ICSA Book Series in Statistics
Series Editors: Jiahua Chen · Ding-Geng (Din) Chen

Dongseok Choi · Daeheung Jang
Tze Leung Lai · Youngjo Lee · Ying Lu
Jun Ni · Peter Qian · Peihua Qiu
George Tiao *Editors*

Proceedings of the Pacific Rim Statistical Conference for Production Engineering

Big Data, Production Engineering and Statistics

Springer

ICSA Book Series in Statistics

The ICSA Book Series in Statistics showcases research from the International Chinese Statistical Association that has an international reach. It publishes books in statistical theory, applications, and statistical education. All books are associated with the ICSA or are authored by invited contributors. Books may be monographs, edited volumes, textbooks and proceedings.

More information about this series at http://www.springer.com/series/13402

Dongseok Choi · Daeheung Jang
Tze Leung Lai · Youngjo Lee
Ying Lu · Jun Ni · Peter Qian
Peihua Qiu · George Tiao
Editors

Proceedings of the Pacific Rim Statistical Conference for Production Engineering

Big Data, Production Engineering and Statistics

Springer

Editors
Dongseok Choi
Oregon Health and Science University
Portland, OR
USA

Daeheung Jang
Pukyong National University
Busan
Korea (Republic of)

Tze Leung Lai
Stanford University
Stanford, CA
USA

Youngjo Lee
Seoul National University
Gwanak-gu, Seoul
Korea (Republic of)

Ying Lu
Stanford University
Stanford, CA
USA

Jun Ni
University of Michigan–Ann Arbor
Ann Arbor, MI
USA

Peter Qian
University of Wisconsin–Madison
Madison, WI
USA

Peihua Qiu
University of Florida
Gainesville, FL
USA

George Tiao
University of Chicago
Chicago, IL
USA

ISSN 2199-0980 ISSN 2199-0999 (electronic)
ICSA Book Series in Statistics
ISBN 978-981-10-8167-5 ISBN 978-981-10-8168-2 (eBook)
https://doi.org/10.1007/978-981-10-8168-2

Library of Congress Control Number: 2017964576

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
part of Springer Nature
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The Pacific Rim area is one of the key manufacturing sites in the world, and applications of statistical thinking and methods for production engineering have never been more important with big data. To address the need, a statistical conference for production engineering was first proposed by Prof. George Tiao, The University of Chicago, during his opening remarks at 2014 Joint Applied Statistics Symposium of the International Chinese Statistical Association and Korean International Statistical Society in Portland. The first conference was held at Shanghai Center for Mathematical Sciences located in Fudan University in December 2014. The main goal was to bring researchers and practitioners in statistics and engineering from academe and industry to promote collaborations and exchange the latest advancements in methodology and real-world challenges among participants. Following the success of the first conference, the 2nd Pacific Rim Statistical Conference for Production Engineering was held at Seoul National University in December 2016. These proceedings present the selected papers based on the presentations at the first and second Pacific Rim Statistical Conferences for Production Engineering. We hope that this effort can stimulate further collaborations between academe and industry in production engineering.

The conference series has become a major joint event of the International Chinese Statistical Association and Korean International Statistical Society. We welcome those who are interested in this endeavor to join the third conference that will be held at National Tsing Hua University in Taiwan in 2018.

Portland, USA	Dongseok Choi
Busan, Korea (Republic of)	Daeheung Jang
Stanford, USA	Tze Leung Lai
Seoul, Korea (Republic of)	Youngjo Lee
Stanford, USA	Ying Lu
Ann Arbor, USA	Jun Ni
Madison, USA	Peter Qian
Gainesville, USA	Peihua Qiu
Chicago, USA	George Tiao

Contents

Part I
Design and Collection of Big Data

Chapter 1
Bottom-Up Estimation and Top-Down Prediction: Solar Energy Prediction Combining Information from Multiple Sources

Youngdeok Hwang, Siyuan Lu and Jae-Kwang Kim

Abstract Accurately forecasting solar power using the data from multiple sources is an important but challenging problem. Our goal is to combine two different physics model forecasting outputs with real measurements from an automated monitoring network so as to better predict solar power in a timely manner. To this end, we consider a new approach of analyzing large-scale multilevel models for computational efficiency. This approach features a division of the large-scale data set into smaller ones with manageable sizes, based on their physical locations, and fit a local model in each area. The local model estimates are then combined sequentially from the specified multilevel models using our novel bottom-up approach for parameter estimation. The prediction, on the other hand, is implemented in a top-down matter. The proposed method is applied to the solar energy prediction problem for the US Department of Energy's SunShot Initiative.

1.1 Introduction

Solar energy's contribution to the total energy mix is rapidly increasing. As the most abundant form of renewable energy resource, solar electricity is projected to supply 14% of the total demand of Contiguous United States by 2030, and 27% by 2050,

Y. Hwang (✉)
Department of Statistics, Sungkyunkwan University, Seoul, Korea
e-mail: yhwang@skku.edu

S. Lu
IBM Thomas. J. Watson Research Center, Yorktown Heights, NY, USA
e-mail: lus@us.ibm.com

J.-K. Kim
Department of Statistics, Iowa State University, Ames, IA, USA
e-mail: jkim@iastate.edu

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_1

respectively (Margolis et al. 2012). Having a high proportion of solar energy in the electric grid, however, poses significant challenges because solar power generation has inherent variability and uncertainty due to varying weather conditions (Denholm and Margolis 2007; Ela et al. 2011). Moreover, the uncertainty of solar power often obliges system operators to hold extra reserves of conventional power generation at significant cost. Accurate forecasting of solar power can improve system reliability and reduce reserve cost (Orwig et al. 2015; Zhang et al. 2015). Applying statistical methods on the forecasts from these numerical models can significantly improve the forecasting accuracy (Mathiesen and Kleissl 2011; Pelland et al. 2013).

Computer models have advanced beyond scientific research to become an essential part of industrial applications. Such expansions need a different methodological focus. To take advantage of the availability of such computer models, matching the model output with the historical observations is essential. This task is closely related to model calibration (Gramacy et al. 2015; Wong et al. 2016) to choose the optimal parameters for the computer model.

In this work, we consider a general framework to exploit the abundance of physical model forecasting outputs and real measurements from an automated monitoring network, using multilevel models. Our method addresses the aforementioned challenges for large-scale industrial applications. The proposed bottom-up approach has a computational advantage over the existing Bayesian method in computation for parameter estimation, because it does not rely on the Markov chain Monte Carlo (MCMC) method. Our approach is a frequentist based on the Expectation-Maximization (EM) algorithm.

1.2 Global Horizontal Irradiance

In this section, we describe our solar energy application and the overall problem. Our goal is to improve Global Horizontal Irradiance (GHI) prediction over the Contiguous United States (CONUS). GHI is the total amount of shortwave radiation received by a surface horizontal to the ground, which is the sum of Direct Normal Irradiance (DNI, the amount of solar radiation received by a surface perpendicular to the rays that come from the direction of the sun), Diffuse Horizontal Irradiance (DHI, the amount received by a surface that has been diffused by the atmosphere), and ground-reflected radiation. GHI forecast is of main interest of the participants in the electricity market.

To monitor the GHI, sensors are located over CONUS. The collected observations are obtained from the sensor locations marked on Fig. 1.1. The GHI readings are recorded at 1,528 locations in 15-min intervals. Hence, the data size grows very quickly; every day, thousands of additional observations are added. The data from each site are separately stored in the database indexed by the site location. The readings are obtained from various kinds of sensors, which may cause some potential variability among different locations. In our application, we consider two models to

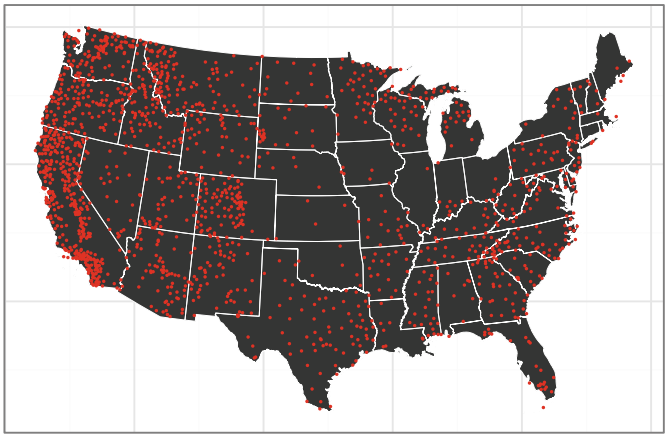


Fig. 1.1 The map of the 1,528 monitoring network locations, marked by dots

forecast GHI: Short-Range Ensemble Forecast (SREF, Du and Tracton 2001) and North American Mesoscale Forecast System (NAM, Skamarock et al. 2008). They share a common overall trend; however, there are certain discrepancies between the two model outputs. The model outputs are available at any location in a pre-specified computational domain, which covers the entire CONUS. The model output is stored at every hour, but can be matched with 15-min interval measurement data after post-processing.

1.3 Model

In this section, we present the basic setup and our proposed method. A model with three levels is considered in this paper, but the number of levels can be arbitrary.

1.3.1 Multilevel Model

Assume that the sensors are divided into H exhaustive and non-overlapping groups. For group h, measurements are collected at n_h sensors. From the ith sensor in group h, the measurements y_{hij} are available, as well as the output from computer models as the covariates $\mathbf{x}_{hij}$, for $j = 1, \ldots, n_{hi}$. Information at sensor or group level, $\mathbf{c}_h$ and $\mathbf{c}_{hi}$, is also available. Note that the covariates $\mathbf{x}$ are often more widely available than y_{hij}'s; in our application in Sect. 1.4, the computer model output is available not only at monitoring sites but also everywhere in the spatial domain of interest. We assume that n_h can be relatively small while n_{hi} is usually large, because managing

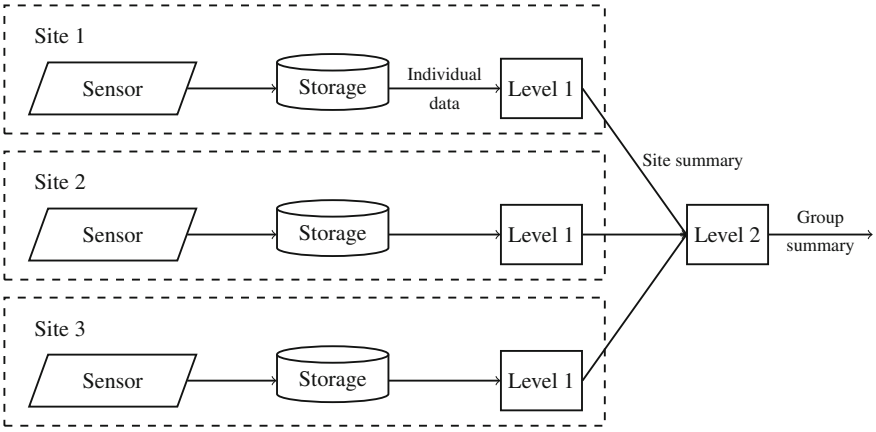


Fig. 1.2 Overall description of the data storage and modeling structure, where the data are stored separately for each site

the existing sensors and taking additional measurements from them usually do not cost much, while deploying new monitoring sensors often causes considerable cost.

Figure 1.2 shows the overall data storage and modeling structure of our proposed method to achieve these goals. Our so-called bottom-up approach builds up a hierarchy with the measurements by taking the following three steps.

The first step is *summarization*. There is no direct measurement for the kth level model ($k \geq 2$), so we use the observations from the lower level model to obtain a 'measurement' and construct an appropriate measurement model. The second step is *combination*; we combine the measurement model and structural model to build a prediction model using Bayes' theorem. The third step is *learning*, in which we estimate the parameters by using the EM algorithm. In the bottom-up approach, the computation for each step uses a summary version to ease the storage of data and spare the use of computer memory despite the large amount of data. In the subsection below, we describe each step in detail.

1.3.2 Bottom-Up Estimation

In this section, we give a detailed description of the estimation procedure. First, consider the level one and level two models,

$$\mathbf{y}_{hi} \sim f_1(\mathbf{y}_{hi}|\mathbf{x}_{hi}; \boldsymbol{\theta}_{hi}), \tag{1.1}$$

$$\boldsymbol{\theta}_{hi} \sim f_2(\boldsymbol{\theta}_{hi}|\boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h), \tag{1.2}$$

where $\mathbf{y}_{hi} = (y_{hi1}, \ldots, y_{hin_{hi}})^\top$ and $\mathbf{x}_{hi} = (\mathbf{x}_{hi1}^\top, \ldots, \mathbf{x}_{hin_{hi}}^\top)^\top$ are the observations and covariates associated with the ith sensor in the hth group for the level one model, respectively, and $\boldsymbol{\theta}_{hi}$ is the parameter in the level one model. In (1.2), $\boldsymbol{\theta}_{hi}$ is treated as a random variable and linked to the unit-specific covariate $\boldsymbol{c}_{hi}$ and parameter $\boldsymbol{\zeta}_h$ in the level two model.

To estimate $\boldsymbol{\zeta}_h$ in (1.2), we use the three-step approach discussed in Sect. 2.1. In the summarization step, for each sensor, we treat $(\mathbf{x}_{hi}, \mathbf{y}_{hi})$ as a single data set to obtain the best estimator $\hat{\boldsymbol{\theta}}_{hi}$ of $\boldsymbol{\theta}_{hi}$, a fixed parameter. Define $g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi})$ to be the density of the sampling distribution of $\hat{\boldsymbol{\theta}}_{hi}$. This sampling distribution is used to build a measurement error model, where $\hat{\boldsymbol{\theta}}_{hi}$ is a measurement for the latent variable $\boldsymbol{\theta}_{hi}$, while (1.2) is a structural error model for $\boldsymbol{\theta}_{hi}$.

The sampling distribution $g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi})$ is combined with the level two model f_2 to obtain the marginal distribution of $\hat{\boldsymbol{\theta}}_{hi}$. Thus, the MLE of the level two parameter $\boldsymbol{\zeta}_h$ can be obtained by maximizing the log-likelihood derived from the marginal density of $\hat{\boldsymbol{\theta}}_{hi}$. That is, we maximize

$$\sum_{i}^{n_h} \log \int g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h) d\boldsymbol{\theta}_{hi} \tag{1.3}$$

with respect to $\boldsymbol{\zeta}_h$, *combining* $g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi})$ with $f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)$. The maximizer of (1.3) can be obtained by

$$\hat{\boldsymbol{\zeta}}_h = \arg\max_{\boldsymbol{\zeta}_h} \sum_{i=1}^{n_h} \mathbb{E}\left[\log\{f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)\} \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h\right]. \tag{1.4}$$

Note that $\boldsymbol{\zeta}_h$ is the parameter associated with the level two distribution, and (1.4) aggregates the information associated with $\hat{\boldsymbol{\theta}}_{hi}$ to estimate $\boldsymbol{\zeta}_h$.

To evaluate the conditional expectation in (1.4), we derive

$$p_2(\boldsymbol{\theta}_{hi} \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h) = \frac{g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)}{\int g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h) d\boldsymbol{\theta}_{hi}}. \tag{1.5}$$

The level two model can be *learned* by the EM algorithm. Specifically, at the tth iteration of EM, we update $\boldsymbol{\zeta}_h$ by

$$\hat{\boldsymbol{\zeta}}_h^{(t)} = \arg\max_{\boldsymbol{\zeta}_h} \sum_{i=1}^{n_h} \mathbb{E}\left[\log\{f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)\} \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h = \hat{\boldsymbol{\zeta}}_h^{(t-1)}\right], \tag{1.6}$$

where the conditional expectation is with respect to the prediction model in (1.5) evaluated at $\hat{\boldsymbol{\zeta}}_h^{(t-1)}$, which is obtained from the previous iteration of the EM algorithm.

When $\hat{\boldsymbol{\theta}}_{hi}$ is the maximum likelihood estimator, we may use a normal approximation for $g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi})$. Asymptotically, $\hat{\boldsymbol{\theta}}_{hi}$ is a sufficient statistic for $\boldsymbol{\theta}_{hi}$ and

normally distributed with mean $\boldsymbol{\theta}_{hi}$ and the estimated variance $\{I_{1hi}(\boldsymbol{\theta}_{hi})\}^{-1}$, where $\{I_{1hi}(\boldsymbol{\theta}_{hi})\}^{-1}$ is the observed Fisher information derived from g_1.

Once each $\hat{\boldsymbol{\zeta}}_h$ is obtained, we can use $\{\hat{\boldsymbol{\zeta}}_h; h = 1, \ldots, H\}$ as the summary of observations to estimate the parameters in the level three model. Let the level three model be expressed as

$$\boldsymbol{\zeta}_h \sim f_3\left(\boldsymbol{\zeta}_h | \boldsymbol{c}_h; \boldsymbol{\xi}\right), \tag{1.7}$$

where $\boldsymbol{c}_h$ are the covariates associated with group h and $\boldsymbol{\xi}$ is the parameter associated with the level three model. Estimation can be done in a similar fashion to the level two parameters. However, $\boldsymbol{\zeta}_h$ is now treated as a latent variable, and $\hat{\boldsymbol{\zeta}}_h$ as a measurement. Similar to (1.3), we maximize

$$\sum_{h=1}^{H} \log \int g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3\left(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \boldsymbol{\xi}\right) d\boldsymbol{\zeta}_h \tag{1.8}$$

with respect to $\boldsymbol{\xi}$ to obtain $\hat{\boldsymbol{\xi}}$, where $g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h)$ is the sampling distribution of $\hat{\boldsymbol{\zeta}}_h$, which is assumed to be normal. The EM algorithm can be applied by iteratively solving

$$\hat{\boldsymbol{\xi}}^{(t)} = \arg\max_{\boldsymbol{\xi}} \sum_{h=1}^{H} \mathbb{E}\left[\log\left\{f_3\left(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \boldsymbol{\xi}\right)\right\} \mid \hat{\boldsymbol{\zeta}}_h; \boldsymbol{\xi} = \hat{\boldsymbol{\xi}}^{(t-1)}\right], \tag{1.9}$$

where the conditional distribution is with respect to the distribution with density

$$p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h; \boldsymbol{\xi}) = \frac{g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3\left(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \boldsymbol{\xi}\right)}{\int g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3\left(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \boldsymbol{\xi}\right) d\boldsymbol{\zeta}_h}$$

evaluated at $\boldsymbol{\xi} = \hat{\boldsymbol{\xi}}^{(t-1)}$. The level three model can be chosen flexibly depending on the usage, as it was in the lower levels.

