions, as well as longer consecutive runs of insertions. This suggests that incorporating only the information contained in the C_α atoms of the structures produces a less desirable alignment, and therefore, MUSTANG either incorporates additional information or is unreliable. Consequently, we do not have much confidence in the artefact testing method to accurately determine whether the anchor residues are an artefact of MUSTANG. A simple test of removing the anchor residues in order to test whether MUSTANG would artefactually introduce more residues concluded in favour of MUSTANG, as no new anchor residues were produced. The median distance matrix also provides evidence in favour of the MUSTANG alignment, owing to the fact that the structure produced by multidimensional scaling of the median distance matrix resulted in a homogeneous trypsin structure.

When another protein family was aligned, we expected the anchor residues not to be apparent if MUSTANG is not introducing bias because it is unlikely that this feature would be observed in every protein family. However, the anchor residues may be a feature of protein evolution rather than an artefact. The divergences in each position were all small, suggesting that anchor residues merely identified areas of the alignment where trypsin aligned well. Since this was not a large area, it appeared to be an interesting result.

While multiple-sequence alignments do not introduce bias, they also do not produce an alignment based on how the structural components are aligned. A reliable structural alignment would be preferred to an alignment based purely on sequence because the protein structure evolves more slowly than sequence.

The Clustal-W sequence alignment results in a similar range of divergences compared to the MUSTANG alignment. However, the MUSCLE sequence alignment is significantly different with a much lower range of divergences overall. We expect differences between the structure and sequence alignments because the structure alignment completely ignores the amino-acid sequence while the sequence alignments only use the amino-acid sequences. MUSTANG ignores the amino-acid sequence in order to align more distantly related proteins; similarly Clustal-W weight sequences based on their similarity. This focus on the evolution of the structures may explain why Clustal-W and MUSCLE produce different alignments.

Out of the tests that were conclusive, many are in favour of MUSTANG. However, some tests identify inconsistencies that lead us to believe that MUSTANG may be unreliable. The most convincing result against the existence of anchor residues arose from aligning another protein family; the distances in the short-chain dehydrogenase protein family have smaller divergences than the anchor residues in every position. This strongly suggests that the anchor residues merely indicate well-aligned regions of structure in the trypsin family. Combined with the result that the anchor residues do not appear to be strongly conserved in sequence or correspond to important functional residues, we conclude that MUSTANG may be introducing bias, but it is also likely that the anchor residues are artefacts of the trypsin family. To support this conclusion, a larger range of protein families from diverse organisms would need to be aligned, both in sequence and structure. There is also scope to subject MUSTANG to further testing to determine its reliability.

References

Bateman A, Coin L, Durbin R, Finn R, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer E *et al.* 2004 The pfam protein families database. *Nucleic Acids Research* **32**(Suppl 1), D138–D141.

Berbalk C, Schwaiger CS and Lackner P 2009 Accuracy analysis of multiple structure alignments. *Protein Science* **18**(10), 2027–2035.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov IN and Bourne PE 2000 The protein data bank. *Nucleic Acids Research* **28**(1), 235–242.

Branden C, Tooze J *et al.* 1991 *Introduction to Protein Structure* vol. 2. Garland, New York.

Edgar RC 2004 Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**(5), 1792–1797.

Henikoff S and Henikoff JG 1994 Position-based sequence weights. *Journal of Molecular Biology* **243**(4), 574–578.

Holm L and Sander C 1993 Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology* **233**(1), 123–138.

Jmol an open-source java viewer for chemical structures in 3D. http://www.jmol.org/

Konagurthu AS, Whisstock JC, Stuckey PJ and Lesk AM 2006 Mustang: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics* **64**(3), 559–574.

Liu W, Srivastava A and Zhang J 2011 A mathematical framework for protein structure comparison. *PLoS Computational Biology* **7**(2), e1001075.

Mardia KV, Kent JT and Bibby JM 1979 Multivariate analysis.

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R and Sander C 2011 Protein 3d structure computed from evolutionary sequence variation. *PLoS One* **6**(12), e28766.

R Core Team 2013 *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria.

Rypniewski W, Perrakis A, Vorgias C and Wilson K 1994 Evolutionary divergence and conservation of trypsin. *Protein Engineering* **7**(1), 57–64.

Saitou N and Nei M 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**(4), 406–425.

Stroud RM 1974 A family of protein-cutting proteins. *Scientific American* **231**(1), 74.

Thompson JD, Higgins DG and Gibson TJ 1994 Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**(22), 4673–4680.

Várallyay É, Lengyel Z, Gráf L and Szilágyi L 1997 The role of disulfide bond c191-c220 in trypsin and chymotrypsin. *Biochemical and Biophysical Research Communications* **230**(3), 592–596.

Zhang Y and Skolnick J 2005 Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Research* **33**(7), 2302–2309.

17

Individualised divergences

Clive E. Bowman
Mathematical Institute, University of Oxford, Oxford, UK

17.1 The past: genealogy of divergences and the man of Anekāntavāda

Fundamental to the analysis of shape is the ability to compare profiles of variables across individual instances in a geometry. Insight into the relatedness of such multidimensional things is by contrasting them in this space. A distance measure over such ‘signatures’ in the context of any group distinction of instances is needed for this operation – and, of course, in practice any individual measurement of reality has error. For a statistician, knowledge is based upon the evidence that stochastic data yields in a designed test of a hypothesis (i.e. an experiment). Evidence is information in a context. Shannon founded the use of information in the theory of communication in 1948 (Shannon 1948), with Kullback and Leibler building upon this, formalising measures of evidence by introducing the concept of information divergences in 1951 (Kullback and Leibler 1951). Knowledge about the size and shape of profiles in the world is driven by the evidence found for hypotheses or claims and assertions about them. Divergences explicitly measure any distinctions in a space of stochastics – yet outside of statistical density comparisons such measures of distance have been little used in shape analysis. Sibson (1969) extended the theoretical field around the 1970s and facilitated its brief impact among the early interest of numerical taxonomy by biologists (Jardine and Sibson 1971). Thereafter, it languished despite its august family tree.

Professor Kantilal Mardia, MSc, PhD, DSc, *‘...a statistician specializing in directional statistics, multivariate analysis, geostatistics, statistical bioinformatics and statistical shape analysis’* (Wikipedia 30th December 2013) has fostered the rediscovery and

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

popularisation of these probabilistic distance measures for the last 10 years. Kanti, as he is known to his friends, has always been far-sighted about under-considered professional advances and the renaissance of long forgotten results – having a keen historical eye and genealogical interest. A sage ‘look-out’, he intuitively identified something when a strange non-academic from an industrial ‘back-water’ (*myself*) offered a poorly written abstract for a talk at LASR2005 (Delrieu and Bowman 2005). This talk outlined the Euclidean geometric decomposition (using SVD) of a smooth universal metric of evidence (*individualised divergences*) to understand contrasts in observed very high dimensional profiles (i.e. patterns and features in size and shape simultaneously over multiple data types). Negentropic at heart, individualised divergences encapsulate the evidence that each point *itself*, among the measurements made, engenders for a question (i.e. a comparison) that a researcher may ask. From that generous opportunity much has flowed over the subsequent time in and around comparison of the shape of single nucleotide polymorphism profiles in precision medicine (Alfirevic et al. 2009; Bowman 2009; Bowman and Delrieu 2009a; Bowman et al. 2006; Charalambous et al. 2008; Delrieu and Bowman 2006b, 2007). Individualised divergences have advanced scientific understanding in the immunogenetics of drug-induced skin blistering disorders (Bowman and Delrieu 2009b, 2009c) and in dissecting the biochemistry of platelet function (Bowman and Jones 2010). A detailed semi-worked disease example is given in Delrieu and Bowman (2006a) and an applied genetic example is in Pirmohamed et al. (2007). Many other practical examples of these evidence methods needed to generate knowledge from medical profile data such as Zhang et al. (2008) are outlined in Bowman (2013). They are now part of grand initiatives such as my Royal Society Industrial Fellowship IF110047 (2012–2016), and a European Union Seventh Framework Programme for Research FP7 grant. The field has been reawakened courtesy of the consideration of humble-born Kanti, a man of Anekāntavāda who exemplifies the ‘Middle Path’ (Mardia 2007): Right speech, right action and right livelihood constituting ethical conduct; Right effort, right mindfulness and right concentration as mental disciplines; and, Right understanding and right thought constituting wisdom. Without Kanti, knowledge and application of divergences could have remained forgotten for many more years.

17.2 The present: divergences and profile shape

Outside of their recent application to contrasting genetic profiles and in Jardine and Sibson (1971), divergences are not yet widely used in shape work (Dryden and Mardia 1998). This is despite shape analyses being comparative (Procrustean) at their heart, that is, based upon a contrast. Accordingly, the basis of the individualisation of divergences for shape analysis needs wider explanation. The subsequent sections give this in detail to complement the practical examples referenced in the previous section. First the theory is explained, then a likelihood formulation is outlined employing parameter estimation, expectation and individualisation. Finally, the whole algorithm is assembled and a brief justification of why it works given together with a new example.

17.2.1 Notation

The starting point is a probability model, $g(x; \theta)$ for a random observation X under a model with parameter θ. For simplicity, we focus on continuous models (so g is a probability density function with respect to dx), but the same conclusions hold in the discrete case with integrals replaced by sums. When we wish to emphasise vector-valued observations with p components, we use bold-face and write $\boldsymbol{x} = (x_1, \ldots, x_p)^T$.

The emphasis in this chapter will be on comparing two models, with parameters θ_1 and θ_2, say. When investigating medical profiles, a common setting is a case-control study to compare two groups of individuals. An example using the profile of high-dimensional SNP divergences is given in Delrieu and Bowman (2005) where a simpler decomposition is required to understand between and within profile sample structure.

Data from these two models will take the form of an $n \times p$ data matrix $\boldsymbol{X}$, where the first n_1 rows come from the first model and the last n_2 rows come from the second model, $n_1 + n_2 = n$. A typical data value is written $x_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, p$.

Throughout the chapter, $\log$ means the base e logarithm. In information theory, base 2 is often used, with information measured in bits. Of course, a base e logarithm can be converted into a base 2 logarithm by multiplying by $\log_2 e$.

17.2.2 Known parameters

In the simplest setting, we imagine a single univariate observation x, coming from a probability model, $g(x; \theta)$. Define the *marginal* self-information Shannon (1948),

$$-\log g(x; \theta).$$

This is a measure of the information content associated with the outcome of a random variable from this distribution. The unit of this information, after multiplying by $\log_2 e$, is the 'bit'. It has also been called surprisal Tribus (1961), as it represents the 'surprise' of seeing the outcome (e.g. a highly probable outcome is not surprising). It is *marginal* in the sense that the dependence of x on any other variables we might measure is not being taken into account here.

The *marginal* differential entropy (or expected surprisal) is

$$-\int g(x; \theta) \log g(x; \theta) dx.$$

This information entropy is a number measuring the uncertainty associated with a random variable. It is a continuous analogue of the discrete Shannon entropy (Shannon 1948). It is a measure of the average information content a recipient is missing when they do not know the value of the random variable. Some authors define the negative of this to be 'Risk'. However, this usage can be confused with the colloquial use of the word 'risk' and is not included in this outline.

These entropies can be used to define divergences. Divergences (Kullback 1959; Kullback and Leibler 1951) are natural measures for analysis which flow from considering any scientific question as a contrast (Mead 1990). All hypotheses posed in science are relative; that is, they are a contrast from a basis (in the Popperian paradigm – from a 'null' hypothesis, null being used colloquially here). Contrasts are compact ways of comparing estimated summary measures by adding or subtracting them and interpreting the answer. So 'Is the cost of a car more than the cost of a bicycle?' is a contrast summarised by the algebra:

$$\text{Cost-of-Car} - \text{Cost-of-Bicycle} > 0.$$

This may be estimated from a cost of a single car and a single bicycle (one instance at a time) or perhaps by some summary measure of costs or ‘typical’ values over many cars and bicycles; the principle is the same. With thought, any experiment or scientific assertion can be posed as a contrast, even if one uses logarithms to move from a multiplicative or ratio space of interest to a linear additive one.

Taking any two populations (with known parameters θ_1 and θ_2), the log probability ratio between the populations is

$$\log\left\{\frac{g(x;\theta_1)}{g(x;\theta_2)}\right\}.$$

It is a measure of the information in x for discriminating between θ_1 and θ_2.

The *marginal* relative entropy (or Kullback–Leibler divergence) between the two populations is a measure of the *directed* (i.e. oriented or asymmetric) difference between the two probability densities. It is defined by the expected value (under population 1) of the log probability ratio,

$$D_{\mathrm{KL}}(\theta_1;\theta_2) = \int g(x;\theta_1)\log\left\{\frac{g(x;\theta_1)}{g(x;\theta_2)}\right\}dx.$$

It is the ‘loss’ on average when the $g(x;\theta_2)$ density is used to approximate the $g(x;\theta_1)$ density. Note that the expectation is over $g(x;\theta_1)$, not $g(x;\theta_2)$. It can be viewed as a measure of the mean information in discriminating between θ_1 and θ_2 using X. It is sometimes confusingly called the (*marginal*) discrimination information function (see Dadpay et al. 2007). It is a contrast (expand the log term as a difference).

The Kullback–Leibler divergence has some advantageous properties as a summary measure.

- It is always non-negative.
- It equals 0 *only* if both distributions are identical.
- The larger the divergence is in value, the further apart are the two densities; small values indicate closeness.
- It is not symmetric; swapping θ_1 and θ_2 generally leads to a different value. Hence, it is *not* necessarily a metric as it stands.
- It is additive for independent random variables; i.e. if X is a bivariate random vector with independent components, then the overall Kullback–Leibler divergence is just the sum of divergences for the two components.
- It is invariant against transformations of the *sample* space of X. That is, if instead of the random variable X, one considers $Y = h(X)$, where h is an invertible function, then the Kullback–Leibler divergence remains the same. In this sense, it is a *geometric* quantity *independent of the choice of the co-ordinate system.*
- It belongs to the class of *f*-divergences (see Table 17.1).

If θ is an m-dimensional vector, and if θ_2 is close to θ_1, $\theta_2 = \theta_1 + \delta\theta$, the Kullback–Leibler divergence can be simplified using a Taylor’s series approximation to

$$D_{\mathrm{KL}}(\theta_1;\theta_2) \approx \frac{1}{2}\delta\theta^T I_{\theta_1}^{\mathrm{Fisher}}\delta\theta.$$

Table 17.1 Various 'f-divergences' (Ali–Silvey distances) between two discrete probability measures $\mu(z)$ and $\pi(z)$ in Euclidean space; see Nyguyen et al. (2005).

Name of distance	Continuous convex function $f(u) \ : \ [0,\infty) \to \Re \bigcup \{+\infty\}$	f-divergence $I_f(\mu,\pi)$
Kullback–Leibler	$u\log(u)$	$\sum_z \mu(z)\log(\frac{\mu(z)}{\pi(z)})$
Variational distance $*$	$\lvert u-1\rvert$	$\sum_z \lvert\mu(z)-\pi(z)\rvert$
Hellinger distance †	$\frac{1}{2}(\sqrt{u}-1)^2$	$\frac{1}{2}\sum_{z\in Z}(\sqrt{\mu(z)}-\sqrt{\pi(z)})^2$

For continuous densities see Barnett et al. (2002). This tableau is focused on distances, but detection and discrimination are 'two faces of the same coin' as are nearness and relatedness; see Tobler (1970). One is a surrogate for the other, just as SVD ordination of objects (aka PCA) can be seen as geometrical transformation of data or as a loss minimisation of mutual inter-data-point distances over a kernel (aka PCOORD); they are duals. Distances detect shapes in ordinations and vice-versa. Note that these f-divergences (widely used in signal processing, that is, stochastic density and distribution detection) can be mapped to an equivalence class of particular convex loss functions in machine learning classification decisions (e.g. support vector machines, boosting and logistic regression). $*$ Variational distance $\equiv$ '[0,1] loss' and 'hinge losses'. † Hellinger distance $\equiv$ 'exponential loss' = [1-Bhattacharyya distance] (Taneja 2005).

Here $I_{\theta_1}^{\text{Fisher}}$ is the usual $m \times m$ Fisher information matrix,

$$I_{\theta_1}^{\text{Fisher}} = -E\left\{\frac{d^2}{d\theta\, d\theta^T}\log g(x;\theta)\right\},$$

where the expectation is taken under $g(x;\theta_1)$.

As a local version of information about the parameters, the Kullback–Leibler divergence is *not* invariant to a change in the *co-ordinate system of the parameters*. That is, if θ is, say, changed to $\phi(\theta)$, then the Jacobian matrix $J_\theta(\phi)$ of partial derivatives comes into play as usual,

$$I_{\phi}^{\text{Fisher}}(\phi) = J_\theta(\phi)^T I_{\theta}^{\text{Fisher}}(\theta(\phi))J_\theta(\phi).$$

The *marginal* Jeffrey's symmetric divergence (Jeffreys 1946) is the average of the two Kullback–Leibler divergences,

$$D_J(\theta_1,\theta_2) = \frac{1}{2}\left\{D_{\text{KL}}(\theta_1;\theta_2) + D_{\text{KL}}(\theta_2;\theta_1)\right\}.$$

This numerical value like its particular variant the Jensen-Shannon divergence (Endres and Schindelin 2003) is an *undirected* (i.e. symmetric) distance and measures the difference between the two probability densities. It is the average 'loss' when each of the densities $g(x;\theta_1)$ and $g(x;\theta_2)$ is used to approximate each other. It is a metric distance if square rooted. It is a type of information radius (Sibson 1969). Note the expectations over the different densities in the two terms mentioned earlier. It is a contrast (expand the log term as a difference). It is always greater than or equal to 0 and equals 0 if and only if the two densities are identical.

For well-behaved distributions, all of these information measures are additive and open to manipulation by standard linear algebra. When simple closed forms are available, these should be used (especially in the following marginal likelihood formulations) in preference to numerical derivations. For instance, consider two full d-dimensional multivariate normals, $N_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Then, with $\theta_1 = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\theta_2 = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, the (asymmetric) Kullback–Leibler divergence is

$$D_{\mathrm{KL}}(\theta_1; \theta_2) = \frac{1}{2}\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \mathrm{tr}(\boldsymbol{\Sigma_2}^{-1}\boldsymbol{\Sigma}_1) - d - \mathrm{logdet}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)\}. \tag{17.1}$$

When $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$, say, then the asymmetry collapses and D_{KL} reduces to

$$D_{\mathrm{KL}}(\theta_1; \theta_2) = \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{17.2}$$

which is known as half the squared Mahalanobis distance.

17.2.3 The likelihood formulation

The previous section covers the distributional theory of information, but experimental science is about dealing with *samples*, presented as an $n \times 1$ data vector $\boldsymbol{x}$ when there is just $p = 1$ variable. If it is possible to move easily from theoretical distributions to observed data, then a linear measure of observed information (cf. evidence) is available for experimenters for any form of the statistical distribution.

To illustrate these ideas, suppose that the two populations represent a treatment group (θ_1) and a control or reference group (θ_2), respectively. The first step is to estimate the parameters. A variety of methods can be used, including standard maximum likelihood techniques, least squares (Dobson 1983) or Bayes' estimates with simple conjugate priors (Congdon 2006). In some cases, explicit forms are available; in other cases, an iterative solution of complex non-linear equations is needed. (Ironically, maximum likelihood estimation itself is equivalent to a minimisation of a Kullback–Leibler divergence between the actual data and an empirical data distribution of Dirac delta functions!)

For this chapter, we limit attention to maximum likelihood estimates, denoted $\hat{\theta}_1$ and $\hat{\theta}_2$, with estimation carried out separately for each group. For each observation $i = 1, \ldots, n$, Delrieu and Bowman (2005) suggested considering the individualised log likelihood ratio,

$$\ell(i) = \log\left\{L(x_i; \hat{\theta}_1)/L(x_i; \hat{\theta}_2)\right\}, \quad i = 1, \ldots, n.$$

Delrieu and Bowman (2005) regard this value as an observed divergence or individualised likelihood ratio. For instance, if the two populations were $N(\mu_1, \sigma_2^2)$ and $N(\mu_2, \sigma_2^2)$, then twice this log-likelihood ratio for x_i would be

$$\log(\hat{\sigma}_2^2) + \frac{(x_i - \hat{\mu}_2)^2}{\hat{\sigma}_2^2} - \log(\hat{\sigma}_1^2) - \frac{(x_i - \hat{\mu}_1)^2}{\hat{\sigma}_2^2}.$$

Taking the expectation over the first population yields the Kullback–Leibler divergence

$$D_{\mathrm{KL}}(\theta_1; \theta_2) = \frac{1}{2}\left\{\log(\sigma_2^2/\sigma_1^2) + \frac{\sigma_1^2}{\sigma_2^2} - 1 + \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}\right\},$$

the one-dimensional version of (17.1), where now the hat notation for estimates has been dropped for simplicity. The one-dimensional symmetric *J*-divergence is

$$D_J(\theta_1; \theta_2) = \frac{1}{2}\{D_{\mathrm{KL}}(\theta_1; \theta_2) + D_{\mathrm{KL}}(\theta_1; \theta_2)\} = \frac{\sigma_1^4 + \sigma_2^4 + (\sigma_2^2 + \sigma_2^2)(\mu_1 - \mu_2)^2}{2\sigma_1^2\sigma_2^2} - 1,$$

see Bowman et al. (2006).

Since the first n_1 observations are assumed to come from the first population and the last n_2 from the second population, we see that

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\log\{L(\hat{\theta}_1;x_i)/L(\hat{\theta}_2;x_i)\}+\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\log\{L(\hat{\theta}_2;x_i)/L(\hat{\theta}_1;x_i)\}$$

is an estimate of the overall symmetrised divergence. In particular when $n_1 = n_2$, this is an entropy approximation to the logarithm of an evidence ratio for testing if two samples have been drawn from the same underlying distribution (Endres and Schindelin 2003).