1.3.3 Top-Down Prediction

In this section, we describe the prediction procedure. In contrast to the bottom-up approach of Sect. 1.3.2, the prediction is made in a top-down fashion.

To describe the top-down approach to prediction, consider the three-level models in (1.1), (1.2), and (1.7). The bottom-up estimation in Sect. 1.3.2 provides a way of estimating the parameters, $\boldsymbol{\theta}_{hi}$, $\boldsymbol{\zeta}_h$, and $\boldsymbol{\xi}$ by $\hat{\boldsymbol{\theta}}_{hi}$, $\hat{\boldsymbol{\zeta}}_h$, and $\hat{\boldsymbol{\xi}}$, respectively, using EM algorithm or maximizing the marginal likelihood.

Our goal is to predict unobserved y_{hij} values from the above models using the parameter estimates. The goal is to generate Monte Carlo samples of y_{hij} from

$$p(y_{hij} \mid \mathbf{x}_{hij}; \hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) = \frac{\int\int f_1(y_{hij} \mid \mathbf{x}_{hij}; \boldsymbol{\theta}_{hi}) p_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{\zeta}_h, \hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) d\boldsymbol{\zeta}_h d\boldsymbol{\theta}_{hi}}{\int\int\int f_1(y_{hij} \mid \mathbf{x}_{hij}; \boldsymbol{\theta}_{hi}) p_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{\zeta}_{hi}, \hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) d\boldsymbol{\zeta}_h d\boldsymbol{\theta}_{hi} dy_{hij}} \tag{1.10}$$

where $p_2(\boldsymbol{\theta}_{hi} \mid \hat{\boldsymbol{\theta}}_{hi}, \boldsymbol{\zeta}_h, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) = p_2(\boldsymbol{\theta}_{hi} \mid \hat{\boldsymbol{\theta}}_{hi}, \boldsymbol{\zeta}_h)$ and $p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}})$ are the predictive distribution of $\boldsymbol{\theta}_{hi}$ and $\boldsymbol{\zeta}_h$, respectively.

To generate Monte Carlo samples from (1.10), we use the top-down approach. We first compute the predicted values of $\boldsymbol{\zeta}_h$ from the level three model,

$$p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) = \frac{g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \hat{\boldsymbol{\xi}})}{\int g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \hat{\boldsymbol{\xi}}) d\boldsymbol{\zeta}_h}, \tag{1.11}$$

where $g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h)$ is the sampling distribution of $\hat{\boldsymbol{\zeta}}_h$. Also, given the Monte Carlo sample $\boldsymbol{\zeta}_h^*$ obtained from (1.11), the predicted values of $\boldsymbol{\theta}_{hi}$ are generated by (1.5). The best prediction for y_{hij} is

$$\hat{y}_{hij}^* = \mathbb{E}_3 \left[\mathbb{E}_2 \left\{ \mathbb{E}_1(y_{hij} \mid \mathbf{x}_{hij}, \boldsymbol{\theta}_{hi}) \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h \right\} \mid \hat{\boldsymbol{\zeta}}_h; \hat{\boldsymbol{\xi}} \right] \tag{1.12}$$

where subscripts 3, 2, and 1 denote the expectation with respect to p_3, p_2, and f_1, respectively. Thus, while the bottom-up approach to parameter estimation starts with taking the conditional expectation with respect to p_1 and then moves on to p_2, the top-down approach to prediction starts with the generation of Monte Carlo samples from p_2 and then moves on to p_1 and f_1.

To estimate the mean-squared prediction error of $\hat{y}_{hij}^*$ given by $M_{hij} = \mathbb{E}\{(\hat{y}_{hij}^* - y_{hij})^2\}$, we can use the parametric bootstrap approach (Hall and Maiti 2006; Chatterjee et al. 2008). In the parametric bootstrap approach, we first generate bootstrap samples of y_{hij} using the three-level model as follows:

1. Generate $\boldsymbol{\zeta}_h^{*(b)}$ from $f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \hat{\boldsymbol{\xi}})$, for $b = 1, 2, \ldots, B$.
2. Generate $\boldsymbol{\theta}_{hi}^{*(b)}$ from $f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h^{*(b)})$, for $b = 1, 2, \ldots, B$.
3. Generate $y_{hij}^{*(b)}$ from $f_1(y_{hij} \mid \mathbf{x}_{hij}; \boldsymbol{\theta}_{hi}^{*(b)})$, for $b = 1, 2, \ldots, B$.

Once the bootstrap samples of $\mathbf{Y}^{*(b)} = \{y_{hij}^{*(b)}; h = 1, 2, \ldots, H; i = 1, \ldots, n_h; j = 1, \ldots, m_{hi}\}$ are obtained, we can treat them as the original samples and apply the same estimation and prediction method to obtain the best predictor of y_{hij}. The mean-squared prediction error (MSPE) M_{hij} can also be computed from the bootstrap sample. That is, we use

$$\hat{M}_{hij} = \mathbb{E}_*\{(\hat{y}_{hij}^* - y_{hij})^2\}$$

to estimate M_{hij}, where $\mathbb{E}_*$ denote the expectation with respect to the bootstrapping mechanism.

1.4 Prediction of Global Horizontal Irradiance

In this section, we give a detailed description of the available data and the model that we use. We apply the proposed model and compare results to those of the comparators.

1.4.1 Data Description

We use 15 days of data for our analysis (12/01/2014–12/15/2014). There are 1528 sites to monitor GHI, where the number of available data varies between 12 and 517 observations, and the total number of observations is 557,284. To borrow strength from neighboring sites, we formed 50 groups that are spatially clustered by applying the K-means algorithm on the geographic coordinates. We assume the sites belonging to the same group are homogeneous. The number of sites in each group, n_h, varies between 10 and 59. Depending on the goal, one can use other grouping schemes such as the distribution zone described in (Zhang et al. 2015). Calculated irradiance is available at every 0.1 degree and is matched to the monitoring site location.

Since we are interested in the amount of irradiance, we first exclude zeros from both observed measurements and computer model outputs for the analysis. Thus, all values are positive and skewed to the right, and we used the logarithm transformation for both predictors and responses. Hereinafter, all variables are assumed to be log-transformed.

1.4.2 Model

This section presents the model that we used in the data analysis in detail. Let y_{hij} be the jth measurement for the ith sensor in the hth group. Following the multilevel modeling approach described in Sect. 1.3, we first assume that the measurement y_{hij} follows

$$y_{hij} = \mathbf{x}_{hij}\boldsymbol{\theta}_{hi} + e_{hij}, \tag{1.13}$$

with a latent site-specific parameter $\boldsymbol{\theta}_{hi}$, where the covariates $\mathbf{x}_{hij}$ has NAM and SREF model output as predictors including an intercept term, and $e_{hij} \sim t(0, \sigma^2_{hi}, \nu_{hi})$, where σ^2_{hi} is scale parameter and ν_{hi} are the degree of freedom (Lange et al. 1989).

The degrees of freedom are assumed to be five in the analysis, but it also can assumed to be unknown and estimated by the method of (Lange et al. 1989). Assume that the level two model follows

$$\boldsymbol{\theta}_{hi} \sim N(\boldsymbol{\beta}_h, \boldsymbol{\Sigma}_h), \tag{1.14}$$

for some group-specific parameters $\boldsymbol{\beta}_h = (\beta_{h1}, \ldots, \beta_{hp})$ and $\boldsymbol{\Sigma}_h$. For further presentation, define the length H vector of jth coefficients of $\boldsymbol{\beta}_h$ concatenated over H groups

$$\boldsymbol{\beta}_{(j)} = (\beta_{1j}, \ldots, \beta_{Hj}),$$

and similarly define $\hat{\boldsymbol{\beta}}_{(j)}$. The subscript j is omitted hereinafter as we model each parameter separately but in the same manner. To incorporate the spatial dependence that may exist in the data, we assume that the level three model follows

$$\boldsymbol{\beta} \sim N(\boldsymbol{F}\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1.15}$$

where $\boldsymbol{F}$ is a pre-specified H by q model matrix, and $\boldsymbol{\mu}$ is the mean parameter of length q. In the analysis in Sect. 1.4.3, $\boldsymbol{F}$ is chosen to be $\mathbf{1}$, length H vector of 1's and a scalar μ. The spatial covariance $\boldsymbol{\Sigma}$ has its (k, l) th element

$$\Sigma_{kl} = \text{cov}(\beta_k, \beta_l) = \tau^2 \exp(-\rho d_{kl}),$$

where d_{kl} is the distance between the groups. The distance between two groups is defined to be the distance between the centroids of groups. The estimated spatial effect for two coefficients is depicted in Fig. 1.3. Note that a group is formed by collapsing several neighboring sites; hence, the number of groups is less than that of sites. This also reduces the computational burden because the main computation in our spatial model is associated with the number of spatial locations. Hence, it is helpful to introduce the spatial components in the group level instead of the sensor level to provide computational benefit.

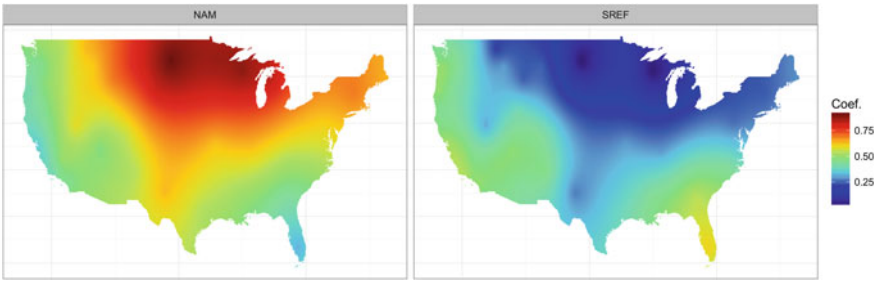


Fig. 1.3 Spatial variation of the group-level coefficients from the second level for two computer models, where the left panel shows the NAM model and the right panel the SREF model

1.4.3 Results

This section presents the data analysis result. Under the linear regression model in (1.13), the best prediction is $\hat{y}^*_{hij}$ in (1.12). We compared the multilevel approach with two other modeling methods: (1) site-by-site model: fit a separate model for each individual site; (2) global model: fit a single model for all sensor locations using the aggregate data combining all sensors. To evaluate the prediction accuracy, we conducted tenfold cross-validation. The data set is randomly partitioned into 10 subsamples. Of these 10 subsamples, one subsample was held out for validation, while the remaining nine subsamples are used to fit the model and obtain predicted values. The cross-validation process is repeated for each fold.

We considered two scenarios: (a) prediction made at observed sites and (b) prediction made at new sites. For scenario (a), we partitioned the time point into ten subperiods, while for (b) the sites into ten subregions.

We compare the accuracy of different methods by the root-mean-squared prediction error (RMSPE), $\{N^{-1}\sum_j (y_{hij} - \hat{y}_{hij})^2\}^{1/2}$, with N being the size of the total data set. Table 1.1 presents the overall summary statistics for the accuracy of each method, calculated from cross-validation. The standard deviation calculated over the subsamples is in parentheses.

The rightmost column shows the overall accuracy. The global model suffers because it cannot incorporate the site-specific variation. On the contrary, the site model suffers from reliability issues for some sites because it does not use the information from neighboring sites. The multilevel approach strikes a fine balance between flexibility and stability. For a comprehensive comparison of each method, we evaluate the accuracy measure divided by the number of available data points for each site. As noted earlier, some stations may suffer from the data reliability problem. As such, the available sample size can vary from station to station, which affects the site-by-site model. When the prediction is made based on few available samples due to the data reliability issues, the inference can be unstable, affecting the accuracy of the prediction. The multilevel method can utilize information from other sites belonging to the same group, so it is particularly beneficial for locations with smaller sample sizes.

Table 1.1 Root-mean-squared prediction error comparison of the different modeling methods, divided by the size of the training sample and overall

Training sample size			
Method	<200	≥ 200	Overall
Multilevel	0.678 (0.129)	0.591 (0.052)	0.594 (0.055)
Site	1.344 (0.764)	0.593 (0.073)	0.632 (0.133)
Global	0.646 (0.038)	0.639 (0.009)	0.639 (0.009)

1.5 Conclusion

With the advances in remote sensing and storage technology, data are now collected over automated monitoring networks at an unprecedented scale. A simple yet efficient modeling approach that can reliably handle such data is of great need.

In this paper, we have developed a general framework using a multilevel modeling approach, which utilizes monitoring data collected to manage a large-scale system. It is presented with a solar energy application, although it can be flexibly modified to incorporate the data structure or overall goal. The computation can be automated with deterministic criteria and be easily distributed. It has been shown that the method can provide improved inference compared to naive approaches. Our methodology can also be extended to incorporate discrete measurements.

Acknowledgements This report was prepared as an account of work sponsored by an agency of the US government. Neither the US government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represented that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the US government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the US government or any agency thereof.

References

Chatterjee, S., Lahiri, P., & Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *The Annals of Statistics*, *36*, 1221–1245. (06)

Denholm, P., & Margolis, R. M. (2007). Evaluating the limits of solar photovoltaics (pv) in traditional electric power systems. *Energy Policy*, *35*, 2852–2861.

Du, J., & Tracton, M. S. (2001). Implementation of a real-time shortrange ensemble forecasting system at ncep: An update. In *Ninth Conference on Mesoscale Processes*, Preprints, Ninth Conference on Mesoscale Processes, Fort Lauderdale, FL., American Meteorological Society.

Ela, E., Milligan, M., & Kirby, B. (2011). Operating Reserves and Variable Generation. NREL/TP-5500-51978. http://www.nrel.gov/docs/fy11osti/51978.pdf.

Gramacy, R. B., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuranz, C. C., Rutter, E., et al. (2015). Calibrating a large computer experiment simulating radiative shock hydrodynamics. *The Annals of Applied Statistics*, *9*(3), 1141–1168.

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*, 221–238.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the T distribution. *Journal of the American Statistical Association*, *84*, 881–896.

Margolis, R., Coggeshall, C., & Zuboy, J. (2012). Integration of solar into the U.S. electric power system. In *SunShot vision study*. Washington, DC: U.S. Department of Energy.

Mathiesen, P., & Kleissl, J. (2011). Evaluation of numerical weather prediction for intra-day solar forecasting in the continental united states. *Solar Energy*, *85*, 967–977.

Orwig, K., Ahlstrom, M., Banunarayanan, V., Sharp, J., Wilczak, J., Freedman, J., et al. (2015). Recent trends in variable generation forecasting and its value to the power system. *IEEE Transactions on Sustainable Energy*, *99*, 1–10.

Pelland, S., Galanis, G., & Kallos, G. (2013). Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model. *Progress in Photovoltaics: Research and Applications*, *21*, 284–296.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., et al. (2008). A description of the advanced research WRF version 3. NCAR TECHNICAL NOTE: NCAR/TNñ475+STR. National Center for Atmospheric Research, Boulder, Colorado, USA.

Wong, R. K. W., Storlie, C. B., & Lee, T. C. M. (2016). A frequentist approach to computer model calibration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. (To appear)

Zhang, J., Florita, A., Hodge, B.-M., Siyuan, L., Hamann, H. F., Banunarayanan, V., et al. (2015). A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, *111*, 157–175.

Zhang, J., Hodge, B.-M., Siyuan, L., Hamann, H. F., Lehman, B., Simmons, J., et al. (2015). Baseline and target values for regional and point PV power forecasts: Toward improved solar forecasting. *Solar Energy*, *122*, 804–819.

Chapter 2
The 62% Problems of SN Ratio and New Conference Matrix for Optimization: To Reduce Experiment Numbers and to Increase Reliability for Optimization

Teruo Mori

Abstract Robust design has been widely adopted during product design to reduce variation and improve quality. However, based on our survey of 171 published case studies using the L_{18} orthogonal array in Japan, 62% of the signal-to-noise ratios (SN) of the optimal design cases concluded from the main effects plots were worse than the best combinations of the existing 18 runs of the L_{18} orthogonal array. This means that current robust design based on SN ratios and the L_{18} cannot predict the optimal conditions accurately and needs further work to improve the analytical prediction accuracy and optimization efficiency. We will show the six causes of 62% problems. Now, we have understood to face the serious problems like global warming, food amounts for increasing population. We need faster and more precise methodology for researching them, and it will be able to reduce experiment numbers and to increase reliability using conference matrix.

2.1 Introduction

The job range of engineers' assignments is wide and includes the basic research on invention and new product development through the improvement of current products and improvement of production processes, etc. They need to meet development goals and to find optimal conditions to reduce product and process performance variation at the same time (Mori 2011, 2009, 1992).

It is too late to conduct the troubleshooting activities to change product design or production conditions to resolve product defect issues after those products are

T. Mori (✉)
The Mori Consulting Office, 871-3 Daitocho, Fujieda 426-0044, Japan
e-mail: tm551017@yahoo.co.jp

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_2

manufactured and shipped to the market and to customers. It is common to use the recall and warranty activities to resolve quality problem issues. Also, it will be happened at a loss to customers. Of course, company guarantees the product and service quality as top priority and is willing to take action to reduce customers' loss due to defective products (Mori 2014).

In this paper, we will review first the problems of the current robust design. Then, we will show to expect the new conference matrix for optimization methods.

2.2 Verification Assessment to Confirm the Optimal Condition to Exceed the Best of L_{18} Trials

After finding the optimal candidate condition, engineers will conduct experiments to confirm and verify that the results of the optimal candidate condition are reproducible. Table 2.1 shows the SN ratio results (Mori 2013) (a) of the optimal conditions for 171 case studies to compare the best SN ratio values (b) of the L_{18} trials. One hundred and six (62%) cases were a<b. Theoretically, the SN ratio results of (a) are as good as or better than (b), because the optimal condition candidates (a) are chosen from many more possible combinations of factor levels than (b).

Unfortunately, 62% of the optimal conditions of (a) are worse than the best values of (b) as illustrated in Table 2.1 (Japan Quality Engineering association 2003–2012).

Engineers who have been trained in statistical modeling of Taguchi may be surprised at the "prediction uncertainty of the optimal design candidates" shown in Table 2.1. Then, they will be requested the more advanced mathematical analysis for improving the prediction accuracy and reduce the uncertainty of the optimal design solutions.

2.3 Investigating the Root Causes for 62% Problems

Assume that the mean of the output response is μ; and that the main effects of four selected experimental factors (A, B, C, and D) are a, b, c, and d.

The interaction terms are expressed using a multiplication term such as ab, ac,…abc…, and abcd for the four factors. Let the summation of experimental error and measurement error be (e). The experimental output response (y) is expressed with the mean value μ, main effects, a, b, c, and d, and the quadratic, interaction and higher terms as shown here:

Experimental output response (Jeff Wu and Hamada 2009)

$y = \mu + a + b + c + d + aa + ab + ac + \cdots + cd + abc + \cdots + bcd + abcd + (e)$

= Mean value + main effects + quadratic + interaction effects + higher terms + error

Table 2.1 Optimal condition comparison analysis

QES	Total # of case studies	# of case studies where (a<b)
2012	6	3
2011	9	6
2010	15	8
2009	7	6
2008	20	9
2007	14	9
2006	33	24
2005	23	11
2004	22	18
2003	22	12
Total	171	106
(%)	62.0	
QES→	Japan quality engineering symposium	

The optimal conditions (a) of Table 2.1 are selected based on the level averages which were calculated to divide the sum of response y with data numbers. The response graphs are made with the level averages.

On the other hand, orthogonal array tables like L_8, L_9, L_{18} as the design matrix have the linear effect structure, so that it will be expected that the response should consist of linearity components. If the response has nonlinear effect, the response graph will be shifted from the original by contaminating nonlinear effect. Nonlinear effect will consist of quadratic, and interaction between factors and higher other terms. We can estimate the nonlinear effects in the response with the empty column of the orthogonal array table. So, we tried to look for the nonlinear effect from start to finish of the robust process.

2.4 Causes Analysis for 62% Problems

We have done to analyze the causes of 62% problems related to nonlinear effects. We finally detected six nonlinear effects at robust process on Fig. 2.1. It has the marked ①–⑥ on nonlinear in Fig. 2.1.

We will introduce ①–⑥ at Fig. 2.1 to explain the complex mathematics background to get contaminated with nonlinear effects.

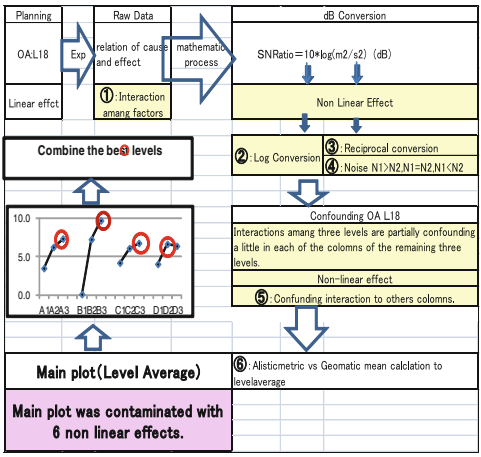


Fig. 2.1 Six nonlinear effects on robust design process

2.5 Multiple Contamination of Six Type of Nonlinear Effect

Six types of nonlinear effects were separately investigated as the cause of 62% problem. Actual optimum cases will be contaminated single or multiple of them. We cannot detect the real causes individually if columns were filled with factors.

However if there were empty columns in orthogonal array tables, we can make a diagnose the degree of contamination the with the empty column factor effects.

We selected the published typical three case studies with empty columns for SN ratio. We showed Figs. 2.2(BGA), 2.3(circuit), 2.4(straw) (Tanabe 2016).

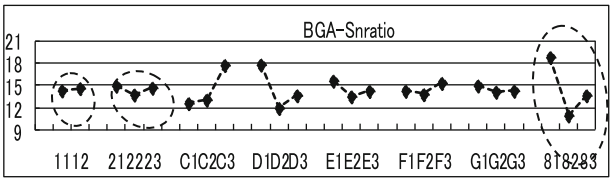


Fig. 2.2 BGA semiconductor structure

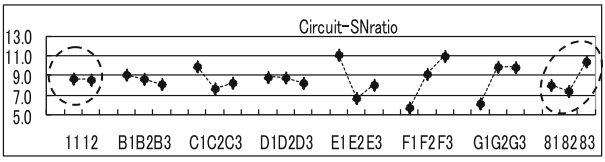


Fig. 2.3 Electro circuit case

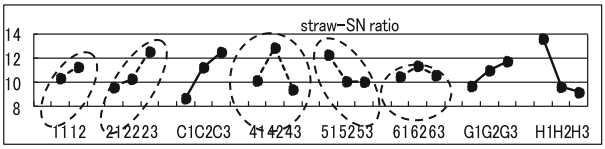


Fig. 2.4 Education training case

At Fig. 2.2 case, the empty column 8 is the largest effect. At Fig. 2.3, this case has six factors, the empty column 8 is fourth largest factor, and it is at middle class position. At Fig. 2.4, this case 3 factor and empty columns are 5 factors. The largest empty column 4 is in second largest position. The empty columns effects might be lower than the layouted columns. Those empty columns are gotten the multiple contaminations of six nonlinear effects. We cannot ignore such six nonlinear effects for contamination to other columns. The current robust design process will not be reasonable for engineering subjects. At least, SN ratio and L_{18} should be replaced the different way to get more reliability of optimization.

2.6 The New Design Matrix for the Next Generation

We started to create the new process to avoid six types of nonlinear effects for robust design. L_{18} has been long time recommended as standard tool at current robust design. Taguchi (1984) explored statistically the optimum condition like black box. The current robust design was not supported that reason is to demand too much experiment number like 108 at dynamic style. If statistician likes to recommend directly catching the optimal solution itself, the trial number might be naturally increased to avoid missing it. If a fisherman tries to catch a fish, he will select the net as much as largest size with the finest mesh to fear failures.

The scientist now is shifting the investing philosophy to confirm the tendency of factor effect to target in mix conditions with design matrix to make the less number trials and higher reliability. The best deign matrix might be "conference matrix." Conference matrix C_4 $(2^1 3^3)$ is the minimum number for three-level matrix with

Table 2.2 $C_4(2^1 3^3)$ and linear term of $L_9(3^4)$

C_4	1	2	3	4
1	0	1	1	1
2	−1	0	−1	1
3	−1	1	0	−1
4	−1	−1	1	0
*	3	0	0	0
		3	0	0
			3	0
				3

L_9	1	2	3	4
1	−1	−1	−1	−1
2	−1	0	0	0
3	−1	1	1	1
4	0	−1	0	1
5	0	0	1	−1
6	0	1	−1	0
7	1	−1	1	0
8	1	0	−1	1
9	1	1	0	−1
*	6	0	0	0
		6	0	0
			6	0
				6

linear term. We compare linear term with $L_9(3^4)$ in Table 2.2. (*) is the sum of product of columns to confirm the orthogonality (Tanaka 2016).

The new process may not use SN ratio with log conversion to avoid six nonlinear effects, and we are testing to adapt the raw data themselves. It will be complete in 2017.

2.7 Conclusion

In this paper, author introduced 62% problems to the current robust design with L_{18} and SN ratio.

Based on author survey, 171 published case studies using the L_{18} orthogonal array (OA) in Japan, and 62% of the signal-to-noise ratios (SN) of the optimal design cases concluded from the main effects plots were worse than the best combinations of the 18 runs of the L_{18} orthogonal array.

Also, author detected six types of nonlinear effects.

①: Interaction among factors: ②: Log conversion for response

③: Reciprocal structure of SN ratio: ④: Diversion (S^2) size of SN ratio

⑤: Confounding type $L_{18}(2^1 3^7)$: ⑥: Geometric level average after log conversion.

Also, author is touching the conference matrix and new concept for new type robust design.

References

Japan Quality Engineering association, *Proceeding* (2003–2012).

Jeff Wu, C. F., & Hamada, M. S. (2009). *Experiments: Planning, analysis, and optimization*., Wiley series in probability and statistics New York: Wiley.

Mori, T. (1992). *Methods for new product and new technology development, trend book*.

Mori, T. (2009). *Taguchi methods- pocket guide book, trend book*.

Mori, T. (2011). *Taguchi methods*: ASME.

Mori, T. (2012). *QES2012, the 20th Annual Proceeding,* Paper No. 63 (Quality Engineering Symposium).

Mori, T. (2013). *QES2013, the 21th Annual Proceeding,* Paper No. 23 (Quality Engineering Symposium).

Mori, T. (2014). *Technical report, Toyota Bousyoku*, vol. 08, pp. 8–19.

Mori, T. (2015). *Mathematic and application of near orthogonal array, Mori Office, Chap. 23*.

Taguchi, G. (1984). *Parameter design for new product development*. Japan: Japan Standards Association.

Tanabe, S. (2016). *QES2016 The 24 Proceeding* (Quality Engineering Symposium) No. 87.

Tanaka, K. (2016). Quality (JSQC) 46(1), 51–54.

Part II
Analytic Methods of Big Data

Chapter 3
Possible Clinical Use of Big Data: Personal Brain Connectomics

Dong Soo Lee

Abstract The biggest data is brain imaging data, which waited for clinical use during the last three decades. Topographic data interpretation prevailed for the first two decades, and only during the last decade, connectivity or connectomics data began to be analyzed properly. Owing to topological data interpretation and timely introduction of likelihood method based on hierarchical generalized linear model, we now foresee the clinical use of personal connectomics for classification and prediction of disease prognosis for brain diseases without any clue by currently available diagnostic methods.

3.1 Introduction

Big data and its handling require refined statistics for clinical application. Examples are (1) physiological monitoring data which can be acquired using smartphone and recently developed soft bioelectronics sensors (Park et al. 2015; Gao et al. 2016), (2) genomics and epigenomics predicting individual's disease predisposition (Rehm et al. 2015) or guiding the N of 1 study using metagenomics (Lillie et al. 2011) or pharmacogenomics for drug selection or avoidance (Relling et al. 2015), and (3) brain connectivity data classifying and predicting the prognosis.

Use of lifelog physiological signals for clinical purposes mandates (1) the introduction of data storage such as cloud and (2) easy mining of the valuable information therein and (3) sleek input and output from the cloud storage. Cheaper and readily available methods are now to be developed for multi-omics data to be used for clinical

D. S. Lee (✉)
Department of Nuclear Medicine, Department of Molecular Medicine and Biopharmaceutical Sciences, College of Medicine, Seoul National University (SNU) and SNU Hospital, Seoul, Korea
e-mail: dsl@plaza.snu.ac.kr

D. S. Lee
Korean Brain Research Institute, Daegu, Korea

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_3

purposes. In contrast, clinical use of brain connectivity data needs a fresh viewpoint other than mapping or graph representation of brain.

3.2 Use of Brain Images as Topography for Clinical Decision

In order to use brain images for classification and prediction of brain diseases, which are yet defined not on these images but only on the basis of psychopathology, scientists had established statistical parametric mapping (SPM) to localize the regional abnormality, and its significance was inferred by strict statistics (Lee et al. 2001). SPM had been used to find the areas of abnormal activity in disease compared with normal controls. The activity ranged from T1 density representing tissue density on T1 MRI to glucose metabolism on fluorodeoxyglucose (FDG) PET. However, we find mostly no abnormality on brain images even if we do refined topographical analysis in many types of brain diseases. Examination of interregional connectivity came to be considered as an alternative, and brain connectivity was defined as brain graphs represented by the correlational activity of many brain regions (Lee et al. 2008).

3.3 Brain Graphs and Inherent Barrier Against Clinical Use

Brain graphs consist of nodes (brain regions or voxels) and edges (connections). Brain graphs were analyzed by the newly introduced method of algebraic topology, especially persistent homology (Zomorodian and Carlsson 2005; Singh et al. 2008). Graphs were filtered by simplicial chain complex filtration. Topological invariants were looked for from this filtration, which yielded barcodes of Betti-0 (Lee et al. 2012). Betti-0 is counting the number of connected components and is looking at the zero-dimensional topological invariant. Recently, this analysis went on to the synthesis of information from multimodal measurements on T1 MRI and FDG PET, which is called multidimensional persistent homology (Lee et al. 2017).

Brain connectivity data had long been analyzed by graph theory (Bullmore and Sporns 2009; Rubinov et al. 2010). Both node and edge data were used for classification according to the classic graph theory (Mucha et al. 2010; Ahn et al. 2010). However, the many parameters of graph theory could not easily elucidate the characteristics of brain graphs, and moreover the representative global and nodal parameters could not be easily compared between diseased and healthy groups, notwithstanding statistical inference. From sample data of two groups, single matrices represented each group for their connectivity. Thus, the statistical inference should have been performed against distribution of pseudorandom matrices acquired by permutation

or bootstrapping of sample data as was used previously in the comparison of tensor maps from population brain data (Thompson et al. 2000). This is in contrast to the successful SPM application of finding regional abnormality using distribution assumptions of Gaussian, chi-square, or F (Worsley et al. 2004). Previously when the investigators used only the regional distribution to speculate connectivity based on SPM analysis, they observed the distribution of regional activity and did statistical inference to find the significant difference of functional connectivity between groups (Worsley et al. 2005).

Scientific community moved on even to adopt Granger causality (Roebroeck et al. 2005). This was based on the vector autoregressive modeling and merged initial structural equation modeling or path analysis which did not gain popular use, partly because of the intimidating complexity of paths and the combination of plausible components. Apparent arbitrariness to choose the variables which were to be put into the models also prevented the propagation of usage when they tried to delineate the disease-specific abnormality of causal connections.

3.4 Brain graphs' Arbitrariness as a Bottleneck Against Clinical Use

The connectivity data of the brain structural and functional images could then be converted to interregional correlation matrices. Voxel-based matrix could yield dozens of thousands rows and columns which should be sparsified with various methods (Lee et al. 2011; Batson et al. 2013; Xie et al. 2017) to enhance the feasibility of data handling. Thresholding was the easiest way to make the matrices sparse. Once the adjacency matrix was acquired, the data were examined for visualization or comparison with norms by taking these data mostly as binary matrices and sometimes as weighted matrices. The choice of threshold and binary/weighted matrices was up to the investigators. This arbitrariness was challenged to cause a critical problem in analysis, as the investigators were changing thresholds, difference between diseased and healthy groups was found significant in some but not in other thresholds (Bassett et al. 2012). This arbitrariness was tried to be ameliorated by setting the number of nodes, i.e., sparsity (Kim et al. 2014). However, any investigator can or will not find the difference to their own will between their groups of interest and the controls. This has been the serious cause of flaw, and we tried to solve this problem from the root (Lee et al. 2012).

3.5 Topological Framework as a Fundamental Solution to the Above Problems

The solution of the problem of arbitrariness was derived from the idea that topological invariant might represent the characteristics of brain graph. This topological concept is based upon persistent homology during the filtration of simplicial chain complex. Modulus of image of the 1-higher dimensional boundary function over kernel of its own dimension represented the topological invariant, which will make a group changing its connected components according to the changing thresholds, i.e., simplicial complex filtration. The changing number of connected components is known to make a barcode as topological invariant of 0-dimension, Betti-0 (Lee et al. 2012).

The next problem was to define the summary matrix for the barcodes, and this was easily done by producing single linkage matrix, which does not consider the merged nodes as separate, which is basically the idea of topology. The uniqueness of the brain graph different from others resides in the fact that brain nodes cannot be shuffled and then the barcode should in fact be the dendrogram. Dendrogram and barcode of persistent homology were exactly equivalent, and single linkage matrix was the matrix representation of these two. Then we just needed to handle the single linkage matrix for further analysis. In usual conditions, single linkage matrix is also equivalent to minimum spanning tree (Lee et al. 2012). From disease and healthy control groups, we now had single matrices and we needed to compare these two single matrices with each other statistically.

3.6 Statistical Inference Using Pseudorandom Data Generated by Permutation

The last problem was to develop the statistical inference methods for comparison of global difference between single linkage matrices and for comparison of the local difference yielding which edges were statistically significantly different. For finding global difference between matrices, we adopted the metric such as Gromov–Hausdorff distance (Lee et al. 2011), bottleneck distance, or Wasserstein distance. Permutation of two groups of disease and controls could easily make 5,000 or 10,000 pseudorandom groups. The above distances were calculated using these pseudorandom groups to make a distribution. The distances between the observed two groups and the difference of these distances were compared with the distribution of the differences of pseudorandom groups to designate p value of type I error. This method was used in all the articles we published upon metabolic connectivity on FDG PET (Lee et al. 2012, 2017; Choi et al. 2014; Im et al. 2016), T1 density connectivity on T1 MRI (Kim et al. 2014), activation fMRI connectivity (Kim et al. 2015), source power connectivity on magnetoencephalography (MEG) (Hahm et al. 2017), and so on.

3.7 Finding the Edges Explaining the Difference Between Groups

Finding the edges of significant difference between disease (or activation) and control (or baseline) groups was solved the same way that distribution of values of all the cells in the single linkage matrices could be depicted, and thus, observed values of the cells were compared with the edge value of 10,000 single linkage matrices calculated using the permuted pseudorandom data. Bonferroni correction was impossible to apply as the number of observation was too large (voxel number, source number, or number of regions-of-interest). Type II error of not finding true positives was also a concern. If we set the N nodes, and when we compared the values (FDG uptake on PET, T1 density on T1 MRI, BOLD signal on fMRI, and power of specific time and frequency on wavelet-transformed MEG signals at the sources), the node activity was a vector ($N \times 1$), however, when we considered connectivity, whichever the matrix, either correlation, partial correlation, or single linkage matrix, was used for further analysis, the data dimension rose to $N \times N$ matrix. What we were dealing with was really matrix and not vector. The information in this matrix was a superposition of edge information over the node values themselves, that is to say, one-dimensional.