For Bernoulli distributed variables (such as genotypes treated each distinct genotype at a time), the observed log-likelihood ratio resolves to the log relative frequency of that genotype occurrence between the two groups, and the expected to a probability weighted version. These surrogate measures, originally called '*lbf*s' (Delrieu and Bowman 2006a), can be inserted in place of each data point as a transformation from data space to evidence space for each observation. For more detail together with the concepts of case-ness and this-ness, see Bowman (2009).

The formulation here does not impose any penalty for complexity if θ_1 and θ_2 have different dimensions, such as in Akaike (1974). Simple closed forms are available for likelihood-based divergences if the models belong to the exponential family of distributions (Delrieu and Bowman 2005, 2007; Bowman et al. 2006) or finite mixtures thereof.

17.2.4 Dealing with multivariate data – the overall algorithm

Next we turn to multivariate models where the data takes the form of an $n \times p$ data matrix $\boldsymbol{X}$. Delrieu and Bowman (2007) use independent parameter estimation per variable, that is, each column separately, but joint estimation over multiple variates is possible.

Thus for each variable $j = 1, \ldots, p$ we replace the data values x_{ij} by the individualised log divergences

$$\ell_j(i) = \log\left\{g(x_{ij};\theta_{1j})/g(x_{ij};\theta_{2j})\right\},$$

where θ_{1j} and θ_{2j} denote the parameters for the marginal distribution of variable j under the two models, and for simplicity we have dropped the hats on the parameter estimates. The new data set then can be analyzed with correlation decomposition methods – see Bowman (2009) for a procedural flowchart. For further shape investigations, the size of an overall profile for each individual can be summarised by the sum over variables

$$\sum_{j=1}^{p}\ell_j(i)$$

or the analogous mean ('Measure M') or variance ('Measure V') over p. Weights allowing for variable interdependencies can be derived from ordinations.

Parameter estimates can be theory-free, treating the observations as a phenomenological 'heap of data' (Bookstein 2014b) or contingent upon prevailing mechanistic theories or models of the underlying biological phenomena within $\hat{\theta}$ (neutral genetic drift, demic diffusion, positive selection, co-adaptation, biochemical dependence etc). Joint parameter estimation may be appropriate for specific *propter hoc* ascertainment or physiological relationships between columnar measurements j_a and j_b, say, and thus entail columnar collapsing (or aggregation). Importantly, parameter estimates may be covariate adjusted (or linearly modelled) with terms not included as columns in the later orthogonal decomposition (for instance when modelling time to event data as proportional hazards, etc). Sensitivity of parameter estimates to the data can be explored using Fisher's (Expected) Information as usual (Cox and Hinkley 1974); standard statistical theory applies.

However, the aim of using *lbf* divergences is to reach evidential conclusions regarding the samples and the observations therein, not parameter inferences. As such, the parameter estimates are useful condensates (and are often summary statistics themselves) but are not of actual interest in themselves – they are a means to an end, not an end in themselves. The end is key insights regarding distinctions between groups and the contributional structure of individual observations within them.

Some of the key questions in this exercise are the following:

- What is the question of interest?
- What is the reference group?
- What sort of group heterogeneity is to be explored?
- How to pose the parameterisation?
- How to estimate the parameters?

Answering these questions will lead the applied experimenter to the appropriate divergence to use.

The experimenter can use their own favourite validation tools for the estimation of the parameter estimates and their impact on where the transformed data sits in evidence space such as leave-one-out cross-validation, k-fold cross-validation, influence and leverage measures (e.g. Hat matrix, Cook's distance, partial leverage, DFFITS, even a datum's Mahalanobis distance itself!) as they desire to establish confidence in parameter estimability.

Then in terms of the robustness of the individualised divergence values themselves, an area to investigate (beyond such parameter estimate validation) is the influence or 'leverage' of any one observation on the divergence measured distinction (and its subsequent decomposition). This could be through a suitably scaled and summed (over individuals) divergence-based comparison of the 'goodness of fit' of set of data with parameter estimates estimated from all observations versus a set of data minus one observation (with parameter estimates estimated from all observations or all observations minus that datum); the summed divergence measure is computed by not including the 'held-out' datum in both cases. Here this divergence- based 'similarity measure' (*sensu* Kwitt and Uhl 2008) is measuring the distinction between the true full set of data and an approximation excluding one datum. In this way, one would not be looking at *where* the individual is in the sample divergence space (i.e. the standard displays in Delrieu and Bowman (2006a), 'Measure M' etc.) but would be looking at the number of bits (or, an average number of bits) that an individual datum contributes in defining what that full set final sample divergence space actually looks like. A datum which on exclusion causes a poor similarity measure is one that clearly has a large impact in defining what the final *lbf* space looks like; that is, it has high 'influence' or 'leverage' on the specified contrast.

Table 17.2 Seven data columns from a small unpublished case-control genetic study on cutaneous adverse reactions to drug treatment covering a total of 78 patients (28 cases and 50 controls) assayed for 241 single nucleotide polymorphisms and 8 HLA loci.

Subject	TNF	TNF	BAT1	...	HLA_A	HLA_B	...	ADR
	A_G_308A	A_G_238A_2	Ch6					hsr v sjs
	Ch6	Ch6	1933100					
	213	451						
CASE 1	A_G	G_G	A_A	...	02_03	08_44	...	hsr
CASE 2	G_G	G_G	A_A	...	03_68	14_44	...	sjs
CASE 3	A_A	G_G	A_A	...	01_01	08_08	...	hsr
↓								
CASE 28	A_G	G_G	A_G	...	01_01	08_38	...	sjs
CONT 1	G_G	G_G	A_G	...	24_26	07_38	...	-
CONT 2	A_G	G_G	A_A	...	01_02	07_44	...	-
↓								
CONT 50	G_G	G_G	A_G	...	02_25	18_44	...	-

For illustration only. ADR = Adverse drug reaction: hsr = hypersensitivity syndrome; sjs = Steven–Johnsons Syndrome (includes Toxic Epidermal Necrolysis in this study).

17.2.5 Brief new example

Table 17.2 gives a partial extract from a small unpublished case-control study on cutaneous adverse drug reactions covering 28 anticonvulsant drug 'Ell'-treated cases and 50 controls, assayed for 241 single nucleotide polymorphisms and 8 HLA loci. It is offered as a 'toy example' of stratification in medicine simply for method illustration.

There are $r = 2$ groups here with $p = 249$ variables. There are $n_1 = 28$ cases and $n_2 = 50$ controls. As in Pirmohamed et al. (2007) carriage of alleles at HLA loci were separated out from genotypes into new columns with three states for each allele (i.e. HLA-B44_57 carriage becomes two new binary carriage columns scored 'HLA-B44 YesNo' and 'HLA-B57 YesNo'). Their frequencies together with the frequencies of each genotype within each SNP column were estimated within cases and controls separately as Delrieu and Bowman (2006a). The observed individualised likelihood divergences (*lbf*s) $\ell_j(i)$ were calculated and each corresponding data point replaced accordingly (see Table 17.3).

Profiles of case-ness evidence are distinctly spiky (Figure 17.1, Lower). Eigen decomposition of a correlation matrix of the transformed data set shows, as in Bowman (2009), the likely importance of IL1 loci proteins in indicating propensity for these syndromes (i.e. they are aligned with case-control direction with loadings to the far right, see Figure 17.1, Upper). Overlay of the type of ADR shows that different syndromes clump in different parts of genetic space within this. Overlaying as a heat-map the carriage of the split-out HLA loci into further aggregated binary serotypes (Y/N) highlights HLA-A*68 carriage as a possible risk indicator together with perhaps the previously published implication of serotype B17 (HLA-B*57 and HLA-B*58) in some specific adverse drug reactions. Note the marked difference in the shapes of the average group profiles in Figure 17.1. Further large rigorous studies would be needed to confirm and prove these illustrative 'toy example' results. No change to medical practice is to be inferred.

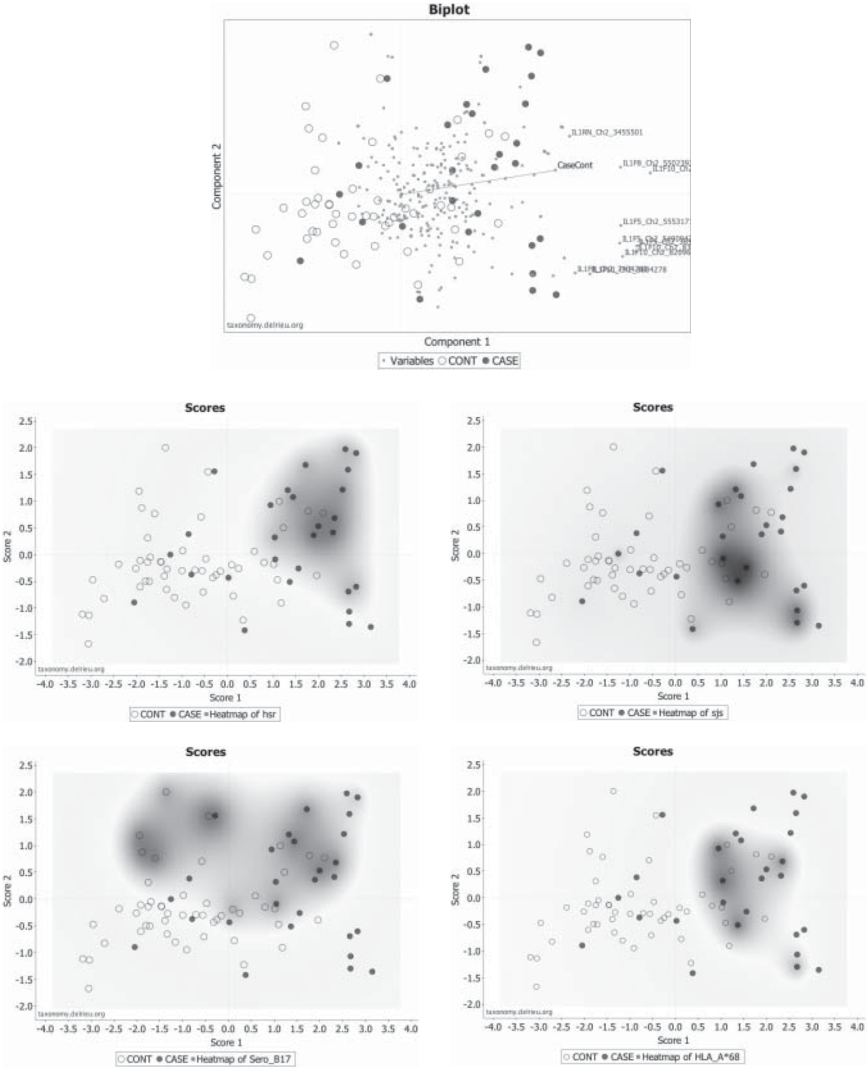


Figure 17.1 Eigen decomposition of evidence (‘toy’ illustrative genetic example from text). Upper biplot showing cases (dark circles), controls (open circles) and loadings (pale dots). Note good separation of cases from controls and in particular labelled loadings for IL1F proteins to the right. Almost horizontal case-control axis suggests robust well-informed study. Second row: to the left – heat-map for hsr cases; to the right – heat-map for sjs cases; Third row: to the left – heat-map for serotype B17 carriage; to the right – heat-map for HLA-A*68 carriage. Lower graphs (next page) show the shape of average case evidence profile over loci (grey mean shape) plotted above average control evidence profile (pale grey mean shape), with mean shape difference (case-control) plotted at foot in black. Shape of the comparative profile is impenetrable unless framed as an ordination of individuals in correlation space within the context of the biology.

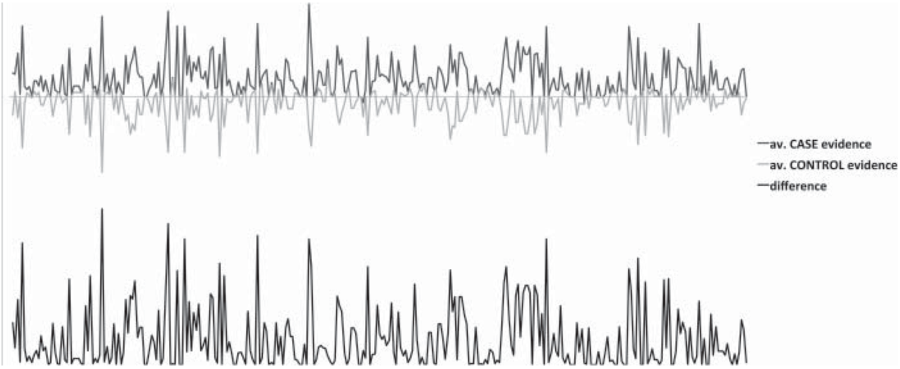


Figure 17.1 (*Continued*)

Table 17.3 Transformed data columns from Table 17.2 now containing surrogate numeric values for the individualised case-ness evidence of group distinction.

Subject	TNF A_G_308A Ch6 213	TNF A_G_238A_2 Ch6 451	BAT1 Ch6 1933100	. . .	HLA_A02	HLA_B44	. . .
CASE 1	0.565	0.134	0.236	. . .	0.054	0.410	. . .
CASE 2	−0.266	0.134	0.236	. . .	−0.060	0.410	. . .
CASE 3	0.565*	0.134	0.236	. . .	−0.060	−0.274	. . .
↓							
CASE 28	0.565	0.134	0.054	. . .	−0.060	−0.274	. . .
CONT 1	−0.266	0.134	0.054	. . .	−0.060	−0.274	. . .
CONT 2	0.565	0.134	0.236	. . .	0.054	0.410	. . .
↓							
CONT 50	−0.266	0.134	0.054	. . .	0.054	0.410	. . .

The original data co-occurrence pattern stays the same. Not all columns shown. Each cell contains the observed individualised divergences (in base 2) corresponding to each original data point (here equivalent to log(relative genotype frequency) comparing cases to controls). * = simply a co-incidence that the *lbf* is the same for A_A as for A_G in this example.

17.2.6 Justification for the consideration of individualised divergences

Why consider individual data points? After all as Abbasi (2012) points out *'case reports are usually deemed to be the lowest form of evidence. What can you really learn from a single case?'*. The key is that within the context of a group distinction, variation between individuals (especially 'sports' or 'outliers') can yield useful insights for the applied experimenter. In the Popperian paradigm, it is the discrepancies from the null hypothesis that matter, not issues of complete agreement with beliefs. Thus one focuses, not on a whole sample of individuals, but only on the ith individual. Assuming the collection of individuals in the populations give a good coverage of each population, then this likelihood approximation is a reasonable indicator of the importance of the evidence which that data point gives to the directed or undirected population comparison. An observed sample χ^2 (in some sense) is fractionated by the ‘contribution’ or importance of individual row observations in the observed data space (or column space) of the contrast. Although this is an analytical approach, it has many features of non-analytical case-based reasoning practised by diagnostic physicians (Norman et al. 2007) – the data ‘speaks’ for itself. Treating the data as a phenomenological ‘pile of data’ relies little on underlying biomedical knowledge (although that can be built in to the parameterisation and choice of contrast). The group distinction is acting like prior experience. Examining each individual is akin to continued (clinical) taxonomic practice and exposure. The scaled evidence deviations of individuals within the posed contrast (aka the hypothetical grouping) offers the structured feedback to prompt the investigator to learn. The deployment of this *lbf* mathematical strategy is mimicking the clinician.

Adding dummy variables into this non-linear kernel projection of the data (see Pirmohamed et al. 2007) reapportions variation in information into context, the final dimension of the problem now being greater than the dimension of the original data. The parametrisation of binary indicator variables can be $[0, 1], [-1, 1], [-\frac{1}{2}, \frac{1}{2}]$ and so on according to convenience of interpretation or display scaling. Conditional characters (Jardine and Sibson 1971) can be dealt with by defining synthetic variables as cross products of dummy indicator variable vectors with the data. Column recoding can be used such as re-expressing genotype data as homozygosity measures (when inbreeding is to be examined). Aggregations of divergences to genes or ontologies or known models or extra hypotheses can be carried out by simple (weighted) columnar summation in *lbf* space (Delrieu and Bowman 2006a) and overlay display or by co-analysis. For SNPs, such aggregations collapse ‘spilt-plot’ variation into the ‘main plot’ gene stratum and reduce noise (yet increase the comparative rugosity of the resultant variate). Aggregations can be linear or non-linear functions, with appropriate covariance adjustment for the resultant concomitant (biased) reduction of variation (outlined in Delrieu and Bowman 2007). Hence non-linearity or curvature *among ordered columns* can be explored through posing three level aggregation weights such as $[-1, 0, 1]$ and $[1, -2, 1]$ across columns. Interactions and residuals can be investigated between variables (so for j columns there are potentially $j(j + 1)/2$ interaction columns (see Bowman and Delrieu 2009a) formed by categorical cross-products before translation into evidence space or by multiplication in data space first, for example, Fisher’s Iris data, where petal surface area (petal length × petal width) is a good group discriminant. Standard statistical tools are available to examine the stability of likelihood-based divergences. For example, let any one of these observed divergence measures mentioned earlier be called τ; then, the slope $\frac{\delta\tau}{\delta\theta}|_{\hat{\theta}}$ and the curvature $\frac{\delta^2\tau}{\delta\theta^2}|_{\hat{\theta}}$ with respect to the estimated parameters $\hat{\theta}$ can be explored for robustness over the estimation space. Since the *lbf* algebra is linear, all standard mathematical tools are available for easy deployment according to need. The algebra is easy to deploy for the applied scientist.

17.3 The future: challenging data

There are many new challenges in the use of these informational individualised divergences measures in simultaneously analysing the evidence of biological and medical profiles.

17.3.1 Contrasts of more than two groups

The preceding analysis was concerned with $r = 2$ groups, typically a case group and a control group. Here we present some new results for comparisons between $r > 2$ groups, represented in the data matrix $\boldsymbol{X}$ as r row blocks. For convenience, we concentrate on the case of $r = 3$ groups represented by samples from multidimensional densities ordered in some (other) covariate(s) space of interest.

Effectively in the information contrast-based outline mentioned earlier, the observed density for a case group (let us now denote this case set by subscript (2)) is measured by a ‘distance’ to that of a control *group* (let us now denote this control set by subscript (0). Now consider intercalating a third density (for instance a ‘mild’ case group) denoted by a (1) subscript; see Table 17.4. The objective is to understand the comparative size and shape of the data in all three groups or *‘row blocks’*.

The two polynomial contrasts (x and y) in the table are orthogonal (as $\sum x_k y_k = 0$), that is, they are the ‘*row*-analogue’ of non-overlapping *columnar* aggregates such as genes in SNP studies. Applying them both to data and simultaneously decomposing the covariance matrix will not induce extra structure in the results as their variances are constant (see Delrieu and Bowman 2007). Rescaling the quadratic contrast in the last row of the table to give the same variance as the unscaled linear contrast * in the table retains the latter’s epistemological interpretation yet does not introduce unequal variances into later decompositions. Then, it is possible to construct new ‘Kullback–Leibler-like’ expected individualised likelihood divergences structured as in Table 17.5. As before, likelihood measures (L) can be appropriately substituted for the densities when dealing with an actual sample of data.

These quadratic contrasts are not obviously *directed* divergences. However, they *are* asymmetric measures of how non-linear or curved the approximating space of densities are between the two extremal ’row block’ samples through the intermediate ’row block’ sample. The typical curved-ness in Table 17.5 is the expected or typical loss of information when g_1 is used to approximate g_2 *and* g_0 *together* (i.e. one is ‘looking outwards’) along a gradient of row blocks. The typical non-linearity in Table 17.5 is the expected or typical loss of information when g_2 and g_0 *each* are used to approximate g_1 (i.e. one is ‘looking inwards’ along a row block gradient). By virtue of the different expectation in the log density ratio inside the Kullback–Leibler divergences, these may yield different results depending on the statistical distribution. It will be easier (and less confusing!) to use only one or the other in any comparative analysis of sample profiles since the two forms are not necessarily orthogonal. In practice, for well-behaved data, they should be broadly equivalent. However, the linear formulation (Delrieu and Bowman 2007) can be simultaneously analysed with either newly posed quadratic form as they remain orthogonal. Again, most importantly the *lbf* individualisation that is carried out on each data column at a time over all the rows

Table 17.4 Proposed framework for contrasts of three groups (*row blocks*) leading to new divergences for profile comparisons.

Contrast				$\sum w_k$	$\sum w_k^2$	Meaning
	Control	‘Mild case’	Case			
	w_0	w_1	w_2			
Linear (x)	-1	0	1	0	2	Case-ness*
Quadratic (y)	1	-2	1	0	6	
*Scaled to match**	$\frac{1}{\sqrt{3}}$	$-\frac{2}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	0	2	Curved Case-ness

See text for detailed explanation. Subscripts denote group (*row blocks*). The quadratic contrast can be considered as the sum of two contrasts $[1, -1, 0]$ & $[0, -1, 1]$. * Base comparison.

Table 17.5 New *Expected* divergences using the contrasts from Table 17.4.