Here, I would like to emphasize that the connection between N nodes are, if the values of the nodes are not null, and also if the values of the edges are not null for all the N nodes, the dimension will be N-dimensional. In reality the definition of graph shall be based on the noise-filtered sparsified node and edge values of the brain graph, we need to assume graph and subgraphs. I assume that brain is working as a disjoint union of brain subgraphs dynamically changing along time. In this sense, brain graph is now to be considered as a dynamic system, which has a regional distribution and the connections thereof, but the dimension is N-dimension. What we are observing is zero-dimensional (point data), one-dimensional (edges between any two brain nodes), or two-dimensional (face of concurrently functionally activated (FDG PET, activation fMRI or MEG) or structurally enlarged (T1 MRI) three brain nodes). The latter is called 2-face, then edge is 1-face and the point is 0-face. 2-face can have a hole, which is open or closed within the triangle. We can also designate three-face of three-dimensional tetrahedron (strong connection between 4 nodes) and so on. This viewpoint is the core concept of simplicial complex in homology of algebraic topology. We filtered our brain graph data and observed only the topological invariants of 0-dimension, the connected components, i.e., point-equivalent. We recently went on to observe hole using one-dimensional topological invariant, Betti-1, and we could find the FDG uptake on PET, metabolic activity of brain nodes were having the one-dimensional hole found in Alzheimer's disease while we were filtering the metabolic brain graph data (Lee et al. 2014). This was expected, however, we did not expect the discovery that mild cognitive impairment patients also had one-dimensional connectivity holes, though the numbers of holes were smaller than Alzheimer's disease.

3.8 Further Problems and Plausible Solutions

Finding global difference of connectivity of grain graphs and localizing the difference-causing edges between the normal (or controls) and disease groups are said to be important, and statistical inference should have been performed. However, we assumed arbitrarily again independence between nodes and set the threshold for finding significant trustworthy edges that the data can allow us to find, at least several edges making the difference between normal and disease. Generation of permuted pseudorandom raw data was inevitable but it immediately limited the result of statistical inference only to the sample population. No matter how the method was threshold-free in the beginning, we cannot be sure that our observation be applied to the population.

3.9 Minimize FNDR of Significantly Different Connections Based on H-Likelihood

To overcome this flaw, we tried and still trying to trim the confounding influence of unrelated parameters. One is considering the spatial adjacency of the nodes of the brain graph, and the other is appropriate modeling to reveal (or conceal) the true positives (or false positives). The latter was solved first by Donghwan Lee and us (Lee et al. 2015). They adopted the hierarchical general linear model (H-GLM) (Lee et al. 2017a, b)-derived extended likelihood method to solve the modeling and the minimization of false non-discovery rate (FNDR) while maintaining false discovery rate (FDR) set a priori. The improvement of decreasing FNDR was validated by the observation that hippocampal metabolic decrease was observed in wider area, as expected, than that obtained by Benjamini–Hochberg FDR correction (Lee et al. 2015). However, this endeavor was about the distribution (zero-dimensional information) of regional activity such as FDG uptake on PET, and not about the edge information from these regional distribution of FDG uptake. The beauty of this method was at the point that we did data-driven modeling simultaneously with the discovery procedure. This endeavor let us decrease FNDR. But we also found that the advantageous effect was only observed when the FNDR was in the moderate range. The application of this method to the single linkage matrix is underway by collaboration of myself, my group and Lee and Lee (2016).

The former might have been brain graph-specific problem. The gist was taking the spatial adjacency into account. Lee and Lee (2016) successfully extended their prior approach (Lee et al. 2015, 2017a, b) so that once the hidden Markov random field took care of spatial adjacency of nodes in Euclidean space, it decreased marginal FNDR down to 0.03 compared with 0.5 without adjacency consideration. This was so impressive that I believe that the same consideration of spatial adjacency of the participating nodes will decrease FNDR and thus yield the true-positive edges having statistical significance to the full account. This extended likelihood method of mini-

mizing FNDR while keeping FDR as smaller than a certain value such as 0.05 will allow us find any edges of difference between groups. This endeavor is underway too.

3.10 Prospects of Using Personal Connectomics Data for Clinical Use

For clinical use of connectivity data characterizing each healthy or diseased subjects, i.e., personal connectomics, we need to overcome the hurdle of group thinking. As the data are to be statistically handled, we tended to merge the data from a group of symptomatically homogeneous patients and understand the signature of a specific condition or pathology. For personal connectomics, Kolmogorov-like test for connectomics should now be established. In the beginning, as a stepping-stone trial or as a leverage, individual PET/resting FMRI is the best option to delineate the individualized connectivity matrix such as single linkage matrix, and then in feedback to the discovered group characteristics, personal connectomics will be interpreted considering these group characteristics as norms finally to find whether that individual belongs to this group. Deep learning approach might help this endeavor. Classification shall now be done on every available data. Thus, clinical use is on the horizon within our reach.

Acknowledgements This study was supported by the National Research Foundation of Korea (NRF) Grant funded by Korean Government (MOE) (No. 2016R1D1A1A02937497), the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIP) (No. 2015M3 C7A1028926, No.2017R1A5A1015626 and No. 2017M3C7A1048079).

References

Ahn, Y. Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, *466*(7307), 761–4.

Bassett, D. S., Nelson, B. G., Mueller, B. A., Camchong, J., & Lim, K. O. (2012). Altered resting state complexity in schizophrenia. *Neuroimage*, *59*(3), 2196–207.

Batson, J., Spielman, D. A., Srivastava, N., & Teng, S. H. (2013). Spectral sparsification of graphs: Theory and algorithms. *Communications of the ACM*, *56*(8), 87–94.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–98.

Choi, H., Kim, Y. K., Kang, H., Lee, H., Im, H. J., Hwang, D. W., et al. (2014). Abnormal metabolic connectivity in the pilocarpine-induced epilepsy rat model: A multiscale network analysis based on persistent homology. *Neuroimage*, *1*(99), 226–36.

Gao, W., Emaminejad, S., Nyein, H. Y., Challa, S., Chen, K., Peck, A., et al. (2016). Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. *Nature*, *529*(7587), 509–14.

Hahm, J., Lee, H., Park, H., Kang, E., Kim, Y. K., Chung, C. K., et al. (2017). Gating of memory encoding of time-delayed cross-frequency MEG networks revealed by graph filtration based on persistent homology. *Scientific Reports*, *7*(7), 41592.

http://www.fil.ion.ucl.ac.uk/spm/

Im, H.J., Hahm, J., Kang, H., Choi, H., Lee, H., Hwang, do W., Kim, E.E., Chung, J.K., Lee, D.S. (2016). Disrupted brain metabolic connectivity in a 6-OHDA-induced mouse model of Parkinson's disease examined using persistent homology-based analysis. *Scientific Reports*, *6*:33875.

Kim, H., Hahm, J., Lee, H., Kang, E., Kang, H., & Lee, D. S. (2015). Brain networks engaged in audiovisual integration during speech perception revealed by persistent homology-based network filtration. *Brain Connectivity*, *5*(4), 245–58.

Kim, E., Kang, H., Lee, H., Lee, H. J., Suh, M. W., Song, J. J., et al. (2014). Morphological brain network assessed using graph theory and network filtration in deaf adults. *Hearing Research*, *315*, 88–98.

Lee, H., Chung, M.K., Kang, H., Kim, B.N., Lee, D.S. (2011). Computing the shape of brain networks using graph filtration and Gromov-Hausdorff metric. *Medical Image Computing and Computer-Assisted Intervention*, *14*(Pt 2), 302–309.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2017a). Generalized linear models with random effects: Unified analysis via H-likelihood. Chapman & Hall/CRC.

Lee, Y., Ronnegard, L., & Noh, M. (2017b). *Data analysis using hierarchical generalized linear models with R*. CRC Press.

Lee, H., Chung, M. K., Kang, H., & Lee, D. S. (2014). Hole detection in metabolic connectivity of Alzheimer's disease using kappa-Laplacian. *Medical Image Computing and Computer-Assisted Intervention*, *17*(Pt 3), 297–304.

Lee, D., Ganna, A., Pawitan, Y., & Lee, W. (2016). Nonparametric estimation of the rediscovery rate. *Statistics in Medicine*, *35*(18), 3203–12.

Lee, H., Kang, H., Chung, M. K., Kim, B. N., & Lee, D. S. (2012). Persistent brain network homology from the perspective of dendrogram. *IEEE Transactions on Medical Imaging*, *31*(12), 2267–77.

Lee, H., Kang, H., Chung, M. K., Lim, S., Kim, B. N., & Lee, D. S. (2017). Integrated multimodal network approach to PET and MRI based on multidimensional persistent homology. *Human Brain Mapping*, *38*(3), 1387–1402.

Lee, D., Kang, H., Kim, E., Lee, H., Kim, H., Kim, Y. K., et al. (2015). Optimal likelihood-ratio multiple testing with application to Alzheimer's disease and questionable dementia. *BMC Medical Research Methodology*, *30*(15), 9.

Lee, D. S., Kang, H., Kim, H., Park, H., Oh, J. S., Lee, J. S., et al. (2008). Metabolic connectivity by interregional correlation analysis using statistical parametric mapping (SPM) and FDG brain PET; methodological development and patterns of metabolic connectivity in adults. *European Journal of Nuclear Medicine and Molecular Imaging*, *35*(9), 1681–91.

Lee, D., & Lee, Y. (2016). Extended likelihood approach to multiple testing with directional error control under a hidden Markov random field model. *Journal of Multivariate Analysis*, *151*, 1–13.

Lee, H., Lee, D. S., Kang, H., Kim, B. N., & Chung, M. K. (2011). Sparse brain network recovery under compressed sensing. *IEEE Transactions on Medical Imaging*, *30*(5), 1154–65.

Lee, D. S., Lee, J. S., Oh, S. H., Kim, S. K., Kim, J. W., Chung, J. K., et al. (2001). Cross-modal plasticity and cochlear implants. *Nature*, *409*(6817), 149–50.

Lillie, E. O., Patay, B., Diamant, J., Issell, B., Topol, E. J., & Schork, N. J. (2011). The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Personalized Medicine*, *8*(2), 161–173.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J. P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, *328*(5980), 876–8.

Park, M., Do, K., Kim, J., Son, D., Koo, J. H., Park, J., et al. (2015). Oxide nanomembrane hybrids with enhanced mechano- and thermo-sensitivity for semitransparent epidermal electronics. *Advanced Healthcare Materials*, *4*(7), 992–7.

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., et al. (2015). ClinGen-the clinical genome resource. *New England Journal of Medicine*, *372*(23), 2235–42.

Relling, M. V., & Evans, W. E. (2015). Pharmacogenomics in the clinic. *Nature*, *526*(7573), 343–50.

Roebroeck, A., Formisano, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, *25*(1), 230–42.

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage*, *52*(3), 1059–69.

Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., & Ringach, D.L. (2008). Topological analysis of population activity in visual cortex. *Journal of Vision*, *8*(8), 11.1–18.

Thompson, P. M., Giedd, J. N., Woods, R. P., MacDonald, D., Evans, A. C., & Toga, A. W. (2000). Growth patterns in the developing brain detected by using continuum mechanical tensor maps. *Nature*, *404*(6774), 190–3.

Worsley, K. J., Chen, J. I., Lerch, J., & Evans, A. C. (2005). Comparing functional connectivity via thresholding correlations and singular value decomposition. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, *360*(1457), 913–20.

Worsley, K. J., Taylor, J. E., Tomaiuolo, F., & Lerch, J. (2004). Unified univariate and multivariate random field theory. *Neuroimage*, *23*(Suppl 1), S189–95.

Xie, J., Douglas, P. K., Wu, Y. N., Brody, A. L., & Anderson, A. E. (2017). Decoding the encoding of functional brain networks: An fMRI classification comparison of non-negative matrix factorization (NMF), independent component analysis (ICA), and sparse coding algorithms. *Journal of Neuroscience Methods*, *15*(282), 81–94.

Zomorodian, A., & Carlsson, G. (2005). Computing persistent homology. *Discrete & Computational Geometry*, *33*(2), 249–74.

Chapter 4
The Real-Time Tracking and Alarming the Early Neurological Deterioration Using Continuous Blood Pressure Monitoring in Patient with Acute Ischemic Stroke

Youngjo Lee, Maengseok Noh and Il Do Ha

Abstract In this paper, we develop a real-time prediction of END (Early Neurological Deterioration) using continuous BP (blood pressure) monitoring and clinical parameters and propose to set up an alarming criterion before END. We identified consecutive ischemic stroke patients hospitalized within 48 h of symptom onset from a prospective stroke registry database. BP data during hospitalization were obtained from the electric medical records. Probability of END at each time point of BP measurement was estimated using a logistic model with covariates, which is derived from two models for clinical information and BP parameters. Here, a model for clinical information was fitted using logistic model with clinical characteristics of patients to predict END. A model for BP was fitted using random effects models allowing for temporal correlations at each time point of BP measurement with irregular intervals. Prediction performance was evaluated by sensitivity and specificity. An alarm criterion for a high probability of END at each time point was defined as being above a cutoff point prior to 24 h.

4.1 Introduction

Approximately 30% of hospitalized patients due to acute ischemic stroke are placed under the risk of Early Neurological Deterioration (END) at their hospital stay. These events constitute serious and adverse problems, such as an extension of hospitalization duration, an increasing demand for more resources, and an aggravation of neurologic disability and death (Ois et al. 2008). As the prevention and timely

Y. Lee (✉)
Department of Statistics, Seoul National University, Seoul, Korea
e-mail: youngjo@snu.ac.kr

M. Noh · I. D. Ha
Department of Statistics, Pukyong National University, Busan, Korea
e-mail: msnoh@pknu.ac.kr

I. D. Ha
e-mail: idha1353@pknu.ac.kr

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_4

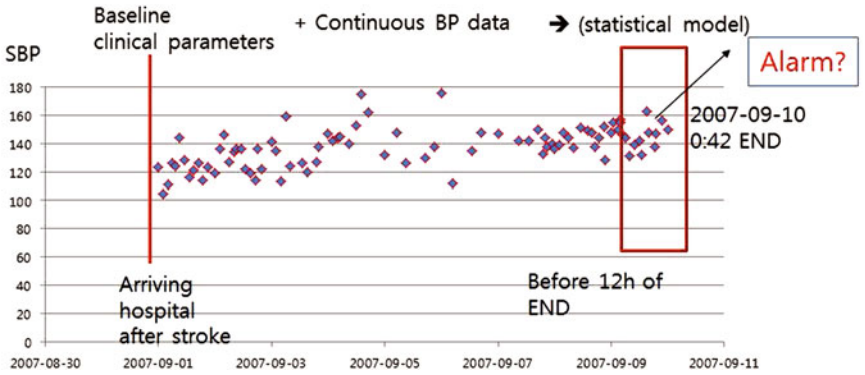


Fig. 4.1 Real-time prediction model of END

intervention of such events would request an intensive monitoring by specifically devoted team, it needs a prioritization of medical services to patients at high risk (Ay et al. 2010).

The real-time risk stratification tool that predicts and alarms before END would be useful for delineating patients at high risk of acute stroke care system. This concept has been developed for management of critically ill patients. That is, it is tracking the patient's condition by automated monitoring vital sign based on electric medical records (EMRs) and is triggering the response when predetermined threshold is reached (Roquer et al. 2008).

For the patients with acute ischemic stroke, real-time tracking of END probability would be actualized by monitoring the individual risk assumptions with continuous BP and known other risk factors. Since BP is associated with not only development of END but also change of physiologic condition, its monitoring would be prompt for tracking the patient (Jenkins et al. 2011).

As shown in Fig. 4.1, in this study, we firstly aim to develop the real-time prediction model of END using continuous tracking BP and baseline clinical parameters. Next, we try to set up an alarming criterion to delineate patients at high risk before 12 h of END by analyzing the cumulating prediction values.

4.2 Methods

4.2.1 Subjects and Measurements

The subjects of study were from the prospective stroke registry that was consecutively enrolled in hospitalized patients diagnosed as ischemic stroke at Seoul National University Bundang Hospital, Republic of Korea. Among them, patients arrived

within 48 h of symptom onset between April 2008 and March 2015 were selected. Here we excluded the subjects whose BP was less than ten.

The demographics and administrative clinical information of subjects of study were collected by reviewing the electric health recording (EHR) and registry database. They consisted of baseline patient characteristics (i.e., age, sex, and vascular risk factors such as hypertension, diabetes, hyperlipidemia, and atrial fibrillation) and index stroke characteristics (i.e., baseline National Institute of Health Stroke Scale (NIHSS) score, stroke subtypes, symptomatic steno-occlusion of cerebral artery, implementation of acute revascularization therapy or not, and result of acute revascularization therapy).

The measurement of BP was regulated by physician's decision based on the current guideline and hospital routine. In general, BP was regularly measured every hour, which was adjusted by conditions of patients using noninvasive BP monitoring device or standard mercury sphygmomanometer on non-hemiparetic arm at supine position. All BP information within 72 h of hospital arrival was obtained from the EHR.

As a part of an institutional quality-of-care monitoring program for hospitalized patients, neurologic deterioration was prospectively monitored and finally adjudicated at regular meeting of stroke team constituted by experienced nurses and physicians. Neurologic deterioration indicated one of the following: increase of more than two points in the total NIHSS score, an increase of more than one point in the level of consciousness or monitor items of NIHSS score, and a newly developed neurologic symptom or sign within 72 h of symptom onset. The END occurred within 3 days of stroke onset was the primary outcome.

4.2.2 Statistical Model

The baseline prediction model was constructed using multivariate logistic regression models with total subjects and predetermined subgroups:

$$\log\{p_i/(1-p_i)\} = x_i^T \gamma,$$

where p_i is the probability for END using the ith patient's covariates x_i for baseline clinical characteristics at admission and γ is a vector of regression parameters corresponding to covariates x_i. For joint modeling of mean and variance of SBP (systolic BP) measurement y_{it} at the tth time point of the ith patient, we consider hierarchical generalized linear models (HGLMs) allowing for temporal correlations with irregular intervals (Lee et al. 2017):

$$\log(y_{it}) \sim N(\mu_{it}, \phi_{it}),$$

where

$$\mu_{it} = \mu_0 + v_{1i} + v_{2t} \text{ and } \log(\phi_{it}) = \log(\phi_0) + w_{1i} + w_{2t}.$$

Here, $v_{1i} \sim N(0, \sigma^2_{v_1})$ and $w_{1i} \sim N(0, \sigma^2_{w_1})$ are random subject effects. Time effects v_{2t} and w_{2t} are defined by $r_1 = A(\rho_1)v_2$ and $r_2 = A(\rho_2)w_2$, where $A(\rho) = I - K(\rho)$, the nonzero elements of $K(\rho)$ are $K_{j+1,j}(\rho) = \rho/|t_{j+1} - t_j|$, and t_j is the measurement time after admission at the jth time point. Here, ρ_1 and ρ_2 are temporal correlations for the mean and variance, respectively. This model can be fitted by using the hierarchical-likelihood approach (Lee and Nelder 1996, 2001).

After fitting of the two joint models above, the real-time prediction model for END within 12 h was developed. Let π_{it} be the probability of END within 12 h at the tth time point of ith patient. We consider the following logistic model:

$$\log\{\pi_{it}/(1 - \pi_{it})\} = \beta_0 + \beta_1 \log(\widehat{p}_i) + \beta_2 \log(\widehat{\mu}_{it}) + \beta_3 \log(\widehat{\phi}_{it}).$$

If a criterion $\pi_{it} > \delta$ was satisfied twice when the tth BP measurement of the ith patient was observed, we predict this patient has high risk of END. The model performances were tested by the area under receiver operating curve (AUROC) with sensitivity and specificity.

4.3 Results

4.3.1 Subjects and Baseline Characteristics

Among 1986 consecutive patients identified, we chose 1805 subjects for development and validation model after excluding 181 subjects with unavailable BP data. Mean age was 67.3 ± 13.0 years and male comprised 60.1%. During hospitalization, 331 patients (18.3%) experienced the END events. Median event time from hospital arrival to END was 19 h (interquartile range, 6 to 41 h). The number of total BP data is observed as approximately 220,000, so that each patient has average of 56 BP data.

4.3.2 Model Development and Alarming Criteria

The baseline prediction model was fitted with the following covariates x_i: age, sex, history of stroke, time to arrival (at hour), baseline NIHSS score, diabetes, initial glucose level, atrial fibrillation, leukocyte count, stroke subtypes, recanalization therapy, and location of symptomatic vessels. If we use only $\widehat{p}_i$ for the END prediction model with $\beta_2 = \beta_3 = 0$, the model performance is evaluated as AUROC = 0.703 with sensitivity = 0.67 and specificity = 0.65.

Substituting the results of HGLM for BP into the prediction model leads to the following fitted model:

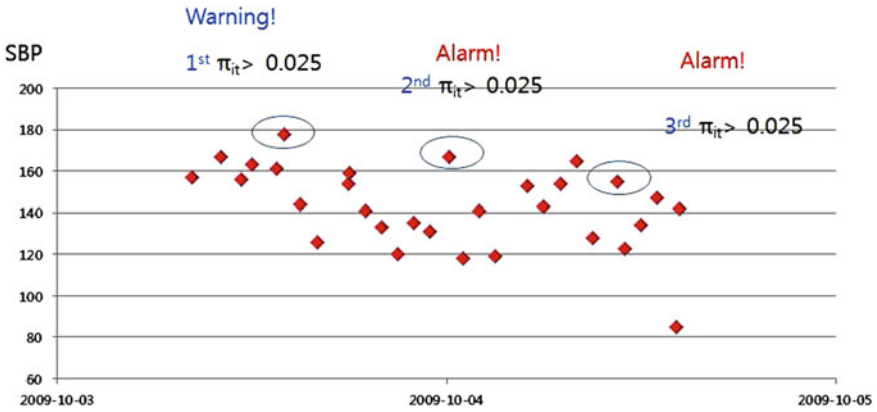


Fig. 4.2 Example of alarming criteria

$$\log\{\widehat{\pi}_{it}/(1-\widehat{\pi}_{it})\} = -14.54(\pm 0.64) + 4.13(\pm 0.15)\log(\widehat{p}_i) + 2.09(\pm 0.13)\log(\widehat{\mu}_{it}) + 0.10(\pm 0.021)\log(\widehat{\phi}_{it}).$$

The estimated effects of mean and variance of SBP are very statistically significant, so that we observe the high mean and high variance of SBP are important risk factors for predicting END.

As shown in Fig. 4.2, we predict END case at the tth SBP measurement if a criterion $\pi_{it} > 0.025$ was satisfied twice. With this alarming criteria, we could have 85.7% of true alarming (sensitivity) and 13.2% of mis-alarming (1-specificity), followed by AUROC = 0.810.

References

Ay, H., Gungor, L., & Arsava, E. M. (2010). A score to predict early risk of recurrence after ischemic stroke. *Neurology*, *74*, 128–135.

Jenkins, P. F., Thompson, C. H., & Barton, L. L. (2011). Clinical deterioration in the condition of patients with acute medical illness in Australian hospitals: improving detection and response. *Medical Journal of Australia*, *194*, 596–598.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society B*, *58*, 619–678.

Lee, Y., & Nelder, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect model and structured dispersion. *Biometrika*, *88*, 987–1006.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2017). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Ois, A., Martinez-Rodriguez, J. E., & Munteis, E. (2008). Steno-occlusive arterial disease and early neurological deterioration in acute ischemic stroke. *Cerebrovascular Diseases*, *25*, 151–156.

Roquer, J., Rodrguez-Campello, A., & Gomis, M. (2008). Acute stroke unit care and early neurological deterioration in ischemic stroke. *Journal of Neurology*, *255*, 1012–1017.

Part III
Operation/Production Decision Making

Chapter 5
Condition Monitoring and Operational Decision-Making in Modern Semiconductor Manufacturing Systems

Dragan Djurdjanovic

Abstract Modern semiconductor manufacturing tools are often complex systems of numerous interacting subsystems that operate in multiple physical domains and often follow highly nonlinear distributed dynamics. In such systems, traditional condition monitoring methods, which rely on a direct link between sensor readings and the underlying condition of the system, cannot be used. Rather, one must acknowledge that the available sensor readings are only stochastically related to the condition of the monitored system, which therefore must be probabilistically inferred from the sensors. This manuscript describes a recently proposed condition monitoring method, based on characterizing the degradation process via a mixture of operation-specific hidden Markov models (HMMs), with hidden states representing the unobservable degradation states of the monitored system, while its observable variables represent the available sensor readings. The new monitoring paradigm was applied to monitoring of several tools operating in major semiconductor fabs over many months, with orders of magnitude better performance than traditional, purely signature-based approaches. The remainder of the paper focuses on describing how Markovian models of degradation of flexible manufacturing equipment, such as those utilized in modern semiconductor manufacturing, can be employed to concurrently optimize the sequence of production operations and schedule preventive maintenance for that machine. It will be shown that integrated decision-making in terms of product sequencing and maintenance operations carries significant potential benefits compared to the more traditional, fragmented decision-making. The manuscript ends with a brief summary of possible future research directions in process monitoring and maintenance decision-making in semiconductor manufacturing.

D. Djurdjanovic (✉)
Department of Mechanical Engineering, University of Texas at Austin, Austin, USA
e-mail: dragand@me.utexas.edu

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_5

5.1 Introduction

In today's competitive, customer-oriented market, companies must provide products and services of the highest possible quality in order to attain and retain a favorable market position. Such pressures are particularly prevalent in semiconductor manufacturing, where intense competition and short product life cycles necessitate continuous innovation and maximal levels of efficiency (Semiconductor Industry Association (SIA) 2015).

The vast majority of equipment maintenance in semiconductor manufacturing today is either purely reactive (fixing or replacing equipment or its components after a failure occurs) or proactive (assuming a certain level of performance degradation, with no input from the equipment itself, and servicing equipment on a routine schedule whether service is actually needed or not). Both scenarios are wasteful and result in costly production or service downtimes. Even though it often seems that a system fails suddenly, each piece of equipment usually goes through a measurable process of degradation before it fails. With the advancement of semiconductor manufacturing technology, the tools in both front-end and back-end operations are becoming increasingly sensorized, with capabilities of collecting a substantial amount of data during the process. Therefore, it is now possible to rapidly and accurately sense performance indicators, and thus assess and predict system degradation states.

Under these circumstances, Condition-Based Maintenance (CBM), based on sensing and assessing the current and sometimes future degradation states of the target system, emerges as an appropriate and efficient tool for achieving near-zero breakdown time through a significant reduction, and, when possible, elimination of downtime due to process or machine failure (Lee et al. 2006, 2013; Djurdjanovic et al. 2003). It is documented that a well-implemented CBM system in a company can save up to 20% of operational costs due to a number of benefits, such as Lee et al. (2013):

- Improved machine availability and productivity due to a decrease in equipment downtime
- Smaller production losses and waste because of the improved quality
- Reduced environmental footprint because of increased manufacturing efficiency and reduced waste
- Decreased costs of maintenance due to the ability to perform non-intrusive maintenance operations synchronized with the production planning
- Cost savings due to improved resource efficiency, decreased spare parts inventory, and maintenance personnel levels
- Improved decision-making with regards to scheduling and sampling

Considering the fact that semiconductor manufacturing is characterized by a relatively high level of technology integration and highly pronounced needs for optimized production flow and quality control, potential savings in semiconductor manufacturing could easily be even higher. Nevertheless, this research direction carries a plethora of challenges unique to the semiconductor manufacturing discipline, which is why

CBM and predictive maintenance efforts are only now gaining momentum in the research and industrial community. This manuscript offers a brief overview of recent key achievements in CBM research in semiconductor manufacturing, with the hope that it could serve as a solid foundation for future endeavors in the advancement of CBM in this area. The remainder of this chapter is organized as follows. Section 5.2 briefly outlines the general concept of CBM and highlights some key challenges it presents to the semiconductor industry. Section 5.3 summarizes key achievements in signal processing and feature extraction, while Sect. 5.4 briefly describes a novel approach for condition modeling in highly complex systems where a direct relation between sensor readings and the system condition could not be established, leading to an innovative framework for fault detection, diagnosis and prediction in semiconductor manufacturing. Section 5.5 offers recent results in operational decision-making in semiconductors fab where simulation-based optimization coordinates condition-based monitoring information with fab operations to yield system-level optimized decisions. Section 5.6 provides concluding thoughts and identifies some key areas for future research in CBM for semiconductor manufacturing.

5.2 Condition-Based Maintenance Paradigm with Key Relevant Challenges in Semiconductor Manufacturing

Condition-Based Maintenance (CBM) can be seen as an integral process of seamless transformation of raw data related to the equipment health and performance, into information about equipment health, and further into decisions that need to be made with respect to that equipment, as illustrated in Fig. 5.1.

Information about the health of any piece of equipment is obtained from the readings of possibly multiple sensors mounted on that equipment. Often, situations exist where sensor readings are augmented with historical knowledge about equipment behavior, engineering models of phenomena occurring in the equipment, or human expertise. Based on these sources of information, features relevant to equipment health are extracted from sensor readings through various forms of sensory *signal processing and feature extraction*. These features form *behavior models of equipment in different health states* (normal behavior and different faulty behavior modes). Those models may be in various forms, including a statistical form (distributions of sensory signatures under normal or various faulty conditions), dynamic model (differential equations describing various health states of the equipment), and others. Based on the models of normal and current equipment behavior, equipment *health assessment* can be accomplished by quantitatively expressing the proximity of the currently observed system behavior to the model describing its normal health state (e.g., fault detection can be done in this manner). Similarly, the presence or absence of any fault can be *diagnosed* through proximity of the model of the currently observed equipment behavior to the behavior model corresponding to a specific fault (fault diagnosis). Finally, the temporal dynamics of signatures extracted from sensor readings

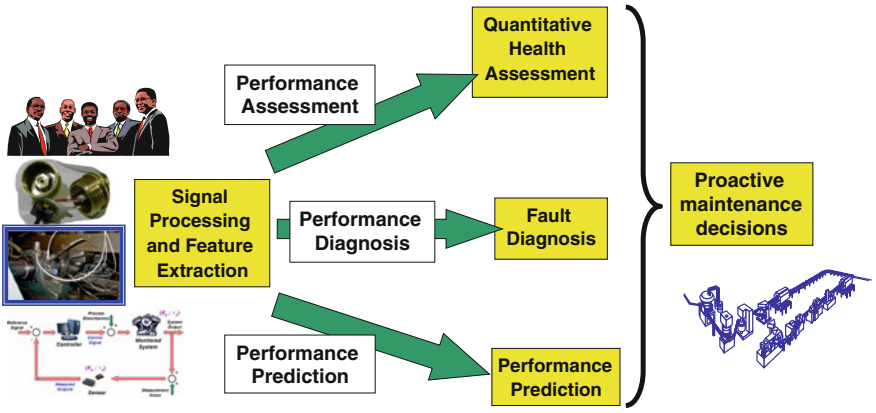


Fig. 5.1 Illustration of the general concept of CBM

can be captured and extrapolated to predict their behavior in the future and thus predict likelihoods of various behavior modes for the equipment. Figure 5.2 illustrates the concepts of quantitative health assessment and diagnosis in CBM based on simple statistical models of various behavioral modes, while Fig. 5.3 illustrates the concept performance prediction in CBM.

The concept of CBM has received significant attention in recent years, especially in the case of sophisticated, expensive, and safety critical systems, such as manufacturing equipment (Rao 1996; Funk and Jackobson 2005), computer networks (Hofmeyr and Forrest 2000; Boukerche et al. 2004; Harmer et al. 2002; Dasgupta and Gonzalez 2002; Yang et al. 2002; Hortos 2003), automotive (Cascio et al. 1999; Marko et al. 1990; Crossman et al. 2003a, b; Hong et al. 2000) and aircraft engines (Beniaminy and Joseph 2002; Gorinewsky et al. 2002; Kobayashi and Simon 2001; Yan et al. 2005; Wegerich 2003, 2004). Such progress of CBM in general areas of engineering represents an opportunity to adopt and/or adapt numerous existing CBM methods to solve the diagnostic and prognostic problems in semiconductor manufacturing. However, enabling the vision of CBM in semiconductor manufacturing requires tackling of some challenges that are very unique to the semiconductor manufacturing industry. Those challenges are present in the stage of transformation of data into information (in the functions of feature extraction, performance assessment, and prediction), as well as in the stage of transformation of information into decisions (in the function of maintenance and operations decision-making).

From the side of *transforming data into information*, key semiconductor manufacturing tools and systems are highly complex machines in which phenomena from quantum physics, electro-mechanics, thermodynamics and fluid mechanics and other domains concurrently play out in highly irregular geometries as patterns of circuitry are successively produced with angstrom-level accuracy across a 300 mm diameter wafer. Traditional methods, based on time-domain or frequency-domain processing of signals and purely statistical interpretation of data based on physical models or

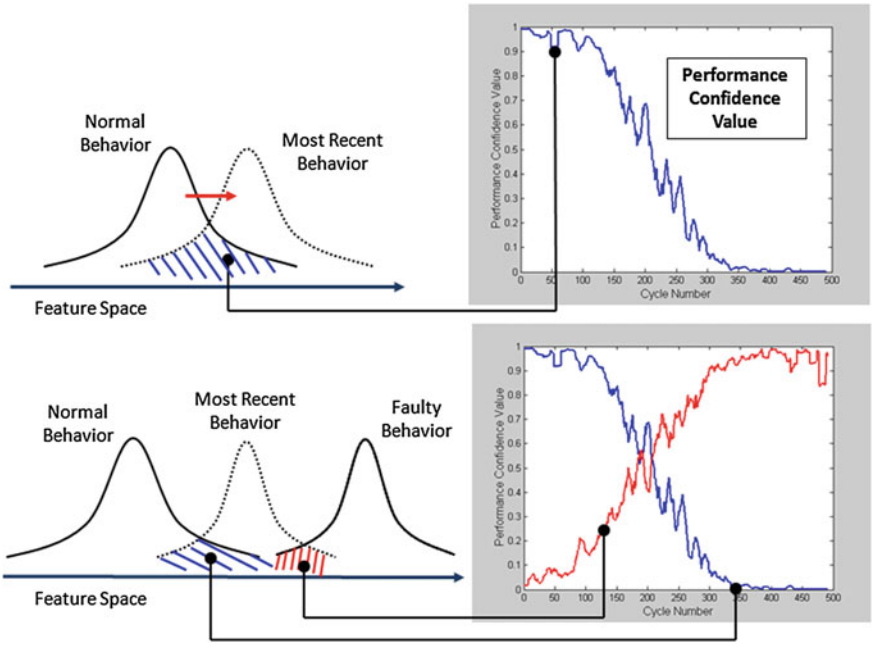


Fig. 5.2 Performance assessment & diagnosis through overlapping of signature distributions

previously seen patterns in the signals, often become utterly unfeasible in such an environment, demanding radically different and innovative information extraction methods.

From the perspective of *operational decision-making*, a typical fab is a highly complex system of interconnected tools and often a vast mixture of products that go through that system. Numerous random effects,[1] as well as complex factory dynamics of interactions between vast numbers of machines, many of which can execute several different operations on different products, lead to a highly intractable problem of synchronizing maintenance and production decisions in a fab. Tractable assumptions about reliability distributions, availability of condition-related information throughout the system, cycle times, inter-machine interactions, and other factors which characterize a great majority of scholarly work are so far from reality in a typical fab environment that significantly new operational decision-making paradigms are needed.

In the next few sections, some key recent accomplishments in advancing information extraction from large amounts of densely sampled sensor data in a fab, as well as system-level optimized operational decision-making in modern semiconductor manufacturing environments, will be discussed.

[1] Random effects exist due to factors such as equipment condition, cycle times, in-process and final product quality, effectiveness of maintenance interventions, availability of spare parts, supply and demand.