Expected divergence	Meaning
$E_2\left[\log(\frac{g_2}{g_0})\right]$	Typical Case-ness
$E_{0\&2}\left[\frac{1}{\sqrt{3}}\log(g_0) - \frac{2}{\sqrt{3}}\log(g_1) + \frac{1}{\sqrt{3}}\log(g_2)\right]$, or $E_{0\&2}\left[\frac{2}{\sqrt{3}}\log(\frac{g_{0\&2}}{g_1})\right]$	Typical Curved-ness
$E_1[E_1\left[-\frac{1}{\sqrt{3}}\log(g_0) + \frac{2}{\sqrt{3}}\log(g_1) - \frac{1}{\sqrt{3}}\log(g_2)]\right]$, or $E_1[E_1\left[\frac{1}{\sqrt{3}}\log(\frac{g_1g_1}{g_0g_2}]\right]$, or $E_1\left[\frac{1}{\sqrt{3}}\log(\frac{g_1}{g_0})\right] + E_1\left[\frac{1}{\sqrt{3}}\log(\frac{g_1}{g_2})\right]$	Typical Non-linearity

See text for detailed explanation. Subscripts denote group (*row blocks*). g are densities from Table 17.4 – these can be replaced by appropriate likelihoods (L). $E_b\left[\log(\frac{g_b}{g_a})\right] =$ expected loss when density g_a in the log function denominator approximates density g_b in the log function numerator. Note how contrast coefficients play out. Also: recall that $E[x] = \mu$ and $E[\mu] = \mu$, so $E_1[E_1[...]] = E_1[...]$; $E_{0\&2}[...] =$ expectation over pooled (w_0, w_2) density; so the quadratic expectations are only over x domain of log ratio numerator in this example of distinct 'row block' densities.

is the same (i.e. there is no change in basis when calculating these linear and quadratic divergences for decomposing the shape relations of real data; see Bowman 2009). Recall that similar shapes in a (relative) log space means that the actual *arithmetic* shapes may be dramatically different (cf. well-known allometry, and the importance of geometric means in differential shapes).

This approach could be used for instance in generating knowledge about the multidimensional genetic shape of distinctions between mild and severe carbamazepine hypersensitivity cases versus controls (see Bowman and Delrieu 2009a; Zhang et al. 2008), that is, as a 'case, (another) case, control study'. Here, different variables indicated by the simultaneous covariance (or correlation) decomposition of both linear and row block quadratic contrasts would suggest non-linearity or curvature, that is, that the mild mid-group was not a centrally placed gradation between the structure of the extremal case and control samples (an inconsistent 'dog-leg'). If no particular curved-ness or non-linearity is detected, then the three densities sit in some sense on the same linear manifold of structure over the ordering (other) covariate(s) of interest; there are consistent shape differences as one moves along the ordering of disease severity , that is, they have similar regular fundamental components that smoothly interpolate. In some sense, the components of $g_1 - g_0$ and the components of $-g_2 + g_1$ are each a fraction $\frac{\sqrt{3}}{2}$ of the components for $g_2 - g_0$ (i.e. smoothly equivalent *size* differences in profile shape), *or*, the shape of the scaled $g_1 - g_0$ distance components *is compensated* by the shape of the scaled $-g_2 + g_1$ distance components to resolve to the overall component shape for $g_2 - g_0$; that is, there is a consistent 'dog-leg' in the space of profile differences.

Some of many open questions concerning these new divergences for profile shapes include the following. What might be the *practical* merit of three-group 'control, *another* control, case multidimensional profile studies'? What do cubic, quartic and so on contrast polynomials look like in an individualised Kullback–Leibler formulation for finer-ordered densities? What is the general form (and is there an asymptote) as the number of distinct row blocks tends to ∞? Could this approach be extended to explicitly compare decompositions on smooth interpolations between the covariance matrices of extremal groups using the result of Dryden et al. (2009) where covariance matrices grade on a geodesic in log space? Can one use the method of Felsenstein (1985) to calculate the contrast coefficients over an ordered tree space of densities? Is there a proof that all or a subset of sub-tree contrasts are orthogonal when applying this approach to contrasting sampled densities arranged in such a notional tree space? If not, how does one adjust for partially overlapping ontologies (in row or even column space) inducing biased covariance structure decompositions? What are the information radii in this polynomial space? What does the average-over-p profile *size* 'Measure M' (Delrieu and Bowman 2006a) mean using these quadratic contrasts? and so on.

17.3.2 Other data distributions

Many medical experiments are based on survival assessments or time to disease expression and so on (for instance, in oncology trials). Some oncology studies measure many such measures, for example, PFS, OS and so on. Kent and O'Quigley (1988) outline an information gain (Kullback–Leibler divergence) measure to compare statistical models over a censored time to event sample. However, this does not decompose the contribution each data point makes. Multi-column data from such experiments could be examined using individualised divergences (*lbf*s), but the issue arises as to what individualised divergence (*lbf*) value to give to a right censored data point in column j of observed variables observed at time t for the ith individual in the kth population, $k = 1, 2$, as its true time of the event is actually unknown (although *per force* must be greater than c_{ij}, the censoring time on the jth time-to-event variable for the ith individual). The obvious choice for $lbf_i(j)$ is the ratio of the two survivor functions at the censoring time c_{ij}, that is, 'all I know about the evidence is what I knew when the individual was lost to observation'. This is taking the same view as the philosophy of the Kaplan–Meier non-parametric estimator of hazard rate (see Kaplan and Meier 1958). However, given the smooth form of the fitted exponential family distribution for that (and all other) group's data, then the *estimated* likelihood functions *are* known rightwards of the censor time c_{ij} and would thus be being ignored (although the real event time will be somewhere in this interval). A dummy conditional variable of censor time may instead highlight issues in any analysis of censored data that inform what sort of modelling of the hidden tail should be. A sensitivity analysis of the results of any profile shape that contains censored data could be done allowing for this extrapolation or not – although in well-posed data, that is, well followed-up data with low censoring, little departures are expected.

Mardia and Jupp (1999) outline the statistics of directional data but publication of divergences for von Mises and other circular multidimensional distributions awaits. This is not just for the integration of say, just directional covariates into analyses. But they are needed for the smooth decomposition of medical image data comparisons of, say, diseases causing head and neck arterial tortuosity *without* the use of summary indices (such as 'VTI' see Morris et al. 2011), arc-chord ratio or curvature and torsion measures of the midline curve across individuals. A further open challenge is to pose individualised divergences of statistical distributions for a unified inverse approach to the aetiology of medical shape differences that contain full or partial reflections between individuals along with traditional deformations (e.g. in *situs inversus*). Many of these gross malformations that, on the face of it, appear to defy simple 2D warp-analysis (see Bookstein 1986) actually have straightforward genetic lesions. Individuals show statistical variation in the placement of these organs – so distributions and an algebra in a smooth space is needed beyond aspects of just bilateral symmetry (see Kent and Mardia 2001; Mardia et al. 2000). Smoothness is needed to accommodate data from *situs ambiguus* or heterotaxy individuals, where *situs* cannot be easily denoted other than by full shape description.

Another open challenge is how to pose individualised divergences to deal with comparing strictly branching structures (such as the bronchi in lungs) or complicated tree-like structures such as anastomoses (as in angiogenesis – the growth of vascular structures within tumours). Both can be extremely complex and have long puzzled biologists but perhaps could be described by simpler decompositions of combinations of *lbf*s from straightforward stochastic processes and coalescents. Particular structures could show both between and within individual variation in estimated parameters much as in simple longitudinal growth profiles and thus yet to be posed multi-level divergences may be required. It also remains open to explicitly pose appropriate divergences for the simultaneous analysis of extreme value data; unaligned sequence data from Next Generation Sequencing initiatives; spike-train data from neuro-physiological recordings; continuous EEG/ECGs and so on without a pre-processing step of motif or feature selection and so on. Range data known to be Gumbel distributed does not have closed analytical form for its density. Divergences for such data are needed that do not rely upon mapping into variances (via that the standard deviation of a sample is approximately equal to one fourth of the range of the data) and then deploying χ^2 divergences derived from those of Gamma variables. Folding multidimensional volatility data from, say, stock-markets as a column into the approach could be via such approximate normal theory or by using exact Wishart divergences. Many opportunities for extension lie around.

17.3.3 Other methods

Plugging in observed values may engender bias in this methodology. A full Bayesian approach could, of course, be used instead. However, this runs the risk of re-introducing potential operational opacity to the applied experimenter. Permutation (Delrieu and Bowman 2006a) or empirical bootstrapping can re-instate the lost stochasticity in part by engendering not only appropriate variation in the parameter estimates (i.e. restore the lost variability in the 'maximum' likelihood data weights) but also variation in the location of the actual data (i.e. compensate for the lost density and the lack of formal integration over the domain). Such an empirical ploy obviates the need for prior data densities and hyper-parameter distributions and a fully rigorous Bayesian approach to the use of individualised divergences for an applied experimenter. Bayesian posterior parameter estimates are possible using a conjugate prior, for example, Poisson, exponential, normal (with known mean), Pareto, Gamma (with known shape parameter), and inverse gamma (with known shape parameter) and have yet to be used. Prior beliefs can be inserted into divergences by appropriate integration. Also, to date nobody has used divergences in analyses based on *three* variate tensors, nor have many others chosen and interpreted equivalent SVD decompositions of such. This could be of use in genetic profile shape comparisons where second-order (i.e. three locus) linkage disequilibrium is of interest. Whether sufficient amount of accurate and reliable data is available to support such an extended approach remains to be seen. And, of course, other decomposition methods than SVD (aka eigenanalysis) could be used, that is, independent components analysis, non-negative matrix decomposition and so on. The field is wide open.

References

Abbasi K 2012 Why patient consent is best practice. *Journal Royal Society of Medicine* **105**(10), 407.

Akaike H 1974 A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**(6), 716–723.

Alfirevic A, Vilar FJ, Alsbou M, Jawaid A, Thomson W, Ollier WER, Bowman CE, Delrieu O, Park BK and Pirmohamed M 2009 TNF, LTA, HSPA1L and HLA-DR gene polymorphisms in HIV positive patients with hypersensitivity to co-trimoxazole. *Pharmacogenomics* **10**(4), 531–540.

Barnett NS, Cerone P, Dragomir SS and Sofo A 2002 Approximating Csiszár f-divergence by the use of Taylor's formula with integral remainder. *Mathematical Inequalities and Applications* **5**(3), 417–434.

Bookstein FL 1986 Size and shape spaces for landmark data in two dimensions. *Statistical Science* **1**(2), 181–242.

Bookstein F 2013 *Measuring and Reasoning: Numerical Inference in the Sciences*. Cambridge University Press.

Bowman CE 2009 Megavariate genetics: what you find is what you go looking for In 19th Altenberg Workshop of Theoretical Biology. Measuring Biology - Quantitative methods: Past and Future. (orgs. Bookstein F and K Schaeffer) Konrad Lorenz Institute, Austria. Published in: *Biological Theory* **4**(1), 21–28.

Bowman CE 2013 Discovering pharmocogenetic latent structure features using divergences. *Journal of Pharmacogenomics & Pharmacoproteomics* **4**(1), doi: 10.4172/2153-0645.1000e134

Bowman CE and Delrieu O 2009a Correlation laplacians, haplotype networks and residual pharmacogenetics In *Statistical Tools for Challenges in Bioinformatics* Gusnanto A, Mardia KV and Fallaize CJ (eds), University of Leeds, pp. 25–31.

Bowman CE and Delrieu O 2009b Immunogenetics of drug-induced skin blistering disorders. Part I - Perspective. *Pharmacogenomics* **10**(4), 601–621.

Bowman CE and Delrieu O 2009c Immunogenetics of drug-induced skin blistering disorders. Part II - Synthesis. *Pharmacogenomics* **10**(5), 779–816.

Bowman C, Delrieu O and Roger J 2006 Filtering pharmacogenetic signals In *Interdisciplinary Statistics and Bioinformatics* Barber S, Baxter P, Mardia K and Walls R (eds), University of Leeds, pp. 41–47.

Bowman CE and Jones CI 2010 Genetic evidence-of-no-interest, pathway Sudoku and platelet function In *High-throughput Sequencing, Proteins and Statistics* Gusnanto A, Mardia K, Fallaize CJ and Voss J (eds), University of Leeds, pp. 53–58.

Charalambous C, Delrieu O and Bowman C 2008 Whole genome scan algebra and smoothing In *The Art and Science of Statistical Bioinformatics* Barber S, Baxter PD, Gusnanto A and Mardia KV (eds), University of Leeds, pp. 21–27.

Congdon P 2006 *Bayesian Statistical Modelling*, 2nd ed. Wiley-Blackwell.

Cox DR and Hinkley DV 1974 *Theoretical Statistics* Chapman and Hall.

Dadpay A, Soofi ES and Soyer R 2007 Information measures for generalised gamma family. *Journal of Econometrics* **138**, 568–585.

Delrieu O and Bowman C 2005 Visualisation of gene and pathway determinants of disease In *Quantitative Biology, Shape Analysis, and Wavelets* Barber S, Baxter PD, Mardia KV and Walls R E (eds), University of Leeds, pp. 21–24.

Delrieu O and Bowman C 2006a Visualising gene determinants of disease in drug discovery. *Pharmacogenomics* **7**(3), 311–329.

Delrieu O and Bowman C 2006b Visualisation of gene by gene interactions in pharmaco-genetics In *International Congress Of Human Genetics*, Brisbane Australia, 6–11th August 2006 (poster).

Delrieu O and Bowman C 2007 On using the correlations of divergences In *Systems Biology and Statistical Bioinformatics* Barber S, Baxter PD and Mardia KV (eds), University of Leeds, pp. 27–35.

Dobson AJ 1983 *An Introduction to Statistical Modelling* Chapman and Hall.

Dryden IL, Koloydenko A and Zhou D 2009 Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging In *Statistical Tools for Challenges in Bioinformatics* Gusnanto A, Mardia KV and Fallaize CJ (eds), Leeds University Press, , Leeds, pp. 43–46.

Dryden IL and Mardia KV 1998 *Statistical Shape Analysis* Wiley-Blackwell, 376 pp.

Endres DM and Schindelin JE 2003 A new metric for probability distributions. *IEEE Transactions on Information Theory* **49**(7), 1858–1860.

Felsenstein J 1985 Phylogenies and the comparative method. *American Naturalist* **125**(1), 1–15.

Jardine N and Sibson R 1971 *Mathematical Taxonomy*. John Wiley & Sons.

Jeffreys H 1946 An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A* **186**, 453–461.

Kaplan EL and Meier P 1958 Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

Kent JT and Mardia KV 2001 Shape, Procrustes tangent projections and bilateral symmetry. *Biometrika* **88**, 469–485.

Kent JT and O'Quigley J 1988 Measures of dependence for censored survival data. *Biometrika* **75**(3), 525–534.

Kullback S 1959 *Information Theory and Statistics*. Dover Books on Mathematics (reprinted 1997).

Kullback S and Leibler R 1951 On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.

Kwitt R and Uhl A 2008 Image similarity measurement by Kullback-Leibler divergences between complex wavelet subband statistics for texture retrieval. In *IEEE International Conference on Image Processing (ICIP 08)*, 4 pp.

Mardia KVM 2007 *The Scientific Foundations of Jainism*, *Lala Sunder Lal Jain Research Series*. Motilal Banarsidass, New Delhi, 142 pp.

Mardia KV, Bookstein FL and Moreton IJ 2000 Statistical assessment of bilateral symmetry of shapes. *Biometrika* **87**, 285–300.

Mardia KV and Jupp PE 1999 *Directional Statistics*. Wiley-Blackwell, 456 pp.

Mead R 1990 *The Design of Experiments: Statistical Principles for Practical Application* Cambridge University Press.

Morris SA, Orbach DB, Geva T, Singh MN, Gauvreau K and Lacro RV 2011 Increased vertebral artery tortuosity index is associated with adverse outcomes in children and young adults with connective tissue disorders. *Circulation* **124**, 388–396.

Nguyen X, Wainwright MP and Jordan MI 2005 On information divergence measures, surrogate loss functions and decentralized hypothesis testing. In *43rd Annual Allerton Conference on Communication, Control and Computing, Allerton, IL* September 2005, 10 pp.

Norman G, Young M and Brooks L 2007 Non-analytical models of clinical reasoning: the role of experience. *Medical Education* **41**, 1140–1145.

Pirmohamed M, Arbuckle J, Bowman C, Brunner M, Burns D, Delrieu O, Dix L, Twomey J and Stern R 2007b Investigation into the multi-dimensional genetic basis of drug-induced Stevens-Johnson syndrome and toxic epidermal necrolysis. *Pharmacogenomics* **8**(12) 1661–1691.

Shannon CE 1948 A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, Part I, 379–423.

Sibson R 1969 Information Radius. *Zeitschrit Wahrscheinlichkeitstheorie verw. Geb.* **14**(2), 149–160.

Taneja IJ 2005 Generalized symmetric divergence measures and inequalities. *arXiv:math.ST/0501301v1* 19 Jan 2005, 1–30.

Tobler WR 1970 A computer movie simulating urban growth in the Detroit region. *Economic Geography Supplement* **46**, 234–240.

Tribus M 1961 *Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications*. D. Van Nostrand Company Inc., New York.

Zhang JE, Alfirevic A, Malhotra R, Park KB and Pirmohamed M 2008 CD28, CTLA4, ICOS and PD1 genes and carbamazepine-induced hypersensitivity. *European Journal of Dermatology* **18**(2) 246, abstract 0067.

18

Proteins, physics and probability kinematics: a Bayesian formulation of the protein folding problem

Thomas Hamelryck[1], Wouter Boomsma[1], Jesper Ferkinghoff-Borg[2], Jesper Foldager[1], Jes Frellsen[3], John Haslett[4] and Douglas Theobald[5]

[1]*Department of Biology, University of Copenhagen, Copenhagen, Denmark*

[2]*Biotech Research and Innovation Center, University of Copenhagen, Copenhagen, Denmark*

[3]*Department of Engineering, University of Cambridge, Cambridge, UK*

[4]*School of Computer Science and Statistics, Trinity College, Dublin, Ireland*

[5]*Biochemistry Department, Brandeis University, Waltham, MA, USA*

18.1 Introduction

Proteins are the workhorses of the living cell. They are responsible for our immune response, propagation of signals through nerves, digestion of food, oxygen transport, and many other vital tasks. In medicine, proteins are the target of most drugs. In biotechnology, proteins are for example used in the production of biofuels, chemicals, food, and feed. Understanding

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

proteins and their behavior in molecular detail is thus an important task in science, medicine, and technology.

Proteins are linear polymers of amino acids. In a watery environment, many proteins spontaneously fold up into a distinct three-dimensional (3D) structure, the so-called *folded state* or *folded conformation* (Dill 1999; Dill and Chan 1997; Dill and MacCallum 2012). The folded state of a protein is in most cases uniquely defined by its amino acid sequence – a discovery that earned the American biochemist Christian B. Anfinsen the Nobel prize in chemistry in 1972 (Anfinsen 1973). Nonetheless, it should be noted that all proteins are, to some extent, dynamic molecules that undergo movements. Indeed, proteins exist that are so flexible that they do not fold into a specific 3D shape (Tompa 2002). However, such proteins might still adopt one or more specific 3D shapes when interacting with binding partners such as other proteins or nucleic acids. In this chapter, we will mostly ignore the dynamical aspects of proteins, though these are also potentially within the scope of the models we discuss here (Boomsma et al. 2014; Harder et al. 2012; Olsson et al. 2013, 2014).

The 3D shape and the dynamical properties of a protein are crucial for its function. In the current genomic era, it has become straightforward to determine the sequences of proteins on a vast scale. The case for their matching 3D structures is unfortunately very different. Today, it is still very expensive and time craving to determine the structure of a protein in atomic detail. Therefore, there is great interest in obtaining the 3D structure of a protein by computational means, starting from the protein sequence. In addition, there are also strategies to obtain a protein structure in atomic detail that fall in between these two extremes, for example, by including low-resolution experimental data that can be easily obtained (Lipfert and Doniach 2007), by using the known structure of related proteins as templates (Dill and MacCallum 2012) or by using evolutionary information to infer plausible amino acid contacts in the folded state (Marks et al. 2011).

The prediction of a protein's structure in atomic detail from its sequence is just one of the many manifestations of what has been called *"the protein folding problem"* (Dill and MacCallum 2012). Indeed, the protein folding problem is not limited to the prediction of the static 3D structure of a protein. It also covers related problems such as protein design – which concerns designing a protein sequence that folds into a given 3D shape – simulating the dynamics of a protein, simulating the folding process in atomic detail, designing drugs that specifically bind disease-related proteins, modeling the way proteins bind to each other and to other biomolecules, and so on.

Current methods take two different roads; they are either knowledge based (Koppensteiner and Sippl 1998; Simons et al. 1997; Sippl 1990) or physics based (Duan and Kollman 1998; Lindorff-Larsen et al. 2011). Knowledge-based methods make use of the database of known protein structures and are – or ideally should be, at least – essentially statistical methods. Physics-based methods make use of physical energy functions to simulate the entire folding process. The latter methods are still too time-consuming for large-scale, routine use in protein structure prediction and currently still far from perfect (Faver et al. 2011). In practice, most methods use a blend of knowledge-based and physics-based approaches. Programs that are based on such methods, including ROSETTA (Bradley et al. 2005; Simons et al. 1997) and I-TASSER (Roy et al. 2010), have demonstrated that prediction of protein structure in the absence of known related structures is now sometimes successful.

Despite frequent claims to the contrary, the protein folding problem is in many respects an open problem. Currently, it is still not routinely possible to predict the structure of an arbitrary protein if no closely related structures are known, or to routinely design protein sequences that fold into a given structure. Nonetheless, triumphant claims that "the protein-folding problems is now solved" have been somewhat of a tradition in the field for decades. A recent example of such claims can be found in an editorial article that appeared in Science in 2008, boldly entitled "Problem solved (sort of)" (Service 2008). However, the current state of the field is rather well summarized by Eugene Shakhnovich's poignant and critical reaction to this article:

> The article "Problem solved (sort of)" highlights a bizarre state of the protein-folding field where some claim that the problem is solved but keep the solution in strict confidence. While useful in emphasizing (albeit not for the first time) the statistical-mechanical aspect of the protein-folding problem, the early phenomenological models mentioned in the article do not provide a solution even in the most approximate sense. The assertion that "proteins fold because their energy landscape is funneled" is hardly satisfactory because neither the protein energy landscape, nor the funnel, are clearly defined, and analogies in science cannot substitute for research. [...] However, coarse-grained statistical-mechanical models did advance our understanding of folding proteins. They explicitly demonstrated that proper sequence selection can guarantee a fast and reliable folding of large model proteins, providing also a conceptual foundation of modern protein design. While important in removing the shroud of miracle from the problem, these early studies provided only a limited insight on how real proteins fold. Recent successful *ab initio* sequence-based all-atom folding of several small proteins showed that mechanistically folding proteins is a far more complex process than suggested by analogies and even by coarse-grained models. It remains to be seen which aspects of the observed protein-folding mechanisms are general and which vary between individual proteins. Emergence of computationally tractable yet realistic protein models combined with enhanced computer power and advanced experimental approaches make it possible, for the first time, to obtain an atomistic picture of statistical folding pathways. Certainly we are at the beginning of the path towards solving the protein-folding problem.