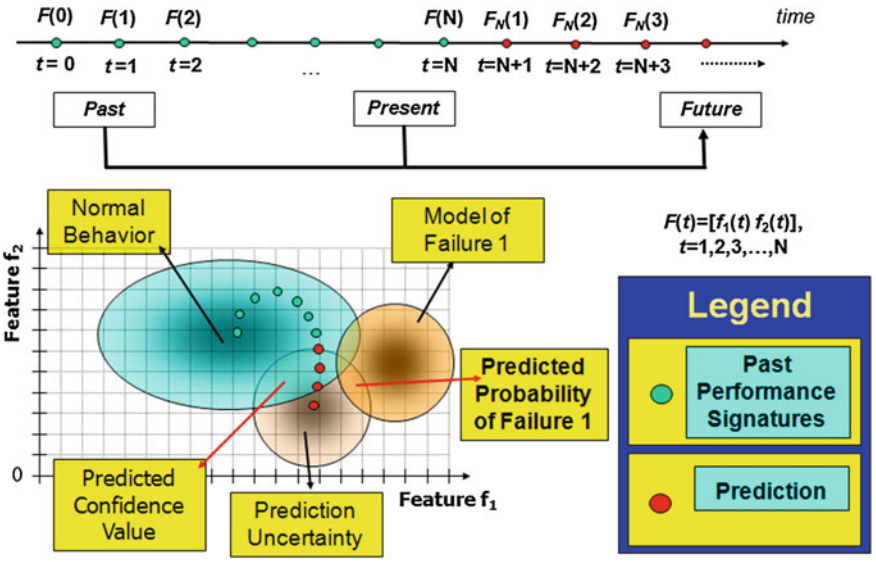


Fig. 5.3 Concept of feature-based performance prediction with prediction confidence intervals. Model of behavior of feature vectors $F(n) = [f_1(n)\ f_2(n)]$, $n = 1, 2, \ldots, N$ can be used to extrapolate their behavior ahead in time and obtain l time steps ahead predictions for those feature vectors $F_N(l)$, along with the associated uncertainties

5.3 Advances in Signal Processing and Extraction of Informative Signatures from Sensors in Semiconductor Manufacturing

A typical semiconductor fab contains hundreds of processing tools interconnected with sophisticated material handling systems, with each of these machines instrumented with hundreds of sensors, which in turn emit multiple data points each second. In this deluge of data, nuggets of useful information about the underlying condition of equipment are buried and extracting them is a major challenge. This section will present several recent advances that enable extraction of useful condition-related information from such data. First, examples of using advanced time–frequency signal analysis to detect and characterize particle-generating features in a material handling device, as well as to monitor slit valve performance in a chamber-based process tool, will be presented. This will be followed by a brief description of a recently introduced method for extraction of dynamics-inspired features from densely sampled signals obtained from semiconductor manufacturing processes, and how such advanced signatures can be used to improve virtual metrology for relevant processes.

5.3.1 Use of Cohen's Class Time–Frequency Distributions for Analysis of Signals in Semiconductor Manufacturing

Most signals in nature are highly non-stationary signals, with frequency content varying over time. An aircraft engine transitioning from one regime of operation into another emits non-stationary vibrations and sounds because excitation caused by variable rotational speeds causes variations in the frequency contents of the signals. Most real-life signals, such as speech, music, machine tool vibration, acoustic emission, are also non-stationary, which places strong emphasis on the need for development and utilization of non-stationary signal analysis techniques, such as wavelets or joint time–frequency analysis.

Most traditional time-domain or frequency-domain-based monitoring techniques for monitoring of dynamic systems (bearings, gears, machine tools, engines, DC/AC motors and drives, etc.) utilize stationary signal characterization methods, such as time series modeling or Fourier domain analysis (modal and spectral analysis) (Marple 1987). These methods assume that frequency content of the signal does not change over time, smearing the information when various frequency components appear or disappear in the signal. In simple terms, one is aware of what frequencies exist in the signal, but not when they existed (Cohen 1995).

Figure 5.4 depicts the inadequacy of applying stationary signal processing techniques, such as Fourier transforms, to non-stationary signals such as simple frequency hopping signals shown in Fig. 5.4. Fourier analysis is able to discern the three sinusoids present in the signals, but is unable to deduce when each one of those sinusoids occurred. Therefore, when the order of sinusoids is altered, the Fourier analysis is unable to detect this change, as indicated in the figure.

More recent work in monitoring and CBM focuses on applications based on wavelet signal transforms (Burrus et al. 1998). Even though wavelet techniques already seem to be a widely accepted method for signal processing and feature extraction in the presence of non-stationary frequency varying signals (Du et al. 1995; Wang et al. 2001), advances in computing technology are slowly allowing a more intensive use of signal processing and feature extraction tools based on the Cohen's class of joint, time–frequency distributions (Cohen 1995; Williams 1996; Djurdjanovic et al. 2002). The origins of this powerful signal description can be traced back to 1930s and advances in quantum physics in the work of Wigner (1932), where he needed to calculate a joint distribution of a particle having a given position and momentum. However, the position and momentum in quantum physics are connected through a Fourier transform, very much in the same way time and frequency contents of a signal are connected in the signal processing theory. This was noticed by a French engineer Ville (1948), who realized that the same approach could be utilized to describe joint distributions of signal energy in both time and frequency.

The Reduced Interference Distribution (RID) time–frequency kernels, developed in mid-1990s at the University of Michigan (Jeong and Williams 1992; Jeong 1990), represent a class of signal-independent, and therefore computationally less demanding, time–frequency kernels that result in time–frequency distributions (TFDs) whose

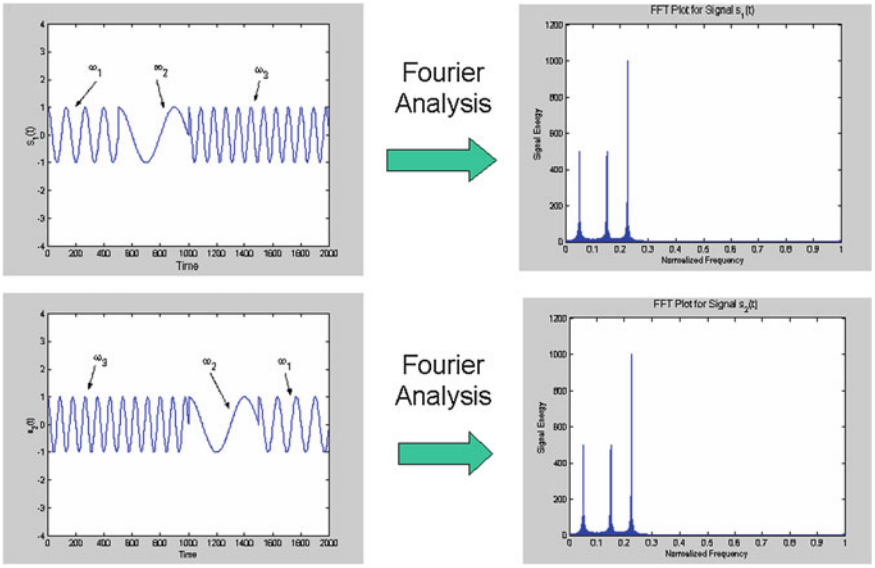


Fig. 5.4 Application of Fourier analysis on two frequency hopping signals

favorable mathematical properties include (Jeong and Williams 1992; Jeong 1990): *time-shift, frequency-shift and scale covariance properties, frequency and time marginal properties* and*instantaneous frequency and group delay properties.* In addition, RIDs have the property of suppressing the TFD cross-terms, which necessarily exist whenever multi-component signals are processed. Cross-terms are sometimes indistinguishable from the auto-terms and can hamper the time–frequency-based signal interpretation and pattern recognition (Williams 1996; Djurdjanovic et al. 2002). Suppression of cross-terms is therefore a desirable mathematical property, and RIDs achieve it in a signal-independent manner, which is computationally quicker to accomplish than the signal-dependent suppression pursued, for example, in Baraniuk and Jones (1993).

Figure 5.5 shows the RID signal energy distribution of the same signals shown in Fig. 5.4. One can readily distinguish the three sinusoids present in the signal, as well as when those sinusoids existed. Figure 5.6 shows applicability of joint time–frequency signal analysis techniques to vibration signatures from a gearbox taken, while gearbox was accelerating. Close observation of energy patterns in the time–frequency plane indicates a series of energy "bumps" that occur closer and closer together, and correspond to the mashing of gear teeth.

Vibrations associated with material handling devices are usually very non-stationary, and utilization of Cohen's class of time–frequency distributions for their analysis carries significant potential benefits. This is especially true if one tries to hunt for particle formation relevant signatures, since particle formation in semiconductor manufacturing systems is an inherently transient, short-lived process that needs to

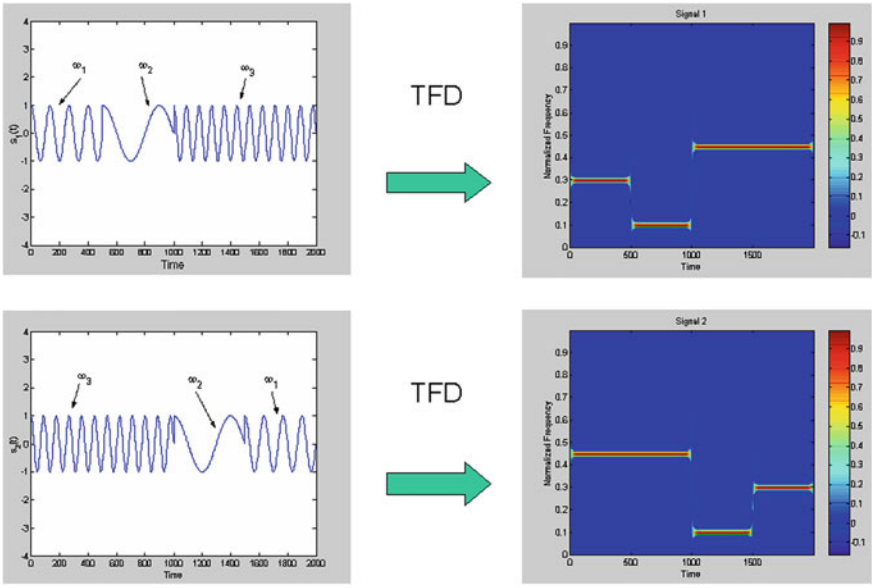


Fig. 5.5 Reduced interference joint time–frequency distribution of the two frequency hopping signals identical to those analyzed in Fig. 5.1

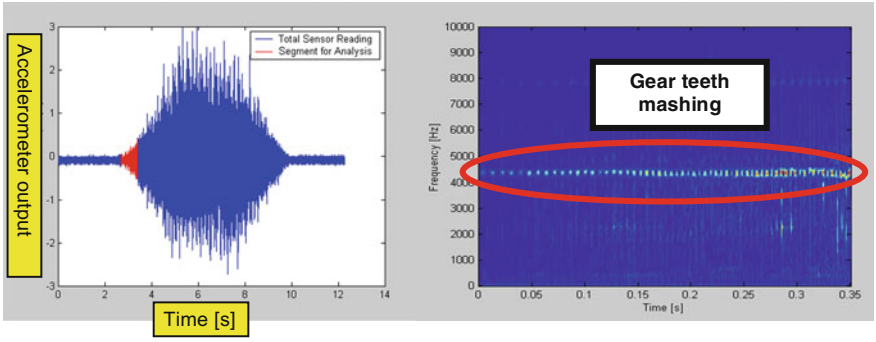


Fig. 5.6 RID of gearbox vibrations emitted during acceleration of the gearbox

be temporally and spatially localized as well as possible. Stationary tools, such as Fourier analysis, or tools with limitations in terms of temporal and frequency resolutions, such as wavelets, may not be able to reveal such minute details buried in often noisy signals. Figure 5.7 shows the binomial kernel-based RID time–frequency distribution (Williams 1996) of vibrations collected during wafer travel on a material handling system known to induce particle formation as wafers passed through it. The strong time and frequency support properties allow one to localize particle-generating features as "bumps" on the material handling guideways, as well as to characterize them (determine their size). Note that after the vibration signals were

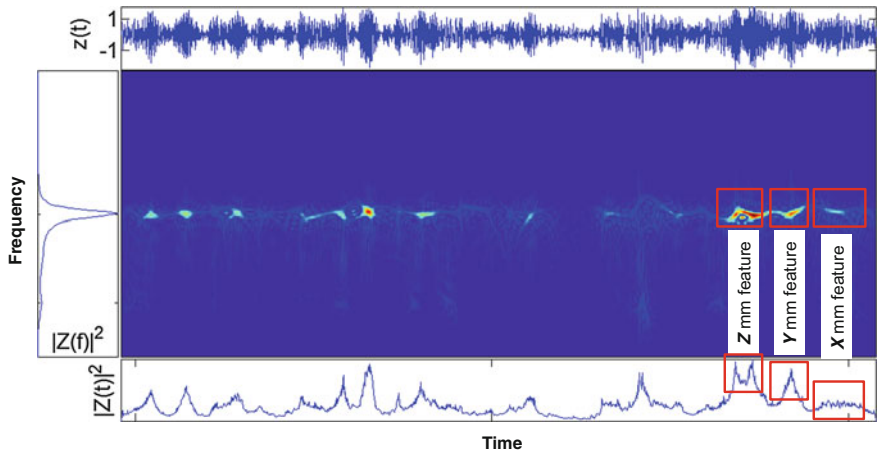


Fig. 5.7 Binomial kernel-based reduced interference time–frequency distribution of vibrations collected from a wafer traveling along a material handling system known to be generating particles as wafers pass through it. Raw time series z(t) of vibrations is shown above the time–frequency distribution, and instantaneous power of the vibrations $|z(t)|^2$ is shown above it, while the corresponding power spectral density $|z(t)|^2$is shown on the right-hand side. Note that no numerical values are reported because of the proprietary nature of the data

collected and analyzed as illustrated in Fig. 5.7, physical inspection of the material handling system confirmed the existence and size of particle-generating features on that system.

5.3.2 *Extraction of Dynamics-Inspired Signatures from Densely Sample Signals Obtained from Semiconductor Manufacturing Tools*

For decades, sensor readings from process tools in semiconductor manufacturing were collected at very low sampling rates, often below 1 Hz. This was sufficient for process control when the underlying semiconductor technology did not require exceptionally tight control and when competition among manufacturers was not as strong as it is today. At such low sampling rates, process dynamics usually could not be observed and therefore, practitioners and researchers focused on characterizing processes via statistical characteristics of the observed signals, including mean values, standard deviations, peak-to-peak values, and occasionally even higher order statistics such as skewness, kurtosis, and entropy. These characteristics were obtained for the entire signal or certain portions of it, as specified by user-defined windows, often requiring significant heuristics of expert knowledge about the process and the machine (for a good survey of sensory signal processing and feature extraction for

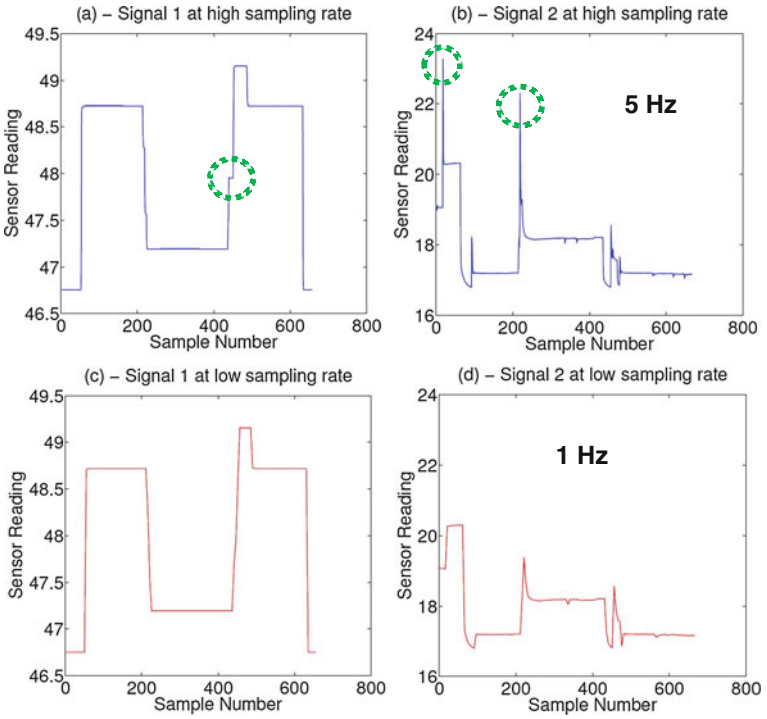


Fig. 5.8 Illustration of significant new features visible at higher sampling grates. Note the additional glitch-like stage in signal 1, as well as notably different transients visible in signal 2, when those signals are sampled at 5 Hz instead of more traditional 1 Hz sampling

CBM in semiconductor manufacturing, one can refer to Chap. 2 of the Ph.D. thesis Yang 2011).