Currently, it is fair to say that none of the knowledge-based methods for protein structure prediction are based on a well-defined Bayesian model. Most knowledge-based methods build on two key methodologies: fragment libraries (Simons et al. 1997) and *knowledge-based potentials of mean force* (KPMFs) (Koppensteiner and Sippl 1998; Sippl 1990).

KPMFs are energy functions that are estimated from the set of known protein structures (Koppensteiner and Sippl 1998; Pohl 1971; Sippl 1990). These potentials typically concern pairwise distances between amino acids in a protein structure. Such knowledge-based potentials should not be confused with the well-justified potentials of mean force as, for example, used in the physics of liquids (Chandler 1987). Rather, KPMFs are *ad hoc* constructions that aim to mimic their rigorous counterparts in physics (Ben-Naim 1997; Thomas and Dill 1996).

Fragment libraries (Simons et al. 1997) are used to assemble plausible protein structures by tying together short fragments of existing protein structures (Simons et al. 1997). They are typically used in Monte Carlo methods as proposal methods, and thereby also bring in an associated – and typically unknown – energy term (Boomsma et al. 2014; Przytycka 2004).

Fragment libraries and KPMFs were proposed two decades ago and still form the backbone of most knowledge-based protein structure prediction methods. Unfortunately, neither of them have an underlying sound probabilistic model. In the case of KPMFs, it was until very recently not even clear why these potentials were to some extent successful in the first place (Ben-Naim 1997; Hamelryck et al. 2010; Thomas and Dill 1996).

Currently, success in protein structure prediction, at least in the absence of a related structure that can be used as a convenient template to start from, seems to be stagnating (Kryshtafovych et al. 2014). This poses a problem, as genome-wide prediction of protein structure certainly requires covering such "orphan proteins." For example, the recently reported genome of the giant Pandora virus revealed more than 2500 putative protein-coding sequences, 93% of which are without recognizable homologue in the structural databases (Philippe et al. 2013). We postulate that the current stagnation can be broken by the development of state-of-the-art, well-justified, and computationally efficient Bayesian models and methods. In addition, such models will also allow to evaluate the precision of the predictions. In this chapter, we outline such a model. In our formulation, protein structure prediction simply corresponds to sampling from a well-defined posterior distribution obtained from applying the Bayesian probability calculus.

The underlying probabilistic model is based on three pillars. First, graphical models, and more specifically dynamic Bayesian networks, represent the sequential nature of a protein as a linear polymer (Boomsma et al. 2008; 2014; Hamelryck et al. 2006; Harder et al. 2010). These models capture the shape of a protein on a local length scale, but without adequately modeling the global features. Second, directional statistics (Mardia and Jupp 2000) – the statistics of angles, directions, and orientations – is used to model the main degree of freedom when representing protein structure, which are the dihedral angles. Finally, inference of protein structure is a *multi-scale problem* (Ferreira and Lee 2007), much like modeling the movements of an animal might involve covering minute movements as well as much larger patterns of migration. Rather than developing one complicated model that covers all scales, we develop a short-range (*local*) and a long-range or global (*non-local*) model. The *reference ratio method* (RR method) – which constitutes the third and final pillar – is used to combine the local and non-local models into a joint posterior distribution (Borg et al. 2012; Frellsen et al. 2012; Hamelryck et al. 2010, 2013; Mardia et al. 2011; Mardia and Hamelryck 2012). The RR method corresponds to a special – and little known – case of Bayesian belief updating called *Jeffrey's conditioning* or *probability kinematics* (Jeffrey 2004).

It remains to be seen how successful this approach will be. Preliminary results are promising and demonstrate that the approach is computationally efficient (Valentin et al. 2014). Although it will take time to bring this method up to par with *ad hoc* methods that have been manually honed and refined for decades by many developers and users, a well-defined Bayesian approach to protein structure prediction – and its many advantages – is now within reach.

18.2 Overview of the article

The structure of the chapter is as follows. First, we outline the nature of the probabilistic model of protein structure that we want to develop. Then, we present a brief overview of protein structure in terms of local and non-local structure and discuss its implications for the development of the probabilistic model. The resulting model consists of a prior distribution concerning local structure and a likelihood concerning non-local structure. The latter two are only briefly discussed, as they have been discussed extensively elsewhere (Boomsma et al. 2008; Valentin et al. 2014).

Due to the nature of the models, combining the prior and the likelihood according to the Bayesian calculus requires a special technique that we called the *reference ratio method* (RR method). We discuss the RR method and why it is needed and give five different ways in which the method can be derived and understood. Finally, we point out how the RR method solves a twenty-year-old conundrum regarding the nature of KPMFs and discuss its interpretation as a maximum entropy method.

18.3 Probabilistic formulation

The problem of protein structure prediction from amino acid sequence (Dill and MacCallum 2012) can be stated in the following way. We want to formulate the following probability distribution:

$$p(\mathbf{x} \mid \mathbf{a}, M_N), \tag{18.1}$$

where $\mathbf{x}$ is a – typically high-dimensional – vector that specifies a protein's structure and $\mathbf{a}$ is the amino acid sequence, which is a vector of symbols chosen from an alphabet with twenty letters. Each letter represents one of the twenty different naturally occurring amino acids. M_N is the underlying model; the subscript N indicates that the model refers to the protein's native folded state.

The subscript N deserves some more explanation. Roughly speaking, a protein can be in an unfolded or folded state. For our purposes, we can say that in the unfolded state, the long-range interactions that keep a protein in a compact, globular folded state are absent.

Later, we will introduce a second model, M_L that concerns this unfolded state. In that case, the subscript L refers to the fact that only the local structure is considered. Specifically, the models express the following hypotheses:

- M_L specifies protein-like local structure but does not consider any long-range features (i.e., non-local structure). Under this hypothesis, the partition of compact, folded conformation has a low probability.
- M_N specifies protein-like local and non-local structure. Under this hypothesis, unfolded conformations have a low probability – at least for compact folded proteins.

Naturally, we want to obtain $p(\mathbf{x} \mid \mathbf{a}, M_N)$ in the form of a well-justified posterior distribution.

If we assume ideal bond angles and bond lengths, which is a reasonable approximation, a protein's structure can be entirely parameterized by a vector $\mathbf{x}$ of dihedral angles (see Figure 18.1). As each dihedral angle corresponds to one point on the unit circle, a protein can be fully parameterized as a sequence of such points (Boomsma et al. 2008; Harder et al. 2010).

18.4 Local and non-local structure

Protein structure can be conveniently understood as consisting of local and non-local features (Figure 18.2). With local structure, we refer to the shape of the protein on a local length scale. Typically, the local structure of a protein is classified into three types – α-helices, β-strands, and coils. With non-local structure, we refer to the contacts between amino acids that are far apart from each other in the sequence, but close together in space in the compact folded conformation. The distinction between local and non-local structure is somewhat artificial. However, it is a very useful concept for the development of methods to predict protein structure, which typically have components that deal with local structure and non-local structure. In the former case, fragment libraries – collections of local fragments excised from known protein structures (Simons et al. 1997) – are typically used. The latter case is typically covered by knowledge-based energy functions such as KPMFs (Sippl 1990).

The solution that we propose here also makes use of a divide-and-conquer approach that distinguishes local from non-local structure (Simons et al. 1997). The probability density that concerns local structure is only accurate on a local length scale and does not capture non-local features. On the other hand, it provides atomic detail, is efficient, and is easy to estimate. This density can be viewed as a prior distribution on the local shape of the protein. The probability density that concerns non-local structure provides information on

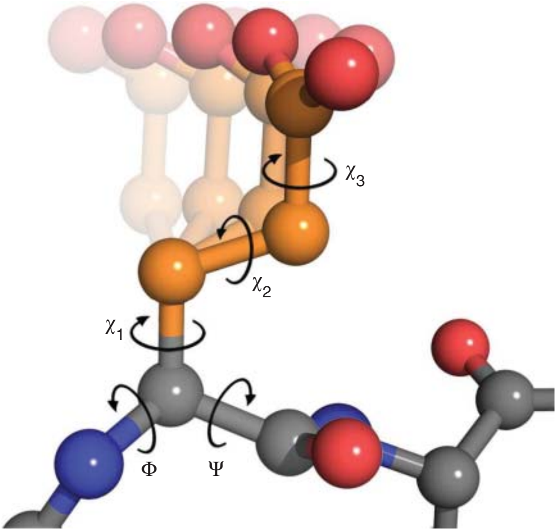


Figure 18.1 When bond angles and bond lengths are considered as fixed to their ideal values, a vector of dihedral angles is the remaining degree of freedom describing a three-dimensional protein structure. The dihedral angles can be subdivided into backbone and side chain angles, respectively involving (ψ, ϕ, ω) triplets and vectors of χ angles. All angles are illustrated in the figure, with the exception of ω, which is typically close to 180°. The number of χ angles varies between zero and four for the twenty standard amino acids. The figure shows a ball-and-stick representation of a single amino acid, glutamate, which has three χ angles, within a protein. The fading conformations in the background illustrate a rotation around χ_1. The figure was made using PyMOL (http://www.pymol.org, DeLano Scientific LCC) (adapted from Harder et al. (2010) http://www.biomedcentral.com/1471-2105/11/306. Used under CC-BY-SA 2.0 http://creativecommons.org/licenses/by/2.0/).

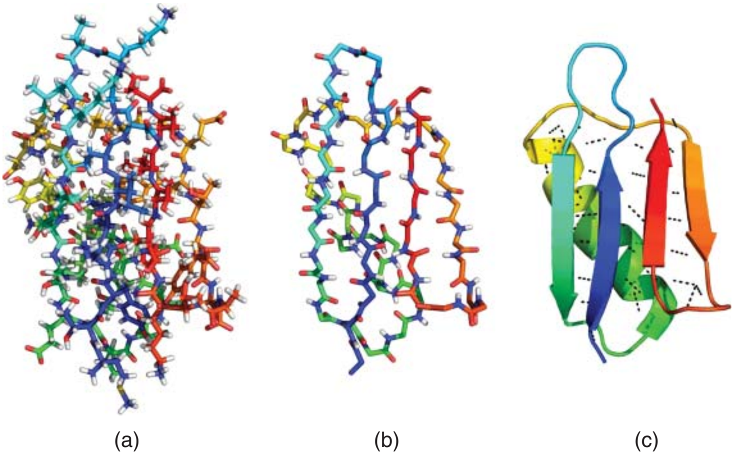


Figure 18.2 Three views of the same protein (protein G; Protein Data Bank code 2GB1). (a) A ball-and-stick representation of the protein, showing all bonds between atoms as sticks. Apart from the dynamics, this view includes essentially all relevant details. (b) Same view, but only showing the linear polymer part of the protein – the so-called main chain. The side chains are not shown in this view. (c) A schematic representation of the protein – called a "cartoon" – which shows an α-helix in the back, a β-sheet consisting of four β-strands (shown as arrows) and the interconnecting coils. The dotted lines show hydrogen bonds, which are some of the features that stabilize the folded conformation. The helices, strands, and coils can be considered "local" features, while the hydrogen bonds shown between the β-strands in the β-sheet can be considered "non-local" features, as they involve amino acids close in space, but relatively distant in sequence. This distinction between local and non-local is somewhat artificial, but can be used to great advantage in the formulation of probabilistic models of protein structure, as discussed in the chapter.

the interactions that are not captured by the local model. This density is *coarse grained*, which means that it involves a lower dimensional variable that does not capture atomic detail. In summary, we have

- One density that captures local, but not non-local, structure and provides atomic detail.
- A second density that captures non-local structure, but does not offer atomic detail.

In the following sections, we briefly describe how these densities are formulated and how they are combined into the desired final model.

18.5 The local model

Local structure concerns protein structure on a local length scale, including α-helices, β-strand, and coils. We will denote the "local" probability density as $p(\mathbf{x} \mid \mathbf{a}, M_L)$. This density is conditioned on the model M_L, which accurately captures local, but not non-local, structure. Conditioning on the imperfect model M_L allows the formulation of the density $p(\mathbf{x} \mid \mathbf{a}, M_L)$ that has very appealing properties, notably regarding estimation and computational efficiency. Conditioning on the desired model M_N, which correctly covers both local and non-local structure, will be added in a next step, as explained subsequently.

As the local model – TORUSDBN – has been described in great detail elsewhere (Boomsma et al. 2008, 2014; Hamelryck et al. 2006, 2012; Harder et al. 2010), we here give a high-level overview of the ideas behind it. To formulate a joint probability density of amino acid sequence and dihedral angles, we combined graphical models with directional statistics. The linear part of a protein – the so-called main chain – can be parameterized as a sequence of dihedral angles pairs. Thus, we used a dynamic Bayesian network consisting of a Markov chain of hidden nodes, to which nodes representing the amino acid symbols and dihedral angle pairs are attached (Boomsma et al. 2008, 2014). The dihedral angle pairs are modeled using a bivariate distribution on the torus – the bivariate von Mises distribution – which was especially developed for this purpose by Kanti Mardia and co-workers (Mardia et al. 2007). The dihedral angles of the so-called side chains – which can be considered as adornments of the main chain – are modeled using a similar approach (Harder et al. 2010).

The use of graphical models featuring Markov chains of hidden variables leads to probabilistic models that are computationally efficient and easy to estimate. However, they have one important shortcoming: a Markov chain has a finite memory along the sequence. Therefore, a Markov chain performs quite well on a local length scale, but cannot capture the many long-range interactions that are important features of the folded conformation of a protein. In other words, sampling from these models results in protein conformations that look like "unfolded" conformations. They are not compact, but locally they look like proteins, featuring α-helices, β-strands and coils, but – for example – not β-sheets, which are non-local features. These shortcomings can be alleviated by formulating a second probabilistic model that accurately covers non-local structure but provides less detail, and by combining the two models.

18.6 The non-local model

Nonlocal structure concerns protein features of a more global nature, including hydrogen bonds (see Figure 18.2(c)), amino acid packing in a hydrophobic core and so on. These interactions are not adequately captured by the local model. To model the non-local features of proteins, we introduce another variable, $\mathbf{y}$. This variable concerns the non-local structure of a protein and can be calculated from the vector of dihedral angles, $\mathbf{x}$,

$$\mathbf{y} = f(\mathbf{x}).$$

The dimensionality of $\mathbf{y}$ is typically much lower than the one of $\mathbf{x}$, and the relationship between $\mathbf{x}$ and $\mathbf{y}$ is many-to-one. We refer to the random variable $\mathbf{y}$ as a *coarse-grained variable*, while $\mathbf{x}$ is referred to as the *fine grained variable* (Borg et al. 2012; Frellsen et al. 2012; Hamelryck et al. 2010, 2013; Mardia et al. 2011; Mardia and Hamelryck 2012). For example, $\mathbf{y}$ could be a single positive, real value, describing the radius of the protein (Hamelryck et al. 2010).

We are interested in two probability densities concerning $\mathbf{y}$, namely $p(\mathbf{y} \mid \mathbf{a}, M_L)$ and $p(\mathbf{y} \mid \mathbf{a}, M_N)$. As we will explain in the next section, we need both densities to formulate our final joint model of protein structure. The first one is the probability distribution over $\mathbf{y}$ as implied by $p(\mathbf{x} \mid \mathbf{a}, M_L)$. The second density models the non-local structure of folded proteins – indicated by the conditioning on M_N – and can be inferred from the database of known proteins. Because $\mathbf{y}$ is a coarse-grained variable, it only provides limited – but accurate – information on the local structure.

For the coarse-grained variable $\mathbf{y}$, we recently proposed to use a low-dimensional vector of energy values that describe various aspects of the non-local structure of a protein, notably hydrogen bonding, hydrophobic interactions, and electrostatic interactions (Valentin et al. 2014). At this point, it is unclear how to optimally estimate this model, but a simple multivariate Gaussian model based on Bayesian linear regression delivered promising results. Another possibility is to infer amino acid contacts from evolutionary information (Marks et al. 2011) and to define $\mathbf{y}$ accordingly.

18.7 The formulation of the joint model

18.7.1 Outline of the problem and its solution

A direct, computationally efficient formulation of $p(\mathbf{x} \mid \mathbf{a}, M_N)$ is intractable. However, as we outlined in the previous two sections, the following probability distributions are available:

$$p(\mathbf{x} \mid \mathbf{a}, M_L), \tag{18.2}$$

$$p(\mathbf{y} \mid \mathbf{a}, M_L), \tag{18.3}$$

$$p(\mathbf{y} \mid \mathbf{a}, M_N), \tag{18.4}$$

where $\mathbf{y}$ is some deterministic function $\mathbf{y} = f(\mathbf{x})$ of $\mathbf{x}$, and $\dim(\mathbf{y}) < \dim(\mathbf{x})$. The second density is defined by

$$p(\mathbf{y} \mid \mathbf{a}, M_L) = \int_{\mathbf{x}: f(\mathbf{x}) = \mathbf{y}} p(\mathbf{x} \mid \mathbf{a}, M_L) d\mathbf{x}.$$

Conceptually, the first two probability densities concern protein structure on a local length scale in terms of the fine grained variable $\mathbf{x}$ and the coarse-grained variable $\mathbf{y}$; the third distribution concerns the non-local structure of actual proteins. Note that the latter distribution thus accurately covers local structure as well, but does not provide enough detail to parameterize it unequivocally. The question is now, how can $p(\mathbf{x} \mid \mathbf{a}, M_N)$ be obtained from the aforementioned three probability densities? The solution is given by

$$p(\mathbf{x} \mid \mathbf{a}, M_N) = \frac{p(\mathbf{y} \mid \mathbf{a}, M_N)}{p(\mathbf{y} \mid \mathbf{a}, M_L)} p(\mathbf{x} \mid \mathbf{a}, M_L).$$

We call this solution the reference ratio method (RR method), because it involves a factor, consisting of a ratio of two probability densities, that modifies $p(\mathbf{x} \mid \mathbf{a}, M_L)$. This solution can be derived in different ways: as the result of Bayesian reasoning, from combining the local and nonlocal models, starting from a conditional independence relationship, as a Jacobian factor resulting from a change of variables, and from marginalization over $\mathbf{y}$. Finally, the expression can also be seen as a special case of Jeffrey's conditioning or probability kinematics – a variant of Bayesian belief updating – and as a maximum entropy method. Protein structure prediction now amounts to sampling from the well-justified posterior distribution $p(\mathbf{x} \mid \mathbf{a}, M_N)$.

In the next sections, we give various derivations of the RR expression and also point out how this expression explains the success of KPMFs that are used ubiquitously in protein structure prediction (Koppensteiner and Sippl 1998; Sippl 1990; Sippl et al. 1996).

18.7.2 Model combination explanation

The outline of the problem and its solution is as follows. We have a probabilistic model $p(\mathbf{x} \mid \mathbf{a}, M_L)$ that covers protein structure on the local level. This model is tractable with respect to estimation, simulation, and computational efficiency. However, we want the probabilistic model $p(\mathbf{x} \mid \mathbf{a}, M_N)$, which covers protein structure on both local and nonlocal levels. The latter model is, however, intractable. The question is, how can the local model be "salvaged" by adding nonlocal information?

A first step to the solution is the introduction of the random variable $\mathbf{y} = f(\mathbf{x})$, – with $\dim(\mathbf{y}) < \dim(\mathbf{x})$ – which can be calculated deterministically from $\mathbf{x}$. For clarity, we will leave out the conditioning on the amino acid sequence $\mathbf{a}$ from now on. By involving $\mathbf{y}$, we can reformulate the local and nonlocal models in terms of a marginal and a conditional distribution as follows:

$$p(\mathbf{x} \mid M_L) = p(\mathbf{x} \mid \mathbf{y}, M_L)p(\mathbf{y} \mid M_L),$$
$$p(\mathbf{x} \mid M_N) = p(\mathbf{x} \mid \mathbf{y}, M_N)p(\mathbf{y} \mid M_N).$$

Conceptually, by choice, $\mathbf{y}$ should constitute a good descriptor of the nonlocal structure of a protein. In addition, by construction, $\mathbf{y}$ is chosen such that

$$p(\mathbf{x} \mid \mathbf{y}, M_L) = p(\mathbf{x} \mid \mathbf{y}, M_N).$$

Thus, the desired probability density is given by

$$p(\mathbf{x} \mid M_N) = p(\mathbf{x} \mid \mathbf{y}, M_L)p(\mathbf{y} \mid M_N),$$

provided that both the marginal and the conditional probability densities are available. In practice, $\mathbf{y}$ is chosen such that the marginal distribution $p(\mathbf{y} \mid M_N)$ can be easily obtained by estimating a probabilistic model from the set of known protein structures.

If the conditional distribution $p(\mathbf{x} \mid \mathbf{y}, M_L)$ of the local model were available, the problem would be solved at this point. However, only $p(\mathbf{x} \mid M_L)$ is available, not $p(\mathbf{x} \mid \mathbf{y}, M_L)$. A tractable solution is obtained by applying Bayes' theorem to the conditional as follows:

$$\begin{aligned} p(\mathbf{x} \mid M_N) &= p(\mathbf{x} \mid \mathbf{y}, M_L)p(\mathbf{y} \mid M_N) \\ &= \frac{p(\mathbf{y} \mid \mathbf{x}, M_L)p(\mathbf{x} \mid M_L)}{p(\mathbf{y} \mid M_L)}p(\mathbf{y} \mid M_N) \\ &= \frac{p(\mathbf{y} \mid M_N)}{p(\mathbf{y} \mid M_L)}p(\mathbf{x} \mid M_L). \end{aligned} \tag{18.5}$$

The factor $p(\mathbf{y} \mid \mathbf{x}, M_L)$ is equal to one as $\mathbf{y}$ is a deterministic function of $\mathbf{x}$ by construction.

The final issue that remains is how to obtain $p(\mathbf{y} \mid M_L)$. In contrast to $p(\mathbf{x} \mid \mathbf{y}, M_L)$, the marginal $p(\mathbf{y} \mid M_L)$ can be fairly easily obtained by

1. sampling a set of samples $\{\mathbf{x}_s\}$ from $p(\mathbf{x} \mid M_L)$,
2. calculating the set $\{\mathbf{y}_s\}$ from the sampled $\{\mathbf{x}_s\}$,
3. and estimating $p(\mathbf{y} \mid M_L)$ from the obtained $\{\mathbf{y}_s\}$.

The final solution given by Equation (18.5) involves a ratio of probability densities that modifies the local model $p(\mathbf{x} \mid M_L)$, hence the name "reference ratio method". We call $p(\mathbf{y} \mid M_L)$ the "reference distribution," for reasons explained in Section 18.9.

18.7.3 Conditional independence explanation

All the derivations we present here share one common feature: they all make use of some conditional independence assumption. In this derivation, this is where we start. Note that the presented independence assumptions are equivalent, as we will show in Section 18.7.6. Specifically, we assume that the conditional distribution of the local and nonlocal models is identical

$$p(\mathbf{x} \mid \mathbf{y}, M_N) = p(\mathbf{x} \mid \mathbf{y}, M_L).$$

This is a reasonable assumption for all choices of $\mathbf{y}$ that parameterize the non-local structure of a protein in adequate detail. The RRM is simply obtained by applying Bayes' rule to both sides of the equation,

$$p(\mathbf{x} \mid \mathbf{y}, M_N) = p(\mathbf{x} \mid \mathbf{y}, M_L) \tag{18.6}$$

$$\Rightarrow \frac{p(\mathbf{y} \mid \mathbf{x}, M_N)p(\mathbf{x} \mid M_N)}{p(\mathbf{y} \mid M_N)} = \frac{p(\mathbf{y} \mid \mathbf{x}, M_L)p(\mathbf{x} \mid M_L)}{p(\mathbf{y} \mid M_L)} \tag{18.7}$$

$$\Rightarrow p(\mathbf{x} \mid M_N) = \frac{p(\mathbf{y} \mid M_N)p(\mathbf{x} \mid M_L)}{p(\mathbf{y} \mid M_L)}. \tag{18.8}$$

Note that $p(\mathbf{y} \mid \mathbf{x}, M_N)$ and $p(\mathbf{y} \mid \mathbf{x}, M_L)$ both are equal to one and cancel because $\mathbf{y} = f(\mathbf{x})$.

18.7.4 Marginalization explanation

Next, we obtained the RRM by marginalization over the coarse-grained variable $\mathbf{y}$, making use of the same conditional independence assumption as in the previous section,

$$p(\mathbf{x} \mid \mathbf{y}, M_N) = p(\mathbf{x} \mid \mathbf{y}, M_L),$$

as follows:

$$p(\mathbf{x} \mid M_N) = \int_{\mathbf{y}'} p(\mathbf{x} \mid \mathbf{y}', M_N)p(\mathbf{y}' \mid M_N)d\mathbf{y}' \tag{18.9}$$

$$= \int_{\mathbf{y}'} p(\mathbf{x} \mid \mathbf{y}', M_L)p(\mathbf{y}' \mid M_N)d\mathbf{y}' \tag{18.10}$$

$$= \int_{\mathbf{y}'} \frac{p(\mathbf{y}' \mid \mathbf{x}, M_L)p(\mathbf{x} \mid M_L)}{p(\mathbf{y}' \mid M_L)} p(\mathbf{y}' \mid M_N)d\mathbf{y}' \tag{18.11}$$

$$= \frac{p(\mathbf{y} \mid M_N)}{p(\mathbf{y} \mid M_L)} p(\mathbf{x} \mid M_L). \tag{18.12}$$

In the last step, the integral disappears because $p(\mathbf{y} \mid \mathbf{x}, M_L)$ is zero if $\mathbf{y}' \neq \mathbf{y} = f(\mathbf{x})$, and one otherwise.

18.7.5 Jacobian explanation

The RR expression can also be derived making use of the Jacobian of a transformation of random variables. We assume that it is possible to augment $\mathbf{y}$ with $\mathbf{z}$, resulting in $\mathbf{v} = (\mathbf{y}, \mathbf{z})$ such that

- There is a one-to-one mapping between $\mathbf{x}$ and $\mathbf{v} = (\mathbf{y}, \mathbf{z})$.
- $\dim(\mathbf{x}) = \dim(\mathbf{y}) + \dim(\mathbf{z})$.

Following the rules regarding transformations of variables, we can write

$$p(\mathbf{x} \mid M_N) = p(\mathbf{y}, \mathbf{z} \mid M_N)\frac{d\mathbf{v}}{d\mathbf{x}} \tag{18.13}$$

$$p(\mathbf{x} \mid M_L) = p(\mathbf{y}, \mathbf{z} \mid M_L)\frac{d\mathbf{v}}{d\mathbf{x}}, \tag{18.14}$$

and thus

$$p(\mathbf{x} \mid M_N) = \frac{p(\mathbf{y}, \mathbf{z} \mid M_N)}{p(\mathbf{y}, \mathbf{z} \mid M_L)} p(\mathbf{x} \mid M_L).$$

After applying the product rule to both factors in the ratio, we obtain

$$p(\mathbf{x} \mid M_N) = \frac{p(\mathbf{z} \mid \mathbf{y}, M_N) p(\mathbf{y} \mid M_N)}{p(\mathbf{z} \mid \mathbf{y}, M_L) p(\mathbf{y} \mid M_L)} p(\mathbf{x} \mid M_L). \tag{18.15}$$

Next, we make the following assumption:

$$p(\mathbf{z} \mid \mathbf{y}, M_N) = p(\mathbf{z} \mid \mathbf{y}, M_L),$$

which reduces Equation (18.15) to the RR expression.

18.7.6 Equivalence of the independence assumptions

In the three derivations given earlier, we have used two seemingly different conditional independence assumptions, namely,

$$p(\mathbf{x} \mid \mathbf{y}, M_N) = p(\mathbf{x} \mid \mathbf{y}, M_L), \tag{18.16}$$

$$p(\mathbf{z} \mid \mathbf{y}, M_N) = p(\mathbf{z} \mid \mathbf{y}, M_L). \tag{18.17}$$

The equivalence of the assumptions is established by

$$p(\mathbf{x} \mid \mathbf{y}, M_N) = p(\mathbf{x} \mid \mathbf{y}, M_L) \tag{18.18}$$

$$\Rightarrow p(\mathbf{y}, \mathbf{z} \mid \mathbf{y}, M_N) = p(\mathbf{y}, \mathbf{z} \mid \mathbf{y}, M_L) \tag{18.19}$$

$$\Rightarrow p(\mathbf{z} \mid \mathbf{y}, M_N) = p(\mathbf{z} \mid \mathbf{y}, M_L), \tag{18.20}$$

which follows from the fact that there is a one-to-one transformation between $\mathbf{x}$ and $\mathbf{v} = (\mathbf{y}, \mathbf{z})$. It is trivial to show the inverse equivalence.

18.7.7 Probability kinematics explanation

Recently, it has also become clear that the RR method can be seen as an application of probability kinematics or Jeffrey's conditioning (Jeffrey 2004). Jeffrey's conditioning was proposed by the American philosopher of probability Richard Jeffrey (1926–2002).[1] Here is a simple illustration of Jeffrey's conditioning, as given by Diaconis and Zabell (1982).

Example 18.7.1: Whitworth's horses

Question: Three horses, A, B, and C, enter a race. Their initial probabilities to win are $\frac{2}{11}$, $\frac{4}{11}$, and $\frac{5}{11}$. We gain extra information, which changes A's probability to win to $\frac{1}{2}$. What are the corresponding probabilities in favor of B now?
Answer: The solution follows, if we assume that the new information on A does not affect the relative probabilities of B and C. The new information diminishes the probability that A loses by $\frac{11}{18}$. Hence, the probabilities of B and C winning are diminished by the same ratio. Thus, the probability of B winning is

$$p'(\text{B wins}) = p(\text{B wins} \mid \text{A loses})p'(\text{A loses}) = \frac{4}{9} \times \frac{1}{2} = \frac{2}{9},$$

where $p(\cdot)$ stands for the previous probabilities and $p'(\cdot)$ stands for the updated probabilities. □

In this section, we use a slightly different notation, following Diaconis and Zabell (Diaconis and Zabell 1982, 1986). We start out with a probability distribution $p(\mathbf{x})$ and a partition $\{E_1, E_2, \ldots, E_n\}$ of the space of $\mathbf{x}$. The partition is assumed to be mutually exclusive and exhaustive. Now, we are given new probabilities $p'(E_i)$ for all the elements of the partition. The question is now, how can we update the probability distribution $p(\mathbf{x})$? The answer is given by Jeffrey's rule of conditioning,

$$p'(\mathbf{x}) = \sum_{i=1}^{n} p(\mathbf{x} \mid E_i)p'(E_i),$$

where the sum runs over all elements of the partition. Since the partition is mutually exclusive and exhaustive, we can also write

$$p'(\mathbf{x}) = p(\mathbf{x} \mid E_{\mathbf{x}})p'(E_{\mathbf{x}}),$$

where $E_{\mathbf{x}}$ is the unique partition to which $\mathbf{x}$ belongs.

By casting the expression in a slightly different way, making use of Bayes' rule, we can easily obtain an expression that corresponds to the reference ratio distribution

$$p'(\mathbf{x}) = \frac{p(E_{\mathbf{x}} \mid \mathbf{x})p(\mathbf{x})}{p(E_{\mathbf{x}})}p'(E_{\mathbf{x}}) = \frac{p'(E_{\mathbf{x}})}{p(E_{\mathbf{x}})}p(\mathbf{x}).$$

The factor $p(E_{\mathbf{x}} \mid \mathbf{x})$ is equal to one, as $\mathbf{x}$ belongs to exactly one partition, $E_{\mathbf{x}}$. The RRM thus corresponds to Jeffrey's conditioning when the conditional probability distribution $p(\mathbf{x} \mid E_{\mathbf{x}})$ is not available or intractable, but one knows $p(E_{\mathbf{x}})$ and $p(\mathbf{x})$ instead.

[1] Not to be confused with Harold Jeffreys (1891–1989), the well-known pioneer of Bayesian statistics.

The valid application of Jeffrey's rule amounts to the assumption that

$$p'(\mathbf{x} \mid E_i) = p(\mathbf{x} \mid E_i),$$

for all i. The condition is called the "J-condition" by Diaconis and Zabell (1982; 1986). In short, the conditional probabilities given an element of the partition stay the same, but the probabilities of the elements themselves are changed. This condition is identical to the independence assumptions discussed in the previous section, notably

$$p(\mathbf{x} \mid \mathbf{y}, M_N) = p(\mathbf{x} \mid \mathbf{y}, M_L).$$

Thus, the introduction of $\mathbf{y}$ is a way to impose a partitioning on the conformational space of a protein. The local density, $p(\mathbf{x} \mid M_L)$ is incorrect in the sense that the probabilities of the partitions are wrong. In other words, $p(\mathbf{y} \mid M_L)$ is incorrect, but $p(\mathbf{x} \mid \mathbf{y}, M_L)$ is correct. The RRM thus corrects the probabilities of the elements of the partition, according to Jeffrey's rule.

18.7.8 Bayesian explanation

The RRM can also be obtained as a result of conventional Bayesian reasoning. In order to show this, we formulate the nonlocal model as the probability density $p(\mathbf{x} \mid N = 1, I)$, where N is a Boolean indicator variable that specifies a folded ($N = 1$) or unfolded ($N = 0$) conformation and I represents the general background knowledge. Conventional Bayesian updating based on the data N results in

$$\begin{aligned} p(\mathbf{x} \mid N, I) &= \frac{p(N \mid \mathbf{x}, I)p(\mathbf{x} \mid I)}{p(N \mid I)} \\ &\propto p(N \mid \mathbf{x}, I)p(\mathbf{x} \mid I). \end{aligned} \tag{18.21}$$

Now we assume that it is possible to chose a random variable $\mathbf{y}$, with $\mathbf{y} = f(\mathbf{x})$ and $\dim(\mathbf{y}) < \dim(\mathbf{x})$, such that

$$p(N \mid \mathbf{x}, I) = p(N \mid \mathbf{y}, I).$$

In addition, $\mathbf{y}$ is assumed to be a good descriptor of a protein's nonlocal structure. Next, we reformulate $p(N \mid \mathbf{y}, I)$ by applying Bayes' theorem, resulting in

$$\begin{aligned} p(N \mid \mathbf{x}, I) &= \frac{p(\mathbf{y} \mid N, I)p(N \mid I)}{p(\mathbf{y} \mid I)} \\ &\propto \frac{p(\mathbf{y} \mid N, I)}{p(\mathbf{y} \mid I)}. \end{aligned} \tag{18.22}$$

Substituting Equation (18.22) into Equation (18.21) results in the expression

$$p(\mathbf{x} \mid N, I) \propto \frac{p(\mathbf{y} \mid N, I)}{p(\mathbf{y} \mid I)} p(\mathbf{x} \mid I),$$

which is equivalent to the RRM if we choose $p(\mathbf{x} \mid M_L)$ for $p(\mathbf{x} \mid I)$. That is, the background knowledge I represents what is known about local protein structure. Thus, it can be seen that the RR expression can be obtained as a result of Bayesian updating of the local model $p(\mathbf{x} \mid M_L)$ in the light of nonlocal information. The modifying ratio in front of $p(\mathbf{x} \mid M_L)$ in the RR expression can be interpreted as a likelihood. Hence, the distribution obtained from the RR expression is a valid posterior distribution from this point of view.

18.8 Kullback–Leibler optimality

The RR method can be interpreted as a maximum entropy modification of $p(\mathbf{x} \mid M_L)$ such that the correct distribution over $\mathbf{y} = f(\mathbf{x})$ is attained. More precisely, the RR expression represents the minimal modification of $p(\mathbf{x} \mid M_L)$ in terms of the Kullback–Leibler (KL) divergence to fulfill the requirement with respect to the marginal distribution of $\mathbf{y}$, which is $p(\mathbf{y} \mid M_N)$.

Consider the set $\mathcal{D}$ of all densities on the space of $\mathbf{x}$ that imply the correct distribution over $\mathbf{y} = f(\mathbf{x})$, that is,

$$\mathcal{D} = \left\{ h(\mathbf{x}) \mid \forall \mathbf{y}' : \int_{\mathbf{x}: f(\mathbf{x}) = \mathbf{y}'} h(\mathbf{x}) \mathbf{d}\mathbf{x} = p(\mathbf{y}' \mid M_N) \right\},$$

where we have used the notation $h(\mathbf{x}) = h(\mathbf{x} \mid \mathbf{y}) h(\mathbf{y})$. Now, we are looking for $\hat{h}(\mathbf{x}) \in \mathcal{D}$ with the minimal KL divergence from $p(\mathbf{x} \mid M_L)$. Using the definition of the KL divergence, and leaving out the conditioning on $\mathbf{a}$ and M_L below the first line for clarity, we have

$$\hat{h}(\mathbf{x}) = \arg \min_{h(\mathbf{x}) \in D} \mathrm{KL}\left[p(\mathbf{x} \mid M_L) \parallel h(\mathbf{x}) \right] \tag{18.23}$$

$$= \arg \min_{h(\mathbf{x}) \in D} \int_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{h(\mathbf{x})} d\mathbf{x} \tag{18.24}$$

$$= \arg \min_{h(\mathbf{x}) \in D} \int_{\mathbf{y}} \int_{\mathbf{x}: f(\mathbf{x}) = \mathbf{y}} \left[p(\mathbf{y}) p(\mathbf{x} \mid \mathbf{y}) \left\{ \log \frac{p(\mathbf{y})}{h(\mathbf{y})} + \log \frac{p(\mathbf{x} \mid \mathbf{y})}{h(\mathbf{x} \mid \mathbf{y})} \right\} \right] d\mathbf{x}\, d\mathbf{y} \tag{18.25}$$

$$= \arg \min_{h(\mathbf{x}) \in D} \int_{\mathbf{y}} p(\mathbf{y}) \log \frac{p(\mathbf{y})}{h(\mathbf{y})} \mathbf{d}\mathbf{y} + \int_{\mathbf{y}} p(\mathbf{y}) \int_{\mathbf{x}: f(\mathbf{x}) = \mathbf{y}} p(\mathbf{x} \mid \mathbf{y}) \log \frac{p(\mathbf{x} \mid \mathbf{y})}{h(\mathbf{x} \mid \mathbf{y})} d\mathbf{x}\, d\mathbf{y} \tag{18.26}$$

$$= \arg \min_{h(\mathbf{x}) \in D} \int_{\mathbf{y}} p(\mathbf{y}) \log \frac{p(\mathbf{y})}{p(\mathbf{y} \mid M_N)} \mathbf{d}\mathbf{y} + \int_{\mathbf{y}} p(\mathbf{y}) \int_{\mathbf{x}: f(\mathbf{x}) = \mathbf{y}} p(\mathbf{x} \mid \mathbf{y}) \log \frac{p(\mathbf{x} \mid \mathbf{y})}{h(\mathbf{x} \mid \mathbf{y})} d\mathbf{x}\, d\mathbf{y}, \tag{18.27}$$

where we have used $h(\mathbf{y}) = p(\mathbf{y} \mid M_N)$ in the last step. The first term is constant. The second term is non-negative according to Jensen's inequality and reaches the minimal value of zero when $h(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{x} \mid \mathbf{y}, M_L) = \frac{p(\mathbf{x} \mid M_L)}{p(\mathbf{y} \mid M_L)}$. Thus, we indeed obtain the RR, given that

$$h(\mathbf{x}) = h(\mathbf{x} \mid \mathbf{y}) h(\mathbf{y}) \tag{18.28}$$

$$= \frac{p(\mathbf{x} \mid M_L)}{p(\mathbf{y} \mid M_L)} p(\mathbf{y} \mid M_N). \tag{18.29}$$

The KL optimality of the RR method – and of Jeffrey's conditioning, for that matter (Diaconis and Zabell 1982; Frellsen et al. 2012) – is an attractive property and provides a clear bridge to physics, where maximum entropy methods are widely used (Jaynes 1957, 1978). The main difference with most maximum entropy applications is that we impose a probability distribution over a many-to-one function of $\mathbf{x}$, rather than adjusting the mean of $\mathbf{x}$ to a new value.

18.9 Link with statistical potentials

The RR method has in fact been used for over twenty years for protein structure prediction, but in an *ad hoc* fashion and without understanding as to why the method works (Borg et al. 2012; Hamelryck et al. 2010). The workhorse for dealing with non-local interactions in protein structure prediction are KPMFs. Typically, these potentials assign an energy to a protein structure based on the pairwise distances between the amino acids (Koppensteiner and Sippl 1998; Sippl 1990). This energy is equal to

$$E_T = \sum_{i,j} E(d_{i,j}, a_i, a_j) - E_R(d_{i,j}, a_i, a_j),$$

where E_T is the total energy, the sum runs over all relevant atom pairs (i, j), $d_{i,j}$ is the distance between amino acids i and j and (a_i, a_j) are the amino acid types. The energies E and E_R correspond to an energy function derived from the known folded proteins and a so-called "reference" energy, which is supposed to model unfolded proteins. These KPMFs are justified by analogy with true potentials of mean force, as used in the physics of liquids (Koppensteiner and Sippl 1998). However, vague analogies are not a substitute for true understanding, and as a result, KPMFs are currently not used to their full potential.

If we reformulate KPMFs in a probabilistic way, by turning energies into probabilities using Boltzmann's law, we obtain the RR expression for $p(\mathbf{x} \mid \mathbf{a})$ uniform. In this case, $\mathbf{y}$ is a vector of pairwise distances calculated from the protein structure $\mathbf{x}$, and it is (incorrectly) assumed that the distances are conditionally independent given the amino acid sequence $\mathbf{a}$, resulting in

$$p(\mathbf{x} \mid \mathbf{a}, M_N) = \frac{p(\mathbf{y} \mid \mathbf{a}, M_N)}{p(\mathbf{y} \mid \mathbf{a}, M_L)} p(\mathbf{x} \mid \mathbf{a}, M_L) \tag{18.30}$$

$$= \frac{\prod_{i,j} p(d_{i,j} \mid a_i, a_j, M_N)}{\prod_{i,j} p(d_{i,j} \mid a_i, a_j, M_L)} p(\mathbf{x} \mid \mathbf{a}, M_L). \tag{18.31}$$

In most KPMF applications, the uniform assumption for $p(\mathbf{x} \mid \mathbf{a}, M_L)$ is incorrect, however. Typically, a non-uniform $p(\mathbf{x} \mid \mathbf{a}, M_L)$ is brought in by sampling from a fragment library. Thus, the fragment library – or any other method that is used for the conformational sampling – will unequivocally determine the denominator in the RR expression, corresponding to the reference energy.

After two decades of successful applications in protein structure prediction and simulation (Sippl 1990) despite much debate about their physical validity (Ben-Naim 1997; Thomas and Dill 1996), KPMFs are now finally explained as statistically well-defined quantities (Borg et al. 2012; Frellsen et al. 2012; Hamelryck et al. 2010). The new insights about the nature of KPMFs are not only of theoretical interest; they readily translate into improved energy functions. Notably, current KPMFs can be improved in two respects:

1. KPMFs are not limited to pairwise distances, but can be extended to any coarse-grained descriptor of protein structure (Hamelryck et al. 2010).
2. The local model used for conformational sampling determines the reference energy. As this fact remains largely unknown, the reference energy is now typically constructed based on *ad hoc* arguments (Borg et al. 2012; Frellsen et al. 2012).