Nevertheless, driven by ever-tightening requirements on the process and product tolerances and by the ever-increasing competition, in the last several years, we can see a strong proliferation of higher sampling rates in modern fabs, reaching 10 Hz and above. At such rates, process dynamics are much more faithfully represented in the signal transients and potentially significant information resides in these parts of the signal, lending value to the ability to automatically and systematically mine those dynamic signatures. Figure 5.8 illustrates remarkable differences in time traces of the same sensor readings obtained from a production tool, in a 300 mm fab, when sampling at 1 Hz and at 5 Hz. Unfortunately, advances in data collection and sampling rates were not accompanied by adequate advances in the processing and utilization of those data. Instead, traditional methods based on expert knowledge-based windowing of signals and extraction of a plethora of statistics from those windows still remain the predominant state-of-the-art technique in fabs today, as illustrated in plot (a) of Fig. 5.9. Though occasionally effective, with 100 s of thousands of such signals now streaming out of a fab, such essentially manual approaches to signal parsing and

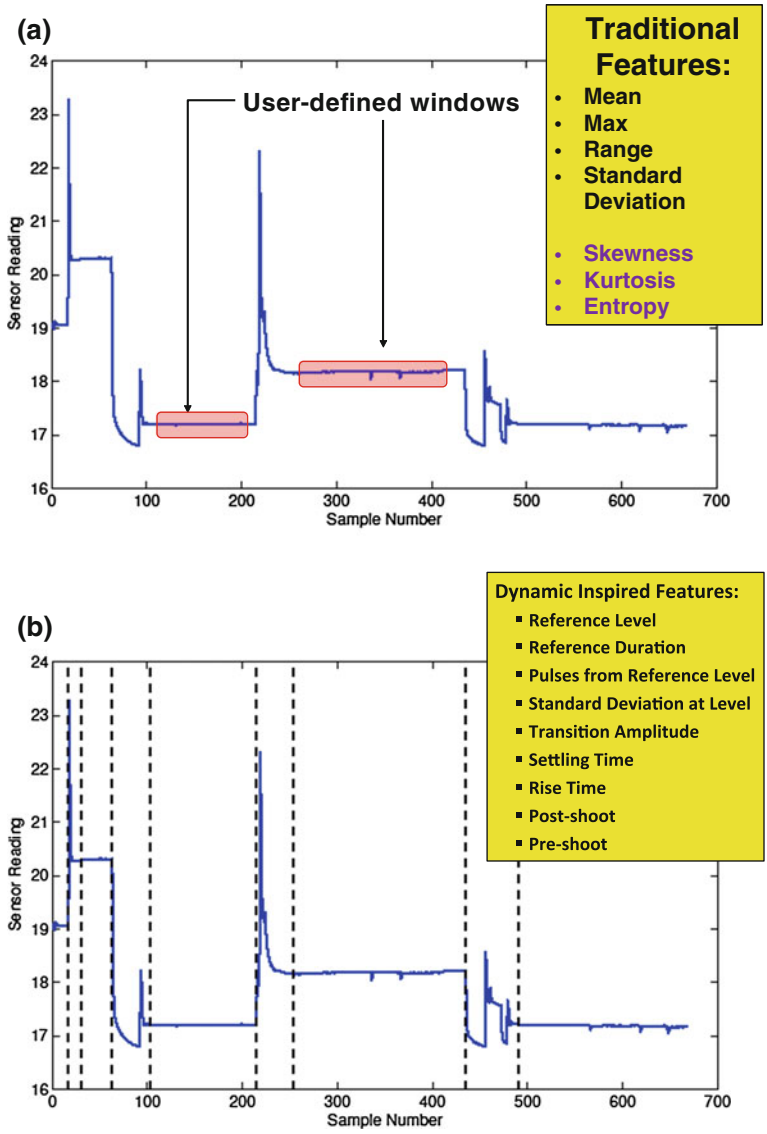


Fig. 5.9 Plot **a** gives illustration of a traditional approach to parsing and extraction of feature from signal waveforms in semiconductor manufacturing (based on process-related knowledge and statistics-inspired features). Plot **b** illustrates the recently proposed automatic method for signal parsing into steady-state and transient segments, with extraction of dynamics-inspired features from the transient signal portions and statistics-inspired features from the steady-state signal portions

extraction of informative features are completely unfeasible and effectively limit one's capabilities for fab-wide data mining for the information that is now available with those high sampling rates.[2]

This significant gap is addressed in a recent paper which introduced a novel methodology for automatic processing of densely sampled signals from semiconductor manufacturing processes into a set of signatures that could be related to the underlying process dynamics (Ul Haq et al. 2016). In that manuscript, the authors propose an elaborate procedure that determines and utilizes noise levels and rates of change (derivatives) of the signal to parse it into sections of steady-state and transient behavior, as illustrated in plot (b) of Fig. 5.9 for the same signal shown in plot (a). Based on that partitioning, traditional statistics-based features, such as mean, standard deviation, kurtosis, skewness, and entropy, could be extracted from the steady-state segments of the signal, while dynamics-based features, such as those defined in IEEE standards (rise-times, overshoots, settling times, duration of transients) (IEEE Standard for Transitions, Pulses and Related Waveforms 2011), could be extracted from the transient signal portions.

The usefulness of this enriched feature set augmented with dynamics-inspired signatures was already illustrated in several applications. Figure 5.10 illustrates superiority of using the features from Ul Haq et al. (2016) for the purpose of tool matching. It shows the most discerning sensory signatures that differentiate three chambers in a thin-film deposition tool used in a major 300 mm fab, as identified using Linear Discriminant Analysis (LDA) (Duda 2001) on the standard feature set extracted using a commercially available software (plot b), and via LDA applied on the augmented feature set obtained using methods from Ul Haq et al. (2016) (plot c). Not only is this separation much clearer when the augmented feature set is used, but the method from Ul Haq et al. (2016) also enables one to clearly identify signal segment that generated the most discerning feature (illustrated in plot a), which can be related to a specific step in the process and then be used to remedy potential problems caused by that mismatch.

Another example of superiority of the feature set obtained using methods proposed in Ul Haq et al. (2016) can be seen in Fig. 5.11. It shows Root-Mean-Squared Errors (RMSE) for several virtual metrology (VM) models constructed using traditional, commercially available features, and the same models constructed using features from Ul Haq et al. (2016). It is clearly visible in Fig. 5.11 that, regardless of what VM model is used, the newly available features yield improvements in terms of VM RMSE. In Ul Haq and Djurdjanovic (2017), a more thorough analysis of VM models in several applications and in terms of several metrics was reported, with all studies consistently pointing to the augmented feature set obtained using methods from Ul Haq et al. (2016) yielding better information for the VM models, as compared to the traditional features.

[2]Informal conversations by the author with several semiconductor manufacturing tool suppliers reveal that they could easily provide chip-makers with even higher sampling rates. Nevertheless, they are unable to see additional value in such high sampling rates because there are no ways to extract additional value from such deluge of data. Effectively, the data avalanche would just become more pronounced, potentially causing more harm than good.

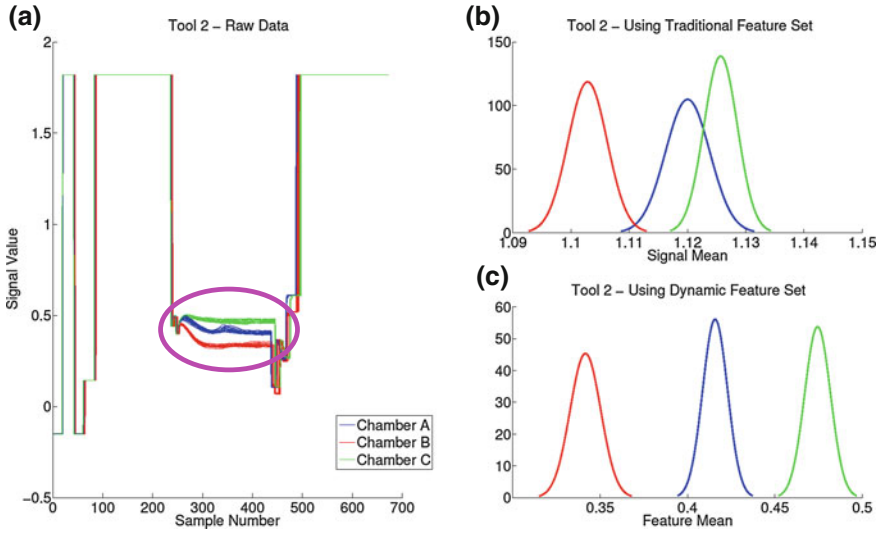


Fig. 5.10 Matching of three chambers in a thin-film deposition tool used in a major 300 mm fab. Plot **a** shows all the raw signals used for tool matching. Plot **b** shows the Gaussian distributions fit to the most discerning feature, as identified by LDA applied to the commercially available feature set. Plot **b** shows the Gaussian distributions fit to the most discerning feature, as identified by LDA applied to the augmented feature set obtained using methods described in Ul Haq et al. (2016). Finally, the magenta ellipse in plot **a** shows the signal segment that generated the most discerning feature obtained using the augmented feature set

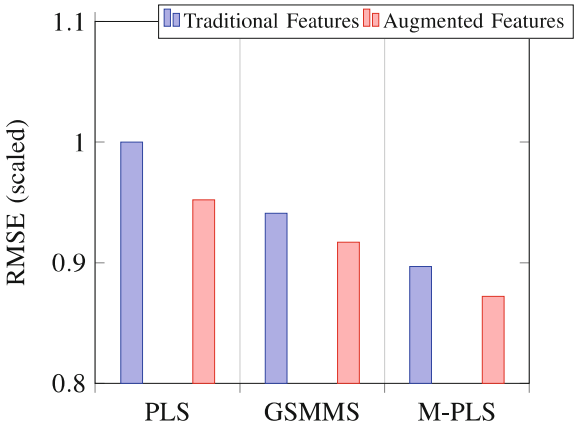


Fig. 5.11 Root-Mean-Squared Errors (RMSE) for several virtual metrology (VM) models applied to estimation of a critical dimension (CD) in an etch process used in a major 300 mm fab. It can be seen that regardless weather the Partial Least Squares (PLS) regression (Höskuldsson and Hoskuldsson 1988), or Growing Structure Multiple Model System (GSMMS) (Bleakie and Djurdjanovic 2016) or Multiple-PLS (M-PLS) (Ul Haq and Djurdjanovic 2017) model form is used for VM, the augmented features yield lower RMSE compared to the traditional features. Note that RMSEs were scaled due to the proprietary nature of the data

5.4 Advances in Condition Modeling in Systems with Partially Observable Conditions

Traditional approaches to modeling of system condition in CBM are based on relating some key informative sensory signatures, such as characteristic frequency of bearing vibrations, root-mean-square value of acoustic emissions obtained from a machine tool, with the condition of the underlying system (Lee et al. 2006). Based on that, abnormal behavioral regimes (fault detection) could be accomplished by quantifying and tracking departure of those signatures away from where they reside during normal behavioral regimes, faults could be diagnosed by matching the current behavior key condition-related signatures with behaviors related to various faults, while prediction of system condition could be accomplished by capturing and extrapolating temporal dynamics of sensory signatures, as illustrated in Figs. 5.2 and 5.3 (Djurdjanovic et al. 2003).

Unfortunately, such a condition-modeling paradigm is unfeasible when we have a highly complex system, such as plasma processes in various semiconductor manufacturing tools (Hutchinson 2002). In such a system, sensors provide information about the condition of a three-dimensional field, but only at discrete points. Thus, even if more sensors can be installed, a full picture about the state of the monitored system between the sensorized points can only be resolved using a highly detailed model, describing the behavior of that field and its interactions with other surrounding subsystems. However, in semiconductor manufacturing, reliable and detailed multi-physics models of an entire plasma-based tool, such as plasma-based etcher or plasma-based deposition tool, do not exist yet. The inability to reliably deduce the full condition of a distributed phenomenon (plasma) leads to situations in which two machines may exhibit very similar sensory signatures, but their conditions are dramatically different—one may be operating normally, while the other one produces poor products. Such situations are indeed encountered in semiconductor manufacturing industry and significantly hamper the adoption of the CBM paradigm in that industry.

Instead, the intuitive relation between the sensor readings and the underlying machine condition can be modeled probabilistically, by associating probabilities of the various levels of system degradation with the observed signatures extracted from the sensor readings. Figure 5.12 illustrates this concept; any given vector of sensory features could be associated with good or bad equipment condition, just with different probabilities. More formally, condition of the monitored system can be modeled using the concept of a hidden Markov model (HMM) (Rabiner 1989), with observable variables of the HMM modeling the signatures extracted from the sensors mounted on the monitored machine, while the hidden states of the HMM model the conditions of that machine. Hence, the machine condition ends up being modeled as a random process that is not directly observable, but is probabilistically related to the available sensor readings (observable HMM variables).

The main challenge in such modeling of system conditions is the need to estimate HMM parameters that describe the dynamics of state transitions and their stochastic

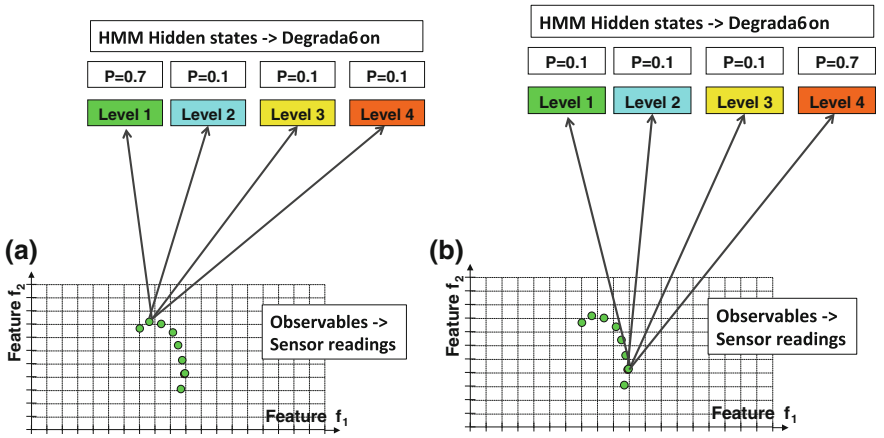


Fig. 5.12 Illustration of HMM-based modeling of degradation dynamics in complex systems. Actual degradation states are hidden and are probabilistically related to the observable variables representing sensor readings obtained from the relevant equipment. Any given observable sensor pattern could be related to any hidden condition of the system, but with different probabilities. Condition monitoring is possible because some sensor signatures are more likely to be related to "good conditions" (plot a), while some are more likely to be related to "bad conditions" (plot b)

relations with the observable variables, using only realizations of observable variables. Once those parameters are estimated, one can use such a model for detection of abnormal system behaviors, diagnosis of reasons for such behavior, and prediction of future system conditions. This is a highly multi-dimensional and multimodal estimation problem, not amenable to traditional gradient-based search methods readily available in the literature (Rabiner 1989). In Cholette and Djurdjanovic (2014), a hybrid optimization relying on a Genetic Algorithm and gradient-based search was proposed for HMM parameter estimation via maximization of the likelihood of the observed symbols. Though significant improvements over the commonly used Baum-Welch approach could be observed, the method did not offer estimates of uncertainties in HMM parameter estimations and thus, model confidence was lacking. This drawback was addressed in Zhang et al. (2016), where a Bayesian approach to HMM parameter estimation was proposed, further improving the parameter estimation, while naturally offering confidence information on the HMM parameter estimates.

In both (Cholette and Djurdjanovic 2014; Zhang et al. 2016), the HMM-based degradation modeling paradigm was applied to monitoring of thin-film deposition process in a Plasma Enhanced Chemical Vapor Deposition (PECVD) tool operating in a major 300 mm fab. A PECVD tool utilizes plasma to lower temperatures at which thin films can be deposited onto silicon wafers. As illustrated in Fig. 5.13, it is composed of a reaction chamber where a set of sapphire spheres hold the wafer above the pedestal, Radio Frequency (RF) plasma generation system, gas delivery system, pendulum valve that controls the chamber pressure, wafer load locks, and

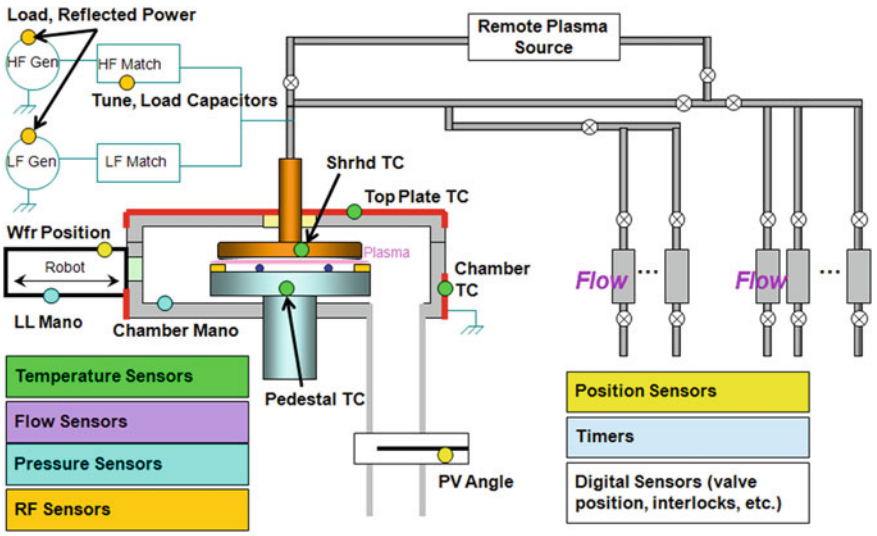


Fig. 5.13 Schematic representation of the PECVD tool considered in (Cholette and Djurdjanovic 2014; Zhang et al. (2016)), and of the sensor readings collected from it

a robotic arm to carry wafers to and from the tool. Load and tune capacitors form matching networks that maximize the power delivered by the RF system to the chamber. RF power characteristics, voltages above the load and tune capacitors, gas flows, top plate, chamber and pedestal temperatures, chamber pressure and the pendulum valve angle are all simultaneously collected at 10 Hz sampling rate (this is order of magnitude higher than standard data collection rates in modern 300 mm fabs). In addition, the valve switching times, and start and end times of all operations are also recorded and aligned with signal traces from all tool systems. The data was continuously collected in a fab for multiple months, along with relevant wafer metrology and information about all maintenance actions done on the tool. It is obviously a highly complex system in which quantum mechanics, fluidics, thermodynamics, and electromagnetic phenomena all occur and interact in an irregular geometry, leading to immense difficulties when traditional Statistical Process Control (SPC) methods are used for monitoring of the tool.[3]

It is therefore not a surprise that the HMM-based degradation model obtained using methods from (Cholette and Djurdjanovic 2014; Zhang et al. 2016) significantly outperforms the purely SPC-based monitoring. Figure 5.14 (adapted from Zhang et al. 2016) shows Receiver Operating Characteristic (ROC) curves for fault detection on the aforementioned PECVD tool data set, as obtained using an HMM-based degradation model from Zhang et al. (2016), and Hotelling's T^2 statistics-based SPC

[3]The data set contained records of 4 fault events, 2 of which kept the tool down for more than 3 weeks each due to difficulties in detecting the abnormal behaviors and finding their root causes.