18.10 Conclusions and outlook

In this chapter, we have outlined a tractable, computationally efficient, and well-justified Bayesian model of protein structure, which can be used for inference of protein structure from sequence (Boomsma et al. 2008; Hamelryck et al. 2010; Valentin et al. 2014). The key idea behind the model is to formulate complementary models – some effective on a local scale and others on a non-local scale – and to tie them together in an unusual, but well-defined, Bayesian way.

The local model is computationally efficient and detailed, but only valid on a short-range scale (Boomsma et al. 2008). This is because it is based on a graphical model that is essentially a Markov model. On the other hand, the global model lacks detail but provides long-range information (Valentin et al. 2014). Glueing the two models together results in a joint model that represents the best of both worlds: the model is detailed and computationally efficient, and valid on both local and global scales (Valentin et al. 2014).

The way in which the two models are combined is quite interesting. It involves modifying the local model with a factor – a ratio of two densities concerning the non-local structure – that brings in the global information. The method can be understood as resulting from updating the local model with non-local information and can be derived in various ways. Notably, the method can be seen as an example of Jeffrey's updating or probability kinematics – a specific way of belief updating first proposed by the philosopher of probability Richard Jeffrey (Jeffrey 2004). The method can also be interpreted as a maximum entropy method and as resulting from conventional Bayesian updating.

This is the first time that a well-defined Bayesian model of protein structure in atomic detail is formulated. We believe that it has great potential and that it might provide a new impetus to the field of protein structure prediction.

Developing a sound probabilistic model of something as complicated as a protein structure, consisting of thousands of atoms, poses a fascinating statistical challenge. It involves large amounts of data, inference on unusual manifold such as hypertori and hyperspheres, (Boomsma et al. 2008; Hamelryck et al. 2006), and stringent demands on computational efficiency. The method we developed to tackle this challenge adopts a unique strategy that can be applied to a wide range of unrelated problems. Many statistical problems are of a multi-scale nature – they involve the modeling of phenomena on local and global scales. A tractable strategy can be to develop local and global models – each of them mainly covering essentially one scale – and to combine them with Jeffrey's conditioning into a model that covers all scales.

In short, the strategy consists of the following three aspects:

- Develop a detailed probabilistic model $p(\mathbf{x} \mid M_L)$ that accurately covers the short-range scale of the problem, but not the long-range scale. Focusing solely on the short-range scale often makes it possible to formulate computationally efficient, yet detailed models, for example, by making use of Markov models, at the expense of modeling long-range features.

- Develop a second model $p(\mathbf{y} \mid M_N)$ that covers long-range aspects of the problem. By leaving out the short-range details, it is often possible to develop a computationally efficient and adequate model of the long-range features. Specifically, the first model concerns a detailed feature vector $\mathbf{x}$, while the second model concerns a coarse-grained feature vector that is a deterministic function of $\mathbf{x}$, that is, $\mathbf{y} = f(\mathbf{x})$.
- We now have two models that cover different ranges of the problem. One model concerns the short-range scale and provides detail, but does not cover the long-range scale. The other model covers the long-range scale but does not cover details. By combining the models using a variant of Jeffrey's conditioning, one obtains a final model, in the form of a posterior distribution

$$p(\mathbf{x} \mid M_N) = \frac{p(\mathbf{y} \mid M_N)}{p(\mathbf{y} \mid M_L)} p(\mathbf{x} \mid M_L),$$

  that covers both short- and long-range scales, provides detail on all scales and yet remains computationally efficient.
- This approach can be readily extended beyond two models, in order to cover multiple scales, which is reminiscent of strategies adopted in deep learning (Bengio 2009).

The RR method is an excellent example of how tackling a challenging, real-life problem can lead to exciting new statistical methods and concepts that can potentially be widely applied. It is quite surprising that a fundamental and potentially widely applicable concept such as Jeffrey's conditioning is so little known by the statistical and machine learning communities. Previously, Ferreira and co-workers proposed to use Jeffrey's conditioning to create multi-scale random field models (Ferreira and Lee 2007) . It has been suggested that the brain approximates Bayesian methods in its computations (Friston 2010). Perhaps the brain also makes use of the multi-scale modeling possibilities offered by probability kinematics? In any case, we look forward to see this elegant method pop up in other applications and contexts.

Acknowledgments

We thank Daniel M. Roy for pointing out the similarity between Jeffrey's rule and the reference ratio method, and John T. Kent and Kanti V. Mardia for clarifying discussions on the reference ratio method. JF was supported by an individual postdoc grant from the Danish Council for Independent Research | Natural Sciences. WB was supported by the Danish Council for Independent Research and the Villum Foundation. TH acknowledges support from the UCHP 2016 Excellence Programme for Interdisciplinary Research (UCPH2016-DSIN)

References

Anfinsen CB 1973 Principles that govern the folding of protein chains. *Science* **181**(96), 223–230.

Bengio Y 2009 Learning deep architectures for AI. *Foundations and Trends in Machine Learning* **2**, 1–127.

Ben-Naim A 1997 Statistical potentials extracted from protein structures: are these meaningful potentials? *Journal of Chemical Physics* **107**, 3698–3706.

Boomsma W, Mardia K, Taylor C, Ferkinghoff-Borg J, Krogh A and Hamelryck T 2008 A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences of the United States of America*, USA **105**, 8932–8937.

Boomsma W, Tian P, Ferkinghoff-Borg J, Hamelryck T, Lindorff-Larsen K and Vendruscolo M 2014 Equilibrium simulations of proteins using molecular fragment replacement and NMR chemical shifts. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 13852–13857.

Borg M, Hamelryck T and Ferkinghoff-Borg J 2012 On the physical relevance and statistical interpretation of knowledge-based potentials In *Bayesian methods in structural bioinformatics* Hamelryck T, Mardia K and Ferkinghoff-Borg J). Springer-Verlag, Heidelberg, Berlin, pp. 97–124.

Bradley P, Misura KM and Baker D 2005 Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871.

Chandler D 1987 *Introduction to Modern Statistical Mechanics*. Oxford University Press, USA.

Diaconis P and Zabell S 1982 Updating subjective probability. *Journal of the American Statistical Association* **77**, 822–830.

Diaconis P and Zabell S 1986 Some alternatives to Bayes's rule In *Proceedings of the 2nd University of California, Irvine, Conference on Political Economy*, pp. 25–38.

Dill KA 1999 Polymer principles and protein folding. *Protein Science* **8**, 1166–1180.

Dill K and Chan H 1997 From Levinthal to pathways to funnels. *Nature Structural Biology* **4**, 10–19.

Dill K and MacCallum J 2012 The protein-folding problem, 50 years on. *Science* **338**, 1042–1046.

Duan Y and Kollman P 1998 Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744.

Faver J, Benson M, He X, Roberts B, Wang B, Marshall M, Sherrill C and Merz K 2011 The energy computation paradox and ab initio protein folding. *PLoS ONE* **6**, e18868.

Ferreira M and Lee H 2007 *Multiscale Modeling: a Bayesian Perspective*. Springer-Verlag.

Frellsen J, Mardia K, Borg M, Ferkinghoff-Borg J and Hamelryck T 2012 Towards a general probabilistic model of protein structure: the reference ratio method In *Bayesian Methods in Structural Bioinformatics* Hamelryck T, Mardia K and Ferkinghoff-Borg J (eds). Springer-Verlag, Heidelberg, Berlin, pp. 125–134.

Friston K 2010 The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* **11**, 127–138.

Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frellsen J, Andreetta C, Boomsma W, Bottaro S and Ferkinghoff-Borg J 2010 Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS ONE* **5**, e13714.

Hamelryck T, Haslett J, Mardia K, Kent J, Valentin J, Frellsen J and Ferkinghoff-Borg J 2013 On the reference ratio method and its application to statistical protein structure prediction In *LASR2013 - Statistical Models and Methods for Non-Euclidean Data with Current Scientific Applications* Mardia K, Gusnanto A, Riley AD and Voss J (eds). Leeds University Press, Leeds, UK, pp. 53–57.

Hamelryck T, Kent J and Krogh A 2006 Sampling realistic protein conformations using local structural bias. *PLoS Computational Biology* **2**(9), e131.

Hamelryck T, Mardia K and Ferkinghoff-Borg J (eds) 2012 *Bayesian Methods in Structural Bioinformatics*, *Statistics for Biology and Health*. Springer-Verlag, Heidelberg, Berlin.

Harder T, Boomsma W, Paluszewski M, Frellsen J, Johansson K and Hamelryck T 2010 Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinformatics* **11**, 306.

Harder T, Borg M, Bottaro S, Boomsma W, Olsson S, Ferkinghoff-Borg J and Hamelryck T 2012 An efficient null model for conformational fluctuations in proteins. *Structure* **20**, 1028–1039.

Jaynes E 1957 Information theory and statistical mechanics. *Physical Review* **106**, 620–630.

Jaynes E 1978 Where do we stand on maximum entropy? In *The Maximum Entropy Formalism Conference* Levine RD and Tribus M). MIT Press, Cambridge, MA, pp. 15–118.

Jeffrey R 2004 *Subjective Probability: The Real Thing*. Cambridge University Press.

Koppensteiner W and Sippl M 1998 Knowledge-based potentials–back to the roots. *Biochemistry (Mosc)* **63**, 247–52.

Kryshtafovych A, Fidelis K and Moult J 2014 CASP10 results compared to those of previous CASP experiments. *Proteins* **82**, 164–174.

Lindorff-Larsen K, Piana S, Dror R and Shaw D 2011 How fast-folding proteins fold. *Science* **334**, 517.

Lipfert J and Doniach S 2007 Small-angle X-ray scattering from RNA, proteins, and protein complexes. *Annual Review of Biophysics and Biomolecular Structure* **36**, 307–327.

Mardia K, Frellsen J, Borg M, Ferkinghoff-Borg J and Hamelryck 2011 A statistical view on the reference ratio method. In *LASR2011 - High-Throughput Sequencing, Proteins and Statistics* Gusnanto A, Mardia K and Fallaize C (eds), Leeds University Press, Leeds, pp. 55–61.

Mardia K and Hamelryck T 2012 Discussion to 'Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation' by P. Fearnhead and D. Prangle. *Journal of the Royal Statistical Society, Series B* **74**, 462–463.

Mardia KV and Jupp P 2000 *Directional Statistics*, 2nd ed. John Wiley and Sons, Ltd.

Mardia K, Taylor C and Subramaniam G 2007 Bivariate von Mises densities for angular data with applications to protein bioinformatics. *Biometrics* **63**, 505–512.

Marks D, Colwell L, Sheridan R, Hopf T, Pagnani A, Zecchina R and Sander C 2011 Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**(12), e28766 EP–.

Olsson S, Frellsen J, Boomsma W, Mardia K and Hamelryck T 2013 Inference of structure ensembles of flexible biomolecules from sparse, averaged data. *PLoS ONE* **8**(11), e79439.

Olsson S, Vögeli B, Cavalli A, Boomsma W, Ferkinghoff-Borg J, Lindorff-Larsen K and Hamelryck T 2014 Probabilistic determination of native state ensembles of proteins. *Journal of Chemical Theory and Computation* **10**, 3484–3491.

Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C *et al.* 2013 Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**, 281–286.

Pohl FM 1971 Empirical protein energy maps. *Nature New Biology* **234**, 277–279.

Przytycka T 2004 Significance of conformational biases in Monte Carlo simulations of protein folding: lessons from Metropolis–Hastings approach. *Proteins: Structure, Function, and Bioinformatics* **57**, 338–344.

Roy A, Kucukural A and Zhang Y 2010 I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols* **5**, 725–738.

Service R 2008 Problem solved (sort of). *Science* **321**, 784–786.

Simons KT, Kooperberg C, Huang E and Baker D 1997 Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology* **268**, 209–225.

Sippl MJ 1990 Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology* **213**, 859–883.

Sippl MJ, Ortner M, Jaritz M, Lackner P and Flockner H 1996 Helmholtz free energies of atom pair interactions in proteins. *Folding and Design* **1**, 289–98.

Thomas PD and Dill KA 1996 Statistical potentials extracted from protein structures: how accurate are they? *Journal of Molecular Biology* **257**, 457–469.

Tompa P 2002 Intrinsically unstructured proteins. *Trends in Biochemical Sciences* **27**, 527–533.

Valentin J, Andreetta C, Boomsma W, Bottaro S, Ferkinghoff-Borg J, Frellsen J, Mardia K, Tian P and Hamelryck T 2014 Formulation of probabilistic models of protein structure in atomic detail using the reference ratio method. *Proteins: Structure, Function, and Bioinformatics* **82**, 288–299.

19

MAD-Bayes matching and alignment for labelled and unlabelled configurations

Peter J. Green[1,2]

[1] *School of Mathematics, University of Bristol, Bristol, UK*

[2] *University of Technology, Sydney, New South Wales, Australia*

19.1 Introduction

When Professor Mardia presented a seminar on protein structural bioinformatics in Bristol in February 2003, I was fascinated by one of the problems he described, about matching and alignment, impressed by his visual aids (the things you could do with overlaid acetates!), but rather unsatisfied by the inferential approach he took. The basic problem (which is properly introduced subsequently) involves two key unknown quantities – the matching between unspecified subsets of two data clouds and the geometrical transformations the clouds had each been subjected to – and it seemed to me essential to treat these two things simultaneously, not sequentially: if that is accepted, then it is natural to use a Bayesian treatment. I think that I said something to this effect in discussion and followed it up later with a proposed model framework, which Professor Mardia and I investigated, with the results eventually becoming a *Biometrika* paper, Green and Mardia (2006).

Some subsequent developments of this idea appear in Mardia et al. (2007) (using the formal Bayesian fitting algorithm as a numerical technique for refining a non-Bayesian

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

solution), Ruffieux and Green (2009) (extending the idea to alignment of multiple configurations), Green et al. (2010) (largely a review article, but describing broader classes of biomolecular matching and alignment problems, and anticipating extensions to the modelling) and Fallaize et al. (2014) (employing a 'gap prior' to use sequence information when it is available).

I have also enjoyed robust, but friendly, conversations about the approach with both of the Editors of this volume, each of whom has also made significant contributions to understanding and addressing the problem, including Kent et al. (2004) and Kenobi and Dryden (2012).

This paper revisits inference based on the models such as those in Green and Mardia (2006) and Fallaize et al. (2014), using MAD-Bayes, a new perspective on fast approximate inference due to Broderick et al. (2013). This view might help to reconcile rival paradigms applied to this problem: it turns out to nicely bridge the gap between Bayesian and optimisation approaches to inferring matching and alignment.

19.2 Modelling protein matching and alignment

A mathematical abstraction of a certain problem in protein alignment involves a form of unlabelled shape analysis: we observe two point configurations $x = \{x_j : j = 1, 2, \ldots, m\}$ and $y = \{y_k : k = 1, 2, \ldots, n\}$ in $\mathcal{R}^d$ (typically $d = 2$ or 3); unknown subsets of each configuration are assumed to be matched, apart from noise, but the two configurations have been subject to different unknown geometrical transformations. These transformations are assumed to lie in prescribed families, for example, translations, rotations, rigid-body or affine transformations, or perhaps there has been some nonlinear warping. The problem is to make simultaneous inference about the alignment and the (relative) transformations. In turn, this abstraction can be set up in various ways: to preserve symmetry in the treatment of x and y, Green and Mardia (2006) supposed both configurations to be transformed from some latent configuration in another space, after being subject to both thinning and the addition of noise.

For the case of affine transformations, Green and Mardia (2006) assumed that the x configuration lies in the same d-dimensional space as the latent points, while the y configuration needs transforming to $Ay + \tau$ to lie in this space. The noise is assumed zero-mean spherical Gaussian with variance σ^2, independently for each point. The alignment between the configurations is represented by the binary (0/1) matrix M, where $M_{jk} = 1$ if and only if x_j and y_k are matched. Each point can be matched at most once, so there is at most one non-zero entry in each row and each column of M. We will write $\{j \underset{M}{\sim} k\}$ for the set of (j, k) pairs matched according to M, that is, $\{(j, k) : M_{jk} = 1\}$.

In Green and Mardia (2006), a stochastic model for point configurations and their alignment is derived, leading to a posterior distribution of the form

$$p(M, A, \tau | \sigma, x, y) \propto |A|^n p(A) p(\tau) \prod_{j \underset{M}{\sim} k} \left(\frac{\rho \phi\{(x_j - Ay_k - \tau)/\sigma\sqrt{2}\}}{\lambda(\sigma\sqrt{2})^d} \right) \tag{19.1}$$

over the unknown parameters A, τ and M, assuming here that σ is fixed, where ϕ is the standard normal density.

In the modelling, the distribution of the alignment M arises indirectly through a thinned-hidden-point formulation, and the induced prior for M has the form

$$p(M) \propto \left(\frac{\rho}{\lambda v}\right)^L, \tag{19.2}$$

where $L = \sum_{jk} M_{jk}$ is the number of matches. It follows that all feasible alignment matrices M with the same value for L have the same prior probability: $M|L$ is uniformly distributed. Of course, the number of different M with the same value of L varies greatly with the value of L – in fact it is $m!n!/[L!(m-L)!(n-L)!]$ (Green and Mardia 2006).

Expressions similar to (19.1) can arise from other underlying formulations by other authors, perhaps with $\sigma\sqrt{2}$ replaced by σ, and perhaps with $\rho/\lambda v$ expressed as a single parameter.

Green and Mardia (2006) build a methodology using the posterior distribution (19.1), concentrating primarily on the case of a rigid-body motion in 2- or 3-dimensions, where A is a rotation matrix, modelled a priori by a matrix Fisher distribution. Posterior sampling can be accomplished with a relatively straightforward Markov chain Monte Carlo (MCMC) sampler. This uses Gibbs updates for σ^2 and τ, Metropolis–Hastings updates for M (in which addition, deletion, or switching of matches are proposed), and, in the 3-D case, a novel Metropolis sampler for the matrix Fisher distribution for updating A.

For Bayesian point estimation of the alignment, we can take a decision theory approach based on a loss function that is additive over (j, k) pairs and exchangeable with respect to indexing. This turns out to require only the pairwise posterior match probabilities $P\{M_{jk} = 1|x, y\}$, which are readily estimated by direct enumeration from an MCMC sample. The resulting optimisation computation is equivalent to a mathematical programming assignment problem, and standard methods can be used to solve it.

These methodologies were illustrated by application to alignment of 2-D protein gels and of 3-D configurations of active sites. The MCMC methodology is in principle vulnerable to mixing problems caused by multi-modality in the posterior distribution, although such problems are not apparent in the examples shown.

19.3 Gap priors and related models

When sequence information is available, it is appealing to consider using it, and an attractive approach is to use a ‘gap prior’ of the form

$$p(M) \propto \exp(-U(M))$$

using the so-called gap penalty $U(M)$ given by

$$U(M) = gS(M) + h\sum_{i=1}^{S(M)}(l_i - 1), \tag{19.3}$$

where $S(M)$ is the number of instances where a new gap in the alignment is opened, l_i is the length of the ith gap, and g, h are positive hyperparameters, with commonly, $g > h$. See Rodriguez and Schmidler (2010) and Fallaize et al. (2014). Informally, the effect of using this prior with $g > h > 0$ compared to $g = h = 0$ is that among alignments with the same likelihood, preference is given to those where consecutively numbered atoms are matched, and where this fails, preference goes to those where the unmatched atoms are consecutive.

Using this prior in place of that used by Green and Mardia (2006), with other modelling details unchanged, leads to the posterior

$$p(M, A, \tau|\sigma, x, y) \propto |A|^n p(A) p(\tau) v^L \exp(-U(M)) \prod_{j \underset{M}{\sim} k} \left(\frac{\phi\{(x_j - Ay_k - \tau)/\sigma\sqrt{2}\}}{(\sigma\sqrt{2})^d} \right). \tag{19.4}$$

Although the gap penalty is commonly expressed in the form (19.3), this form is arguably ambiguous, and it can be helpful to express it more explicitly (Fallaize et al. 2014). Let M be a binary $m \times n$ matrix with L 1s, located in entries $(j_i, k_i), i = 1, 2, \ldots, L$, where the js and ks are consistently ordered: $j_1 < j_2 < \cdots < j_L$ and $k_1 < k_2 < \cdots < k_L$. This represents, of course, the matching of x_{j_i} and y_{k_i}, for $i = 1, 2, \ldots, L$. Then, the gap penalty can be written

$$U(M) = \sum_{i=1}^{L+1} [f(j_i - j_{i-1}) + f(k_i - k_{i-1})], \tag{19.5}$$

where $f(1) = 0$ and for $r \geq 2$, $f(r) = g + (r-2)h$. Here, we write $j_0 = k_0 = 0$ and $j_{L+1} = m + 1$, $k_{L+1} = n + 1$. We take $U(M) = +\infty$ if the js and ks cannot be consistently ordered, that is, if the alignment M is inconsistent with sequence ordering; such M have zero prior probability under this model.

In Fallaize et al. (2014), the MCMC algorithm of Green and Mardia (2006) is adapted to sampling for the posterior distribution for the gap prior model. The resulting algorithm relies upon proposing stepwise updates to M corresponding to adding, removing or switching a match. These are particularly easy to implement for the gap prior. If we insert a new match $(j^\star, k^\star)$ between (j_i, k_i) and (j_{i+1}, k_{i+1}), then the reduction in total gap penalty is the sum of two terms, one from the js and one from the ks. The term from the js is equal to

$$\begin{cases} g & \text{if } j_{i+1} - j_i = 2, \\ h & \text{if } j_{i+1} - j_i > 2 \text{ and } j^\star = j_i + 1 \text{ or } j_{i+1} - 1, \text{ and} \\ 2h - g & \text{otherwise.} \end{cases}$$

These three possibilities correspond to filling (and so eliminating) a gap, shortening a gap, or splitting a gap into two. The term from the ks has the same form.

A feature of this gap model that some might feel unappealing intuitively is that, conditional on the number of matches and the number of gaps, the indices of the x and y points forming those matches are a priori independent. In fact, the penalty $U(M)$, and hence the probability $p(M)$, depends only on L and S, where S is the total number of gaps in the two sequences combined; S is the number of blocks of consecutive all-zero rows or columns in M. To be explicit,

$$U(M) = (g - h)S + h(m + n - 2L). \tag{19.6}$$

Thus, for example, if there are three matches and the x indices are $(4, 5, 9)$, then under this model the y indices $(7, 8, 12)$ are exactly as probable as $(7, 11, 12)$. Indeed, if $m = 9$ and $n = 15$, this probability is also the same as that the x indices $(1, 2, 3)$ match y indices

$(2, j, 14)$, for any $j = 4, 5, \ldots, 12$ as all of these situations give $L = 3$ and $S = 5$. The penalty (19.3) should, therefore, more accurately termed a 'gap-count' penalty!

Changing the specification of $U(M)$ to better match intuition, or different scientific judgement, about likely patterns of insertion and deletion would often still yield a distribution amenable to posterior sampling using an appropriately modified MCMC algorithm. This would be especially straightforward if the penalty remained a sum over the individual gaps, but all that is really needed is that the change to the penalty when a match is deleted, added or switched is cheaply computed, meaning in practice that it uses only information that is local to the revision in M. Two possibilities that come immediately to mind are to use (19.5) but with a function f that is strictly concave but still increasing for positive gap lengths, or to use a form where the penalty is a decreasing function of the correlation between the matched (j, k) indices – with the effect that in the first example mentioned earlier, the x indices $(4, 5, 9)$ are less likely to be matched to the y indices $(7, 11, 12)$ than to $(7, 8, 12)$.

19.4 MAD-Bayes

MAD-Bayes (MAP-based Asymptotic Derivations from Bayes) is a novel methodology for fitting complex stochastic models due to Broderick et al. (2013). It was devised to meet the sometimes contradictory desiderata of complying with the Bayesian paradigm and delivering practical methodology that can be executed very quickly even on large data sets.

MAD-Bayes is essentially a simple framework for delivering small-variance asymptotic approximations to MAP (*maximum a posteriori*) estimation, yielding results that, while not usually of closed form, are nevertheless typically amenable to solution using fast optimisation techniques. It exploits the fact that in many statistical models, when the likelihood is taken to a 'small-variance' limit, a non-trivial limit is obtained for the MAP estimator, provided that hyperparameters in the prior are also taken to appropriate limits. Except in the simplest of cases, there may be more than one way to do this, giving different non-trivial limits, so some judgement is needed.

Although MAD-Bayes was conceived as a perspective to take in the presence of nonparametric priors and models with discrete allocation structures such as mixtures and clustering, the idea can be more simply illustrated and understood with a toy example from parametric Bayes. Suppose $y \sim N(X\beta, \sigma^2)$ with a normal prior: $\beta \sim N(\beta_0, \tau^2 I)$. Then, of course, the posterior is

$$\beta | y \sim N\left(\{\sigma^{-2}X^TX + \tau^{-2}I\}^{-1}\{\sigma^{-2}X^Ty + \tau^{-2}\beta_0\}, \{\sigma^{-2}X^TX + \tau^{-2}I\}^{-1}\right). \quad (19.7)$$

The posterior mean and mode are both $\{X^TX + \alpha I\}^{-1}\{X^Ty + \alpha\beta_0\}$, the value minimising $||y - X\beta||^2 + \alpha||\beta - \beta_0||^2$ over β, where $\alpha = \sigma^2/\tau^2$. This is a non-trivial combination of data and prior information, providing $0 < \sigma^2/\tau^2 < \infty$ strictly. Unlike the other applications of the MAD-Bayes principle for approximating the posterior mode and later the posterior distribution, discussed later in this chapter, these results hold exactly for any positive σ^2.

The canonical example of MAD-Bayes presented by Broderick et al. (2013) provides an extension to the classical K-means clustering algorithm that they call DP-means. They propose clustering multivariate data $(x_1, x_2, \ldots, x_n)$ by partitioning the index set $\{1, 2, \ldots, n\}$

as a disjoint union $\bigcup_{j=1}^{K} C_j$, where K, $\{C_j\}$ and cluster means $\{\mu_j\}$ are chosen to minimise

$$\sum_{j=1}^{K} \sum_{i \in C_j} ||x_i - \mu_j||^2 + (K-1)\lambda^2, \tag{19.8}$$

λ being a regularisation constant. This approach, intuitively reasonable in itself, can be derived by a MAD-Bayes argument approximating the MAP estimate of the clustering under a Dirichlet/Chinese restaurant process mixture model (Lo 1984). As with α in the normal linear model example discussed earlier, the constant λ^2 is the ratio of the variance σ^2 to a function of a hyperparameter in the prior, so the asymptotic framework again demands that the prior concentrates as the variance decreases. Broderick et al. (2013) further illustrate the idea applied to feature learning, particularly exploiting other Bayesian nonparametric prior models such as the Indian buffet process, and various extensions. The idea has more recently been used in feature learning for studying tumour heterogeneity by Xu et al. (2014).

A different kind of recent application is to image segmentation. Pereyra and McLaughlin (2014) apply a MAD-Bayes argument to the posterior arising from an image model based on a hidden Potts–Markov random field. Computing the MAP estimate in this problem is NP-hard, but a convex relaxation is possible, leading ultimately to an objective function of the form

$$\sum_{j=1}^{K} \sum_{i \in C_j} \{||y_i - x_i||^2 + ||x_i - \mu_j||^2\} + \beta ||\nabla x||_1, \tag{19.9}$$

to be minimised over $x, \mu, \{C_j\}$ and K, given a data image y. Here $||\nabla x||_1$ is the ℓ_1 norm of the first-order discrete gradient of the hidden image x, a convexification of the $||\nabla x||_0$ arising formally from the model. The minimisation over x is equivalent to a total-variation denoising problem of a kind which has been extensively studied in the recent optimisation literature and that can be solved very efficiently even in very high-dimensional scenarios using parallel proximal splitting methods. The minimisation over the other variables involves K-means clustering.

19.5 MAD-Bayes for unlabelled matching and alignment

To develop a MAD-Bayes method for matching and alignment, we use (19.1) to obtain, ignoring additive constants in the log-posterior,

$$\begin{aligned} &- 4\sigma^2 \log p(M, A, \tau|\sigma, x, y) = -4\sigma^2 \log\{|A|^n p(A) p(\tau)\} \\ &- 4\sigma^2 L \log(\rho/\lambda) + 4\sigma^2 dL \log(\sigma/\sqrt{2}) + 2\sigma^2 \log 2\pi + \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2. \end{aligned} \tag{19.10}$$

According to the MAD-Bayes approximation paradigm of Broderick et al. (2013), we should examine this function in the small-variance limit, as $\sigma^2 \to 0$. For a non-degenerate limit in this asymptotic analysis, the prior cannot be held fixed. Suppose $\rho/\lambda = \exp(\alpha/4\sigma^2)$ for some real constant α. Then as $\sigma \to 0$ in (19.10) we obtain

$$-4\sigma^2 \log p(M, A, \tau|\sigma, x, y) \to -\alpha L + \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2.$$

Thus, finding the MAP estimate of M, A, τ, the values maximising the log posterior, is asymptotically equivalent to minimising the penalised sum-of-squares

$$-\alpha L + \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2. \tag{19.11}$$

The similarity in general form between (19.11) and (19.8) or (19.9) is clear.

When $\alpha > 0$ there is a non-trivial solution, and the optimisation serves to limit the number of matches L; informally, with A and τ held fixed for simplicity of the argument, including an additional match (j', k') will decrease the penalised sum-of-squares if and only if $||x_{j'} - Ay_{k'} - \tau||^2 < \alpha$.

The parameter α controls the behaviour of the prior parameter ρ/λ in the small-variance limit: positive α implies that $\rho/\lambda \to \infty$ as $\sigma^2 \to 0$, at a particular rate. This is easy to understand qualitatively: if the noise variance is reduced so that matches become harder to find, that must be compensated by concentrating the prior for M on higher numbers of matches L.

In summary, this simple analysis of MAP inference in our Bayesian model has reduced to an optimisation problem, penalised least-squares, one with a fairly simple structure by the standards of problems addressable by modern optimisation techniques. For fixed A, τ, optimisation over M is an instance of a weighted matching problem for a bipartite graph, for which the Hungarian algorithm (Jacobi 1890; Munkres 1957) provides a solution; this is usually posed as a maximisation problem and the weight on edge (j, k) to be used would be simply $\max\{0, \alpha - ||x_j - Ay_k - \tau||^2\}$. For fixed M, optimisation over A and τ (say, in the case of rigid-body transformation) is an example of Procrustes analysis. It is easy to see (since each step reduces the value of the criterion (19.11) and because the set of possible alignments is finite) that alternating between these two steps defines an algorithm that converges to a possibly local optimum in a finite number of iterations. We stress that this may not be a global optimum as complex models often lead to multi-modal posteriors; we comment further on multi-modality in Section 19.11.

This simple idea could no doubt be improved using techniques from modern optimisation methodology. But even without such improvements, this algorithm runs very quickly. Without making any attempt to optimise coding of the outer loop, an implementation in `R`, using function `solve_LSAP` from package `clue` and function `procOPA` from package `shapes`, provides an algorithm that runs in 0.03 seconds on a 3.20 GHz processor for the small problem in Section 4.2 of Green and Mardia (2006), to be compared to 10.85 seconds for 10^6 sweeps of the MCMC sampler on the same problem (but which, of course, provides a much richer inference).

Note that use of the Hungarian algorithm, or other code for the assignment problem, guarantees that the inferred alignment is feasible in the sense that no point is simultaneously matched to more than one point in the other configuration, in contrast to the formally somewhat similar method using the EM algorithm to compute the maximum likelihood estimate of the alignment (see for example Kent et al. 2004).

There is a related approach called ‘Softassign Procrustes’ to this problem due to Rangarajan et al. (1997). This proceeds by first relaxing the constraint that M is a binary matrix to set up an iterative deterministic annealing algorithm using Lagrange multipliers that alternates between updating the geometrical parameters and updating M; the method appeals to a theorem of Sinkhorn (1964) to deliver a solution in which M is in fact binary. The Softassign Procrustes algorithm has been given an EM-like interpretation by Kent et al. (2010).

19.6 Omniparametric optimisation of the objective function

An interesting perspective on the optimisation of (19.11) allows simultaneous consideration of all $\alpha \in (0, \infty)$, delivering what is often called a 'regularisation path' (for example, in the context of the Lasso (Efron et al. 2004)). Picture a two-dimensional scatter plot of points, each representing a possible alignment M, with horizontal coordinate $L(M)$ and vertical coordinate $\sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2$.

The optimal M according to (19.11) corresponds to the point where a line of slope α is a lower tangent to the scatter of points, and the set of all M that are optimal for some α is represented by the lower convex hull of the configuration. Because there are only finitely many possible values of M, this lower convex hull is a polygonal line, so there exists a finite grid of values of α, say $\alpha_0 > \alpha_1 > \alpha_2 > \cdots$, such that for all $\alpha \in (\alpha_{i+1}, \alpha_i)$, $i = 0, 1, \ldots$, the optimal M is constant, say $\hat{M}_i$. Note that $L(\hat{M}_i)$ will decrease with i. One approach to constructing the α_i and $\hat{M}_i$, following a suggestion of the referees, is to proceed sequentially for $i = 0, 1, \ldots$, using each α_i as a starting point for determining α_{i+1}.

The setup also invites comparison with that of Lau and Green (2007), who discussed optimal Bayesian point estimation of a clustering (of gene expression profiles) based on a pairwise-coincidence loss function. 'Omniparametric' optimisation of the expected loss over all values of the parameter in the loss function was implemented in a fast heuristic algorithm, which might be used to inspire a similar approach to the present problem. Following that paradigm would suggest iteratively refining the grid (α_i), starting with an initial pair of (low, high) values; the recursive step to split an interval (α_{i+1}, α_i) would search for an alignment M whose representative point in this diagram lies outside the line segment determined by the interval endpoints.

19.7 MAD-Bayes in the sequence-labelled case

In the sequence-labelled case, the points in each configuration are numbered in sequential order (along a protein, in typical application), and we can use this numbering in specifying a prior on the alignment matrix M. This leads to the posterior (19.4) instead of (19.1). The 'energy function' $U(M)$ in the prior for M may take the gap penalty form (19.3) or something more general, either with the same intention of promoting or insisting upon sequence order being maintained, or with some other purpose.

For such a posterior, we obtain

$$\begin{aligned} -4\sigma^2 \log p(M, A, \tau | \sigma, x, y) &= -4\sigma^2 \log\{|A|^n p(A) p(\tau) v^L\} \\ &+ 4\sigma^2 U(M) + 4\sigma^2 dL \log(\sigma/\sqrt{2}) + 2\sigma^2 \log 2\pi + \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2. \end{aligned} \tag{19.12}$$

Since all of the other terms vanish as $\sigma^2 \to 0$, we need for a non-trivial limit that $U(M)$ or its parameters scale in such a way that $4\sigma^2 U(M)$ has a non-trivial limit. For example, in the case of the gap penalty (19.3), if $8\sigma^2 h \to \alpha$ and $4\sigma^2(g - h) \to \beta$, then according to (19.6) the resulting optimisation problem is to minimise

$$-\alpha L + \beta S + \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2. \tag{19.13}$$

Intuitive interpretation of this objective function is less straightforward: adding a match always increases L by 1, but the associated change in S may be $+2, +1, 0, -1$ or -2. Optimisation over the alignment for fixed A and τ is no longer a weighted matching problem, taking this setup out of reach of the Hungarian algorithm; as suggested by the referees, there may be a role here for dynamic programming.

19.8 Other kinds of labelling

In their Section 3.6, Green and Mardia (2006) propose a way to extend the model leading to (19.1) to allow simultaneous model-based inference about alignment when the points in the observed configurations are recorded as belonging to different clusters, or 'colours', and pairs of points where both belong to the same cluster are more likely to be matched. An example in protein bioinformatics arises when the amino acids characterising the observed points are categorised as hydrophobic or hydrophilic (possibly subdivided into charged, polar and glycine). The model extension achieving this amounts to modifying the prior on the alignment matrix M to favour like-coloured matches, so provides a general mechanism for handling 'partially labelled' configurations, where labels are not unique.

The modified prior on M that was proposed has the form

$$p(M) \propto \left(\frac{\rho}{\lambda v}\right)^L \prod_{j \underset{M}{\sim} k} \exp(\gamma I[r_j = s_k] + \delta I[r_j \neq s_k]),$$

where x_j is coloured r_j and y_k coloured s_k, instead of (19.2). This modification needs only trivial changes to the Metropolis–Hastings updating of M in the posterior simulation.

It is easy to see that such modified priors also lead to a simply modified MAD-Bayes objective function. The penalised sum-of-squares (19.11) is replaced by

$$-\alpha L + \sum_{j \underset{M}{\sim} k} \{||x_j - Ay_k - \tau||^2 + \gamma' I[r_j = s_k] + \delta' I[r_j \neq s_k]\}, \tag{19.14}$$

where $\gamma' = 4\sigma^2\gamma$ and $\delta' = 4\sigma^2\delta$.

Numerical optimisation of (19.14) can again in principle be addressed by alternating between optimising over M and over A and τ, and again the former step is an instance of a weighted matching problem, since the objective function can be expressed as a sum over $\{(j, k) : M_{jk} = 1\}$.

The extensions in this section and the previous one can readily be combined, simultaneously penalising gaps and favouring like-coloured matches, and giving the objective function

$$-\alpha L + \beta S + \sum_{j \underset{M}{\sim} k} \{||x_j - Ay_k - \tau||^2 + \gamma' I[r_j = s_k] + \delta' I[r_j \neq s_k]\}.$$

19.9 Simultaneous alignment of multiple configurations

Ruffieux and Green (2009) generalised the two-configuration methodology of Green and Mardia (2006) to handle the case of multiple configurations. They argue that information is lost by treating the configurations pairwise; the truth of this is most easily seen in the kind of latent-true-configuration model they use (since we should want to use all information at once in the implicit inference about the positions of the latent points), but the point will be generally true. Kenobi and Dryden (2012) match multiple configurations using a model that considers them only two at a time. The ideas illustrated in this chapter will continue to apply *mutatis mutandis* to the multiple-configuration case, although I do not know whether the discrete optimisation algorithms that would be needed for implementation are still instances of standard optimisation theory problems.

19.10 Beyond MAD-Bayes to posterior approximation?

The motivating example in the Gaussian case delivered the whole posterior (19.7) not only the posterior mode. Could we extend the MAD-Bayes perspective to deliver at least an approximation to the posterior, by slightly refining the asymptotic argument? In this section, I attempt only a preliminary, speculative answer to this question, which seems a promising subject for further investigation.

For the unlabelled case, leaving aside technicalities for the moment, the argument leading to the penalised least-squares objective function (19.11) equally well delivers the formal approximation, valid as $\sigma^2 \to 0$,

$$p(M, A, \tau|\sigma, x, y) \approx e^{\alpha L/4\sigma^2} \exp\{(-1/4\sigma^2) \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2\}. \tag{19.15}$$

Our focus will be to investigate the form of the density on the right-hand side. For definiteness, we take the case of rigid-body transformations, so that A is special orthogonal.

It is possible to make some progress interpreting the approximate joint posterior (19.15) by considering the full conditionals for each of A, τ and M in turn.

For A,

$$\begin{aligned}\sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2) &= \sum_{j \underset{M}{\sim} k} \{||x_j - \tau||^2 + ||y_k||^2 - 2(Ay_k)^T(x_j - \tau)\} \\ &= \sum_{j \underset{M}{\sim} k} ||x_j - \tau||^2 + \sum_{j \underset{M}{\sim} k} ||y_k||^2 - 2\text{tr}\{A^T \sum_{j \underset{M}{\sim} k} (x_j - \tau)y_k^T\}.\end{aligned} \tag{19.16}$$

This reveals that under the approximate distribution (19.15), A given τ and M (and x, y, σ) has a matrix Fisher distribution (Mardia and Jupp 2000, p. 289), as shown in Green and Mardia (2006). The normalising constant of this distribution is known, so that A can be integrated out to give

$$p(M, \tau|\sigma, x, y) \approx e^{\alpha L/4\sigma^2} \exp\left\{(-1/4\sigma^2) \sum_{j \underset{M}{\sim} k} ||x_j - \tau||^2 + ||y_k||^2\right\} {}_0F_1(p/2, (1/16\sigma^4)F^T F), \tag{19.17}$$

where $F = \sum_{j \underset{M}{\sim} k} (x_j - \tau)y_k^T$ depends on both τ and M, as well as the data. This does not seem amenable to further analytic simplification.

Similarly, we can evidently extract from the right hand side of (19.15) the approximate conditional for τ given M and A as

$$\tau|M, A, x, y, \sigma \sim N\left(L^{-1}\sum_{j \underset{M}{\sim} k}(x_j - Ay_k), 2\sigma^2/L\right),$$

while the approximate conditional for M given τ and A is also explicit but hardly tractable.

In an effort to gain more insight into the form of the approximate posterior, we could consider one of the approximations to the matrix Fisher distribution developed by Khatri and Mardia (1977) and Bingham et al. (1992). However, these seem too intricate to use for practical statistical analysis.

So let us consider further approximation: we could try to use a Normal approximation for $p(A|M, \tau, x, y, \sigma)$. Suppose that $A \sim \text{MatrixFisher}(F)$ with F non-singular; note that this demands that the M in question matches sufficiently many (x, y) pairs with coordinates in general position. Now let $K = (F^TF)^{1/2}$ be the elliptical part of F and $N = FK^{-1}$ its polar part (Mardia and Jupp 2000, p. 286). Let $V\Delta V^T$ with $\Delta = \text{diag}(\delta_1, \delta_2, \ldots, \delta_d)$ be the spectral decomposition of K. In the concentrated case, where all δ_i become large (many matches), we have (Peter Jupp, *personal communication*)

$$(A - N) \approx NVSV^T,$$

where S is a skew-symmetric matrix with $(\delta_i + \delta_j)^{1/2}s_{ij} \sim N(0, 1)$, independently.

It seems probable that the argument leading to this can be refined to yield a joint Normal approximation for $p(A, \tau|M, x, y, \sigma)$, although I have not attempted to verify the details. Under such an approximation, the approximate joint posterior (19.15) becomes a Normal mixture distribution, and this seems to be the analysis of (19.15) most likely to be useful for numerical implementation. More work is needed here.

Returning to the mathematical basis for the approximation, a rigorous analysis would need to establish that the approximation of densities that we have investigated really does imply convergence of the probability measures (say, in the sense of total variation norm) under suitable regularity conditions.

19.11 Practical uses of MAD-Bayes approximations

It is hoped that the optimisation-based techniques suggested in this chapter could be developed to make a practically useful contribution to methodology. They seem to offer to supply some of the advantages of the Bayesian approach – notably treating uncertainty about the alignment and the geometrical transformation symmetrically and simultaneously – without having to pay the price of relying on Monte Carlo computation.

However, even neglecting the fact that the Bayesian setup has to be approximated to allow delivery of these optimisation solutions, there are other caveats. In particular, they are not a panacea for the problems of multi-modality that can bedevil MCMC methods. The MAD-Bayes perspective is really blind to the possible existence of modes other than the one under consideration, and numerical optimisation methods need to be special and carefully chosen to deliver optima of multi-modal objective functions reliably, just as MCMC methods have to be specifically designed to handle multi-modal target distributions.

It may be useful to regard optimisation approaches as complementary to posterior sampling – for example, MAD-Bayes might provide a starting point for an MCMC simulation, from which perhaps a rather short MCMC run might be used to assess variability; again this would demand some guarantee about unimodality for reliable inference. This is very much in the spirit of the work of Mardia et al. (2007).

Acknowledgments

I am indebted to Peter Jupp for kindly deriving the high-concentration Normal approximation to the matrix Fisher distribution for me, to the referees for very helpful comments and pointers, to my collaborators on these problems, and above all to Kanti Mardia for his friendship and many stimulating discussions over the years.

References

Bingham C, Chang T and Richards D 1992 Approximating the matrix Fisher and Bingham distributions: applications to spherical regression and Procrustes analysis. *Journal of Multivariate Analysis* **41**, 314–337.

Broderick T, Kulis B and Jordan MI 2013 MAD-Bayes: MAP-based asymptotic derivations from Bayes *Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA. JMLR: W&CP volume 28.* See also arXiv:1212.2126.

Efron B, Hastie T, Johnstone I and Tibshirani R 2004 Least angle regression. *Annals of Statistics* **32**(2), 407–499.

Fallaize CJ, Green PJ, Mardia KV and Barber S 2014 Bayesian protein sequence and structure alignment Current version at arXiv:1404.1556.

Green PJ and Mardia KV 2006 Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika* **93**, 235–254. doi:10.1093/biomet/93.2.235

Green PJ, Mardia KV, Nyirongo VB and Ruffieux Y 2010 Bayesian modelling for matching and alignment of biomolecules *The Oxford Handbook of Applied Bayesian Analysis* Oxford University Press, pp. 27–50.

Jacobi CGJ 1890 De aequationum differentialum systemate non normali ad formam normalem revocando *C.G.J. Jacobi's Gesammelte Werke, fünfter Band* K. Weierstrass, Berlin, Bruck und Verlag von Georg Reimer, pp. 485–513.

Kenobi K and Dryden IL 2012 Bayesian matching of unlabeled point sets using Procrustes and configuration models. *Bayesian Analysis* **7**(3), 547–566.

Kent JT, Mardia KV and Taylor CC 2004 Matching problems for unlabelled configurations In *Bioinformatics, Images, and Wavelets* Aykroyd RG, Barber S and Mardia KV (eds), Leeds University Press, pp. 33–36.

Kent JT, Mardia KV and Taylor CC 2010 An EM interpretation of the Softassign algorithm for alignment problems In *High-Throughput Sequencing, Proteins and Statistics* Gusnanto A, Mardia KV, Fallaize CJ and Voss J (eds), Leeds University Press, pp. 29–32.

Khatri CG and Mardia KV 1977 The von Mises–Fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society, Series B* **39**, 95–106.

Lau JW and Green PJ 2007 Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16**, 526–558.

Lo AY 1984 On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics* **12**(1), 351–357.

Mardia KV and Jupp PE 2000 *Directional Statistics*. John Wiley & Sons, Ltd, Chichester.

Mardia KV, Nyirongo VB, Green PJ, Gold ND and Westhead DR 2007 Bayesian refinement of protein functional site matching. *BMC Bioinformatics* **8**, 257.

Munkres J 1957 Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics* **5**(1), 32–38.

Pereyra M and McLaughlin S 2014 Small-variance asymptotics of hidden Potts-MRFs: Application to fast Bayesian image segmentation In *Proceedings of European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal.

Rangarajan A, Chui H and Bookstein FL 1997 The Softassign Procrustes matching algorithm In *Information Processing in Medical Imaging*, *Lecture Notes in Computer Science*, Vol. 1230, pp. 29–42.

Rodriguez A and Schmidler S 2010 Bayesian protein structure alignment *Ann. Appl. Stat*. **8**(4) (2014), 2068–2095.

Ruffieux Y and Green PJ 2009 Alignment of multiple configurations using hierarchical models. *Journal of Computational and Graphical Statistics* **18**, 756–773. doi:10.1198/jcgs.2009.07048

Sinkhorn R 1964 A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics* **35**, 876–879.

Xu Y, Mueller P, Yuan Y, Gulukota K and Ji Y 2014 MAD Bayes for tumor heterogeneity – feature allocation with exponential family sampling. http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2014.995794#.VWiyKE3bLBQ.

Index

Geometry Driven Statistics, First Edition. Edited by Ian L. Dryden and John T. Kent.

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

The Wiley Series in Probability and Statistics is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

† ABRAHAM and LEDOLTER · Statistical Methods for Forecasting

AGRESTI · Analysis of Ordinal Categorical Data, Second Edition

AGRESTI · An Introduction to Categorical Data Analysis, Second Edition

AGRESTI · Categorical Data Analysis, Third Edition

ALSTON, MENGERSEN and PETTITT (editors) · Case Studies in Bayesian Statistical Modelling and Analysis

ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist

AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data

AMARATUNGA, CABRERA, and SHKEDY · Exploration and Analysis of DNA Microarray and Other High-Dimensional Data, Second Edition

ANDEˇL · Mathematics of Chance

ANDERSON · An Introduction to Multivariate Statistical Analysis, Third Edition

* ANDERSON · The Statistical Analysis of Time Series

ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies

ANDERSON and LOYNES · The Teaching of Practical Statistics

ARMITAGE and DAVID (editors) · Advances in Biometry

BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, Second Edition

BOLLEN · Structural Equations with Latent Variables

BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective

BONNINI, CORAIN, MAROZZI and SALMASO · Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R

BOROVKOV · Ergodicity and Stability of Stochastic Processes

BOSQ and BLANKE · Inference and Prediction in Large Dimensions

BOULEAU · Numerical Methods for Stochastic Processes

* BOX and TIAO · Bayesian Inference in Statistical Analysis

BOX · Improving Almost Anything, Revised Edition

* BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement

BOX and DRAPER · Response Surfaces, Mixtures, and Ridge Analyses, Second Edition

BOX, HUNTER, and HUNTER · Statistics for Experimenters: Design, Innovation, and Discovery, Second Editon

BOX, JENKINS, and REINSEL · Time Series Analysis: Forecasting and Control, Fourth Edition

BOX, LUCEÑO, and PANIAGUA-QUIÑONES · Statistical Control by Monitoring and Adjustment, Second Edition

* BROWN and HOLLANDER · Statistics: A Biomedical Introduction

CAIROLI and DALANG · Sequential Stochastic Optimization

CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science

CHAN · Time Series: Applications to Finance with R and S-Plus®, Second Edition

CHARALAMBIDES · Combinatorial Methods in Discrete Distributions

CHATTERJEE and HADI · Regression Analysis by Example, Fourth Edition

CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression

CHEN · The Fitness of Information: Quantitative Assessments of Critical Evidence

CHERNICK · Bootstrap Methods: A Guide for Practitioners and Researchers, Second Edition

CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences

CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty, Second Edition

CHIU, STOYAN, KENDALL and MECKE · Stochastic Geometry and Its Applications, Third Edition

CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, Third Edition

CLARKE · Linear Models: The Theory and Application of Analysis of Variance

CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, Second Edition

* COCHRAN and COX · Experimental Designs, Second Edition

COLLINS and LANZA · Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences

CONGDON · Applied Bayesian Modelling, Second Edition

CONGDON · Bayesian Models for Categorical Data

CONGDON · Bayesian Statistical Modelling, Second Edition

CONOVER · Practical Nonparametric Statistics, Third Edition

COOK · Regression Graphics

COOK and WEISBERG · An Introduction to Regression Graphics

COOK and WEISBERG · Applied Regression Including Computing and Graphics

CORNELL · A Primer on Experiments with Mixtures

CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, Third Edition

COX · A Handbook of Introductory Statistical Methods

CRESSIE · Statistics for Spatial Data, Revised Edition

CRESSIE and WIKLE · Statistics for Spatio-Temporal Data

CSÖRGO´´ and HORVÁTH · Limit Theorems in Change Point Analysis

Dagpunar · Simulation and Monte Carlo: With Applications in Finance and MCMC

DANIEL · Applications of Statistics to Industrial Experimentation

DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, Eighth Edition

* DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, Second Edition

DASU and JOHNSON · Exploratory Data Mining and Data Cleaning

DAVID and NAGARAJA · Order Statistics, Third Edition

DAVINO, FURNO and VISTOCCO · Quantile Regression: Theory and Applications

* DEGROOT, FIENBERG, and KADANE · Statistics and the Law

DEL CASTILLO · Statistical Process Adjustment for Quality Control

DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables

DEMIDENKO · Mixed Models: Theory and Applications with R, Second Edition

DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression

DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis

DEY and MUKERJEE · Fractional Factorial Plans

DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications

* DODGE and ROMIG · Sampling Inspection Tables, Second Edition

* DOOB · Stochastic Processes

DOWDY, WEARDEN, and CHILKO · Statistics for Research, Third Edition

DRAPER and SMITH · Applied Regression Analysis, Third Edition

DRYDEN and MARDIA · Statistical Shape Analysis

DUDEWICZ and MISHRA · Modern Mathematical Statistics

DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, Fourth Edition

DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations

EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment

* ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis

ENDERS · Applied Econometric Time Series, Third Edition

† ETHIER and KURTZ · Markov Processes: Characterization and Convergence

EVANS, HASTINGS, and PEACOCK · Statistical Distributions, Third Edition

EVERITT, LANDAU, LEESE, and STAHL · Cluster Analysis, Fifth Edition

FEDERER and KING · Variations on Split Plot and Split Block Experiment Designs

FELLER · An Introduction to Probability Theory and Its Applications, Volume I, Third Edition, Revised; Volume II, Second Edition

FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis, Second Edition

* FLEISS · The Design and Analysis of Clinical Experiments

FLEISS · Statistical Methods for Rates and Proportions, Third Edition

† FLEMING and HARRINGTON · Counting Processes and Survival Analysis

FUJIKOSHI, ULYANOV, and SHIMIZU · Multivariate Statistics: High-Dimensional and Large-Sample Approximations

FULLER · Introduction to Statistical Time Series, Second Edition

† FULLER · Measurement Error Models

GALLANT · Nonlinear Statistical Models

GEISSER · Modes of Parametric Statistical Inference

GELMAN and MENG · Applied Bayesian Modeling and Causal Inference from ncomplete-Data Perspectives

GEWEKE · Contemporary Bayesian Econometrics and Statistics

GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation

GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments

GIFI · Nonlinear Multivariate Analysis

GIVENS and HOETING · Computational Statistics

GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems

GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, Second Edition

GOLDSTEIN · Multilevel Statistical Models, Fourth Edition

GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues

GOLDSTEIN and WOOFF · Bayes Linear Statistics

GRAHAM · Markov Chains: Analytic and Monte Carlo Computations

GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing

GROSS, SHORTLE, THOMPSON, and HARRIS · Fundamentals of Queueing Theory, Fourth Edition

GROSS, SHORTLE, THOMPSON, and HARRIS · Solutions Manual to Accompany Fundamentals of Queueing Theory, Fourth Edition

* HAHN and SHAPIRO · Statistical Models in Engineering

HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners

HALD · A History of Probability and Statistics and their Applications Before 1750

† HAMPEL · Robust Statistics: The Approach Based on Influence Functions

HARTUNG, KNAPP, and SINHA · Statistical Meta-Analysis with Applications

HEIBERGER · Computation for the Analysis of Designed Experiments

HEDAYAT and SINHA · Design and Inference in Finite Population Sampling

HEDEKER and GIBBONS · Longitudinal Data Analysis

HELLER · MACSYMA for Statisticians

HERITIER, CANTONI, COPT, and VICTORIA-FESER · Robust Methods in Biostatistics

HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, Second Edition

HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 2: Advanced Experimental Design

HINKELMANN (editor) · Design and Analysis of Experiments, Volume 3: Special Designs and Applications

HOAGLIN, MOSTELLER, and TUKEY · Fundamentals of Exploratory Analysis of Variance

* HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes

* HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis

HOCHBERG and TAMHANE · Multiple Comparison Procedures

HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, Third Edition

HOEL · Introduction to Mathematical Statistics, Fifth Edition

HOGG and KLUGMAN · Loss Distributions

HOLLANDER, WOLFE, and CHICKEN · Nonparametric Statistical Methods, Third Edition

HOSMER and LEMESHOW · Applied Logistic Regression, Second Edition

HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling of Time-to-Event Data, Second Edition

HUBER · Data Analysis: What Can Be Learned From the Past 50 Years

HUBER · Robust Statistics

† HUBER and Ronchetti · Robust Statistics, Second Edition

HUBERTY · Applied Discriminant Analysis, Second Edition

HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis, Second Edition

HUITEMA · The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition

HUNT and KENNEDY · Financial Derivatives in Theory and Practice, Revised Edition

HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory and Practice

HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek – with Commentary

HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data

JACKMAN · Bayesian Analysis for the Social Sciences

† JACKSON · A User's Guide to Principle Components

JOHN · Statistical Methods in Engineering and Quality Assurance

JOHNSON · Multivariate Statistical Simulation

JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz

JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, Third Edition

JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present

JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, Second Edition

JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, Second Edition

JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions

JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, Second Edition

JUREK and MASON · Operator-Limit Distributions in Probability Theory

KADANE · Bayesian Methods and Ethics in a Clinical Trial Design

KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence

KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, Second Edition

KARIYA and KURATA · Generalized Least Squares

KASS and VOS · Geometrical Foundations of Asymptotic Inference

† KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis

KEDEM and FOKIANOS · Regression Models for Time Series Analysis

KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory

KHURI · Advanced Calculus with Applications in Statistics, Second Edition

KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models

* KISH · Statistical Design for Research

KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences

KLEMELÄ · Smoothing of Multivariate Data: Density Estimation and Visualization

KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions, Third Edition

KLUGMAN, PANJER, and WILLMOT · Loss Models: Further Topics

KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions, Third Edition

KOSKI and NOBLE · Bayesian Networks: An Introduction

KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, Second Edition

KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index

KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume

KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1

KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2

KOWALSKI and TU · Modern Applied U-Statistics

KRISHNAMOORTHY and MATHEW · Statistical Tolerance Regions: Theory, Applications, and Computation

KROESE, TAIMRE, and BOTEV · Handbook of Monte Carlo Methods

KROONENBERG · Applied Multiway Data Analysis

KULINSKAYA, MORGENTHALER, and STAUDTE · Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence

KULKARNI and HARMAN · An Elementary Introduction to Statistical Learning Theory

KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling

KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering

LACHIN · Biostatistical Methods: The Assessment of Relative Risks, Second Edition

LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction

LAMPERTI · Probability: A Survey of the Mathematical Theory, Second Edition

LAWLESS · Statistical Models and Methods for Lifetime Data, Second Edition

LAWSON · Statistical Methods in Spatial Epidemiology, Second Edition

LE · Applied Categorical Data Analysis, Second Edition

LE · Applied Survival Analysis

Lee · Structural Equation Modeling: A Bayesian Approach

LEE and WANG · Statistical Methods for Survival Data Analysis, Fourth Edition

LePAGE and BILLARD · Exploring the Limits of Bootstrap

LESSLER and KALSBEEK · Nonsampling Errors in Surveys

LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics

LIAO · Statistical Group Comparison

LIN · Introductory Stochastic Analysis for Finance and Insurance

LINDLEY · Understanding Uncertainty, Revised Edition

LITTLE and RUBIN · Statistical Analysis with Missing Data, Second Edition

Lloyd · The Statistical Analysis of Categorical Data

LOWEN and TEICH · Fractal-Based Point Processes

MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition

MALLER and ZHOU · Survival Analysis with Long Term Survivors

MARCHETTE · Random Graphs for Statistical Pattern Recognition

MARDIA and JUPP · Directional Statistics

MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice

MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods

MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, Second Edition

McCULLOCH, SEARLE, and NEUHAUS · Generalized, Linear, and Mixed Models, Second Edition

McFADDEN · Management of Data in Clinical Trials, Second Edition

* McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition

McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data

McLACHLAN and KRISHNAN · The EM Algorithm and Extensions, Second Edition

McLACHLAN and PEEL · Finite Mixture Models

McNEIL · Epidemiological Research Methods

MEEKER and ESCOBAR · Statistical Methods for Reliability Data

MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice

MENGERSEN, ROBERT, and TITTERINGTON · Mixtures: Estimation and Applications

MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and Regression, Third Edition

* MILLER · Survival Analysis, Second Edition

MONTGOMERY, JENNINGS, and KULAHCI · Introduction to Time Series Analysis and Forecasting

MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, Fifth Edition

MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness

MUIRHEAD · Aspects of Multivariate Statistical Theory

MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks

MURTHY, XIE, and JIANG · Weibull Models

MYERS, MONTGOMERY, and ANDERSON-COOK · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, Third Edition

MYERS, MONTGOMERY, VINING, and ROBINSON · Generalized Linear Models. With Applications in Engineering and the Sciences, Second Edition

NATVIG · Multistate Systems Reliability Theory With Applications

† NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses

† NELSON · Applied Life Data Analysis

NEWMAN · Biostatistical Methods in Epidemiology

NG, TAIN, and TANG · Dirichlet Theory: Theory, Methods and Applications

OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, Second Edition

OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis

PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions

PANJER · Operational Risk: Modeling and Analytics

PANKRATZ · Forecasting with Dynamic Regression Models

PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases

PARDOUX · Markov Processes and Applications: Algorithms, Networks, Genome and Finance

PARMIGIANI and INOUE · Decision Theory: Principles and Approaches

* PARZEN · Modern Probability Theory and Its Applications

PEÑA, TIAO, and TSAY · A Course in Time Series Analysis

PESARIN and SALMASO · Permutation Tests for Complex Data: Applications and Software

PIANTADOSI · Clinical Trials: A Methodologic Perspective, Second Edition

POURAHMADI · Foundations of Time Series Analysis and Prediction Theory

POURAHMADI · High-Dimensional Covariance Estimation

POWELL · Approximate Dynamic Programming: Solving the Curses of Dimensionality, Second Edition

POWELL and RYZHOV · Optimal Learning

PRESS · Subjective and Objective Bayesian Statistics, Second Edition

PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach

PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics

† PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming

QIU · Image Processing and Jump Regression Analysis

* RAO · Linear Statistical Inference and Its Applications, Second Edition

RAO · Statistical Inference for Fractional Diffusion Processes

RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, Second Edition

RAYNER, THAS, and BEST · Smooth Tests of Goodnes of Fit: Using R, Second Edition

RENCHER and SCHAALJE · Linear Models in Statistics, Second Edition

RENCHER and CHRISTENSEN · Methods of Multivariate Analysis, Third Edition

RENCHER · Multivariate Statistical Inference with Applications

RIGDON and BASU · Statistical Methods for the Reliability of Repairable Systems

* RIPLEY · Spatial Statistics

* RIPLEY · Stochastic Simulation

ROHATGI and SALEH · An Introduction to Probability and Statistics, Second Edition

ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance

ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice

ROSSI, ALLENBY, and McCULLOCH · Bayesian Statistics and Marketing

† ROUSSEEUW and LEROY · Robust Regression and Outlier Detection

ROYSTON and SAUERBREI · Multivariate Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables

* RUBIN · Multiple Imputation for Nonresponse in Surveys

RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, Second Edition

RUBINSTEIN and MELAMED · Modern Simulation and Modeling

RUBINSTEIN, RIDDER, and VAISMAN · Fast Sequential Monte Carlo Methods for Counting and Optimization

RYAN · Modern Engineering Statistics

RYAN · Modern Experimental Design

RYAN · Modern Regression Methods, Second Edition

RYAN · Sample Size Determination and Power

RYAN · Statistical Methods for Quality Improvement, Third Edition

SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications

SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis

SCHERER · Batch Effects and Noise in Microarray Experiments: Sources and Solutions

* SCHEFFE · The Analysis of Variance

SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application

SCHOTT · Matrix Analysis for Statistics, Second Edition

SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives

SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization

* SEARLE · Linear Models

† SEARLE · Linear Models for Unbalanced Data

† SEARLE · Matrix Algebra Useful for Statistics

† SEARLE, CASELLA, and McCULLOCH · Variance Components

SEARLE and WILLETT · Matrix Algebra for Applied Economics

SEBER · A Matrix Handbook For Statisticians

† SEBER · Multivariate Observations

SEBER and LEE · Linear Regression Analysis, Second Edition

† SEBER and WILD · Nonlinear Regression

SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems

* SERFLING · Approximation Theorems of Mathematical Statistics

SHAFER and VOVK · Probability and Finance: It's Only a Game!

SHERMAN · Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties

SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions

SINGPURWALLA · Reliability and Risk: A Bayesian Perspective

SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference

SRIVASTAVA · Methods of Multivariate Statistics

STAPLETON · Linear Statistical Models, Second Edition

STAPLETON · Models for Probability and Statistical Inference: Theory and Applications

STAUDTE and SHEATHER · Robust Estimation and Testing

STOYAN · Counterexamples in Probability, Second Edition

STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics

STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods

STYAN · The Collected Papers of T. W. Anderson: 1943–1985

SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research

TAKEZAWA · Introduction to Nonparametric Regression

TAMHANE · Statistical Analysis of Designed Experiments: Theory and Applications

TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory

THOMPSON · Empirical Model Building: Data, Models, and Reality, Second Edition

THOMPSON · Sampling, Third Edition

THOMPSON · Simulation: A Modeler's Approach

THOMPSON and SEBER · Adaptive Sampling

THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets

TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics

TROFFAES and DE COOMAN · Lower Previsions

TSAY · Analysis of Financial Time Series, Third Edition

TSAY · An Introduction to Analysis of Financial Data with R

TSAY · Multivariate Time Series Analysis: With R and Financial Applications