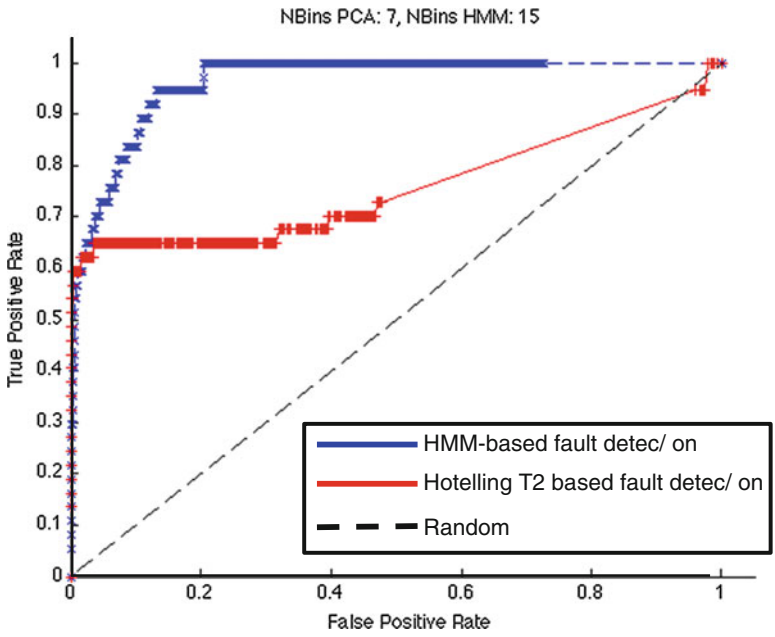


Fig. 5.14 Receiver Operating Characteristic (ROC) curves for fault detection on the PECVD data set described in this chapter, as obtained using HMM-based degradation model obtained using methods from Zhang et al. (2016), and using Hotelling's T^2 SPC chart (Montgomery 2013). Area under the curve (AUC) for the HMM-based fault detection was 0.945, while AUC for the Hotelling T^2-based fault detection was 0.763. Note that ideal AUC is 1, while purely random fault detection yields AUC of 0.5

chart (SPC charting method routinely utilized for multivariate process control in semiconductor manufacturing and in other areas Montgomery 2013).

5.5 Advances in CBM-Based Operational Decision-Making in Semiconductor Manufacturing

Any CBM solution is only as good as the ultimate decision one makes based on that solution. Hence, operational decision-making based on diagnostic or prognostic condition information obtained from the fab floor represents an integral part of CBM-related research. As mentioned earlier, factory physics in a modern fab is highly complex and intractable, which rendered most of the traditional decision-making methods based on tractable assumptions on equipment reliability and interactions between machines obsolete.

Recent dramatic advances in computational technology and ability to sense key variables characterizing factory operations (Work in Progress—WIP levels, cycle

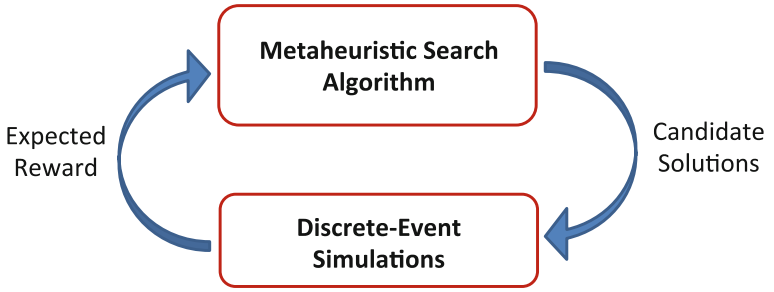


Fig. 5.15 Illustration of simulation-based optimization of operations in manufacturing. System-level cost effects of any candidate decision can be evaluated via multiple replications of discrete event simulations, which can then be used by some metaheuristic to improve operational decisions in the next iteration

times, machine availability status, as well as machine usage levels and age) led to emergence and expansion of research in a fundamentally different direction. Namely, we can see significant recent developments in simulation-based optimization of operations in semiconductor manufacturing, as well as in other areas (for a thorough research survey of operational decision-making in semiconductor manufacturing, with special emphasis on simulation-based optimization, see Chap. 2 of the recent Ph.D. thesis Celen 2016). New, faster computers and parallel computing capabilities enable one to faithfully model the operations without the need for restrictive assumptions, as well as to conduct multiple replications of system simulations and thus evaluate uncertainties associated with any operational decision. This can then be elegantly coupled with metaheuristic optimization methods, such as a Genetic Algorithm, or Tabu search, to guide decisions to ever-improving system-level cost effects, as illustrated in Fig. 5.15. Though such approaches do not guarantee optimality (it is virtually impossible to even characterize optimality under such complex models, let alone guarantee one could reach it), the very character of the simulations and metaheuristics guarantees that over time, one would be improving the decisions, while at the same time having confidence in the effects of those decisions (Celen 2016; Yang 2010).

Of particular interest for semiconductor manufacturing is the ability of such a paradigm to capture interactions between various operational domains in a fab, such as maintenance schedules, production schedules and sequencing, product dispatching, spare part logistics. Sophistication of material handling systems, as well as the ability of most of the tools to execute various operations on various products within very short time spans, makes these interactions in a semiconductor fab much more pronounced than what we see in more traditional manufacturing, such as automotive or petrochemical. On the other hand, ability to model such interactions facilitates concurrent decision-making in those domains, often leading to significant advantageous cost effects of the resulting decisions.

The above-mentioned benefits come at the cost of greatly increased computational requirements, which is why a general method for fab-wide optimization of

operations across all domains still does not exist. Nevertheless, several recent papers did address optimization across multiple operational domains in a semiconductor fab. In the domain of reliability-based maintenance decision-making, seminal works by Sloan and Shantikumar offered an alluring framework for joint decision-making in terms of maintenance and production operations (Sloan and Shanthikumar 2000, 2002), which was extended to simulation-based paradigm[4] in Zhou et al. (2007). The earliest such work in the context of CBM could be found in Li et al. (2007), where dynamics of degradation states of the tool were modeled as perfectly observable Markov chains, with maintenance triggering condition states determined for different product types via discrete event simulations and a Genetic Algorithm-based optimization. The authors report that using product type-dependent CBM policies results in increased yields. However, they overlook the fact that degradation is an operation-dependent process and assume that each operation affects the degradation of the equipment in the same way. That work was extended in Celen and Djurdjanovic (2012) to a multiple-product/multiple-machine manufacturing system, with machine conditions and outgoing product quality degrading according to operating mode-specific models, which better reflects the reality in typical semiconductor fab. Finally, in Celen and Djurdjanovic (2015), the same simulation-based optimization paradigm was applied to joint maintenance and product-sequencing optimization in the same modeling framework proposed in Celen and Djurdjanovic (2012).

Figure 5.16 illustrates the target system considered in Celen and Djurdjanovic (2012, 2015). It is a typical cluster tool frequency encountered in modern fabs, with multiple production chambers and a material handling system that could deliver workpieces (wafers) to any of those chambers at any given time. Each chamber c_i is assumed to be able to conduct several operations o_j, and each product (wafer) type is assumed to require a certain set of operations to be successfully completed in sequence. Condition states of each chamber are assumed to degrade following known operation and chamber-dependent unidirectional Markov chains, with each state being associated with a known probability of completing each of the relevant operations in that chamber (that probability drops with higher levels of degradation). It is assumed that certain numbers of products of each type need to be completed in a mission time T and a simulation-based method was devised to jointly optimize operation and chamber-dependent maintenance triggering states (CBM policy) and a sequence in which the wafers would be released into the system in a way that optimized a customizable cost function rewarding production and penalizing maintenance operations and system downtimes. This policy was combined with the optimal dispatching policy from Li et al. (2007), where it was shown that at any given time, an operation should be dispatched to the chamber (station) that has the highest probability of completing it (this dispatching policy is also quite intuitive).

[4]Note that even though Zhou et al. (2007) did not explicitly address semiconductor manufacturing *per se*, the underlying assumptions about the flexibility of the modeled system and availability of WIP and machine usage information very much make that work highly relevant to semiconductor manufacturing and less to other, more traditional, manufacturing domains.

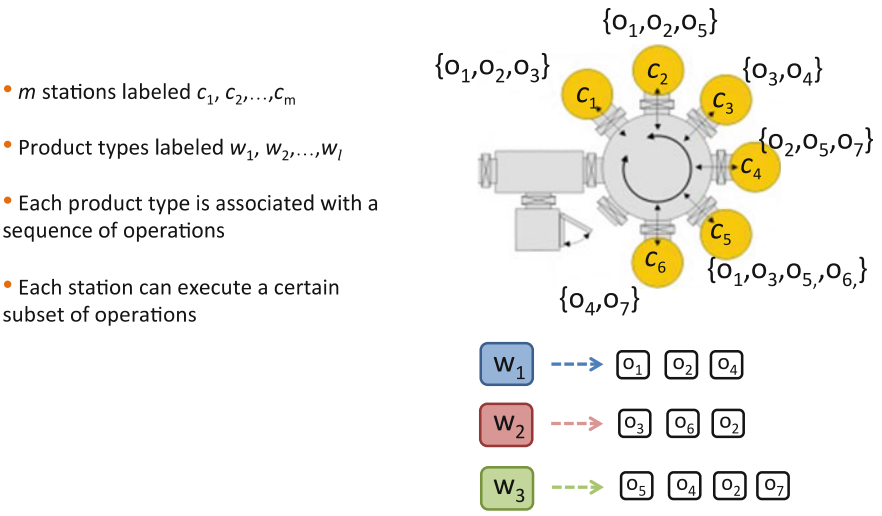


Fig. 5.16 Illustration of a cluster tool considered in Celen and Djurdjanovic (2012, 2015). Each station is assumed to degrade following a station and operation-dependent Markovian chain of condition states, with the probability of successful completion of any given operation in any given station dropping with the progression of degradation states according to a known yield model. Simulation-based optimization was used to concurrently optimize operation-specific and station-specific maintenance triggering states (CBM policy) with the product sequencing (production decision)

Figure 5.17 is adapted from Celen and Djurdjanovic (2015) and shows percentage improvements in terms of system-level operating costs obtained for increasing penalties for unmet production and increasing production goals when the integrated maintenance and product-sequencing decision-making is implemented, as opposed to when the traditional, fragmented decision-making is used (i.e., when product sequence is determined based on some simple heuristic, after which CBM policy is optimized for that product sequence). It is part of an elaborate sensitivity analysis conducted in Celen and Djurdjanovic (2015), which clearly shows that the added decision-making capability within the integrated maintenance and production decision-making framework becomes increasingly beneficial as production goals become more acute (when the system needs to produce more products, or penalties for unmet production are higher). This makes sense because, in the integrated framework proposed in Celen and Djurdjanovic (2015), repair and replacement are not the only actions one can take to fight degradation. Instead, one can reorder production in a way that equipment downtimes due to maintenance can be scheduled at times when they are less intrusive on the production, or when they are cheaper (holidays or weekends).

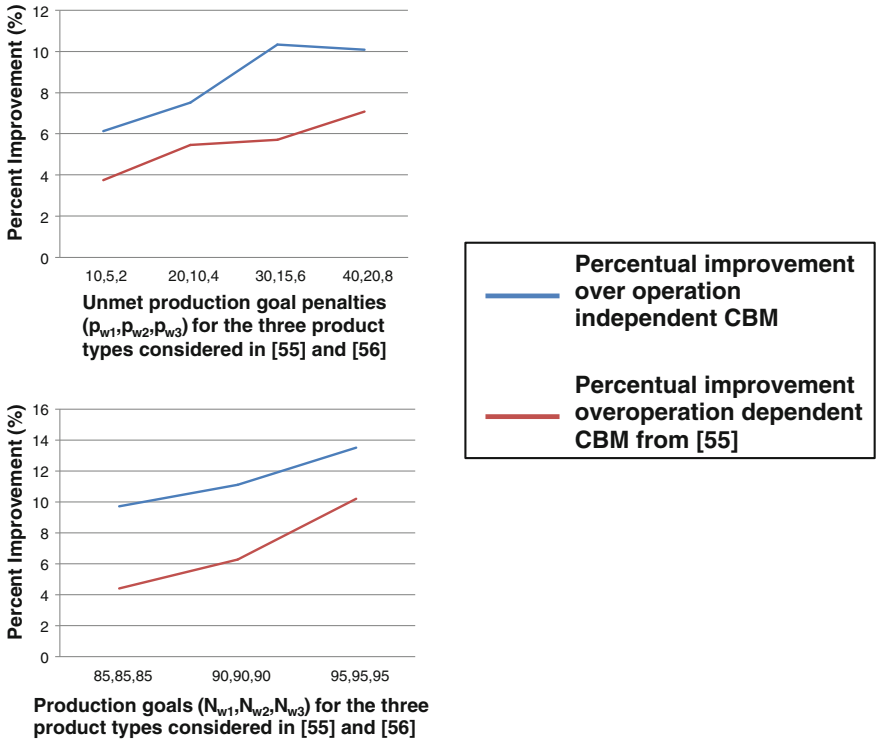


Fig. 5.17 Relative improvements yielded by the integrated maintenance and product-sequencing decision-making framework proposed in Celen and Djurdjanovic (2015) for the generic cluster tool shown in Fig. 5.16, as compared to the traditional operation independent CBM policy, and the operation-dependent CBM policy proposed in Celen and Djurdjanovic (2012). Note that product sequencing for the benchmark policies was obtained separately, by ad hoc grouping product operations. It is visible that relative (percentage) improvements yielded by the integrated operational decision-making increase as production demands become more acute (as penalties for unmet production and production goals become higher)

5.6 Concluding Thoughts

Semiconductor manufacturing of integrated circuits is arguably the most sophisticated and fastest advancing area of high-volume manufacturing, which is why it is laden with all sorts of research challenges. This manuscript offered some of the most recent advances in all aspects of CBM transformation of data to information and further into decisions in this domain. The main message that should be taken from this text is that methods and solutions based on seventeenth- and eighteenth-century mathematics (Fourier, Bernoulli, Jacobi, Newton and others), regardless of how ingenious, cannot always be counted on to solve twenty-first-century problems, such as those encountered in modern fabs. Instead of emphasis on analytical tractability so dearly needed in the olden days, novel directions should use novel computational tech-

nologies to acknowledge and take into account the non-stationary and non-Gaussian nature of signals emanating from sophisticated equipment we see today, model the inherently stochastic relations between the available sensor readings and underlying equipment conditions, understand and estimate uncertainties in models relating sensor readings with equipment conditions and, finally, use those stochastic models along with the associated uncertainties to synchronize information across various portions of the plant into coherent, cost-effective decision. These decisions need to be made with full understanding of the underlying factory physics and interactions between various operational domains (maintenance, production, quality, logistics). The author hopes that this manuscript clearly conveys the aforementioned message and that the interested reader can see numerous avenues for advancing the state of the art.

In the end, it is worth noting that any serious advancement in CBM methodologies and practices in semiconductor manufacturing cannot be done through isolated university research. Truly impacting solutions require that novel methodological advances that usually originate in universities and research institutes be coupled with data and expertise from equipment suppliers, who understand the underlying equipment physics, as well as the fabs who best understand how that equipment is used. All results presented in this chapter came from such collaborations, and there is clear need for such synergies to continue.

References

Beniaminy, I., & Joseph, D. (2002) Reducing the 'No Fault Found' problem: contributions from expert-system methods. In *Proceedings of 2002 IEEE Conference on Aerospace* (Vol. 6, pp. 6-2971–6-2973). Big Sky, MT, USA, 9–16 March 2002.

Bleakie, A., & Djurdjanovic, D. (2016). Growing structure multiple model system for virtual metrology in manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, *29*(2), 79–97.

Boukerche, A., Juca, K., Sobral, J., & Notare, M. (2004). An artificial immune based intrusion detection model for computer and telecommunication systems. *Parallel Computing*, *30*(5–6), 629–646.

Burrus, C. S., Gopinath, R. A., & Haitao, G. (1998). *Introduction to wavelets and wavelet transforms-a primer*. Upper Saddle River, NJ: Prentice Hall.

Cascio, F., Console, L., Gaugliumi, M., Osella, M., Panati, A., Sottano, S., et al. (1999). Generating on-board diagnostics of dynamic automotive systems based on qualitative models. *STC Press, AI Communications*, *12*(1–2), 33–43.

Celen, M. (2016). *Joint maintenance and production operations decision making in flexible manufacturing systems*. Ph.D. Thesis, University of Texas at Austin.

Celen, M., & Djurdjanovic, D. (2012). Operation-dependent maintenance scheduling in flexible manufacturing systems. *CIRP Journal of Manufacturing Science and Technology*, *5*(4), 296–308.

Celen, M., & Djurdjanovic, D. (2015) Integrated maintenance decision-making and product sequencing in flexible manufacturing systems. *ASME Journal of Manufacturing Science and Engineering*, *137*(4), 041006-1–041006-15.

Baraniuk, R. G., & Jones, D. L. (1993). A signal-dependent time-frequency representation: Optimal-kernel design. *IEEE Transactions on Signal Processing*, *41*(4), 1589–1602.

Cholette, M., & Djurdjanovic, D. (2014). Context dependent degradation modeling using hidden markov models. *IIE Transactions on Quality and Reliability Engineering*, *46*(10), 1107–1123.

Cohen, L. (1995). *Time-frequency analysis*. Englewood Cliffs, NJ: Prentice Hall.

Crossman, J. A., Hong, G., & Murphey, Y. L. (2003a). Automotive signal fault diagnostics-part I: Signal fault analysis, signal segmentation, feature extraction and quasi-optimal feature selection. *IEEE Transactions on Vehicular Technology*, *52*(4), 1063–1075.

Crossman, J. A., Hong, G., & Murphey, Y. L. (2003b). Automotive signal fault diagnostics-part II: A distributed agent diagnostic system. *IEEE Transactions on Vehicular Technology*, *52*(4), 1076–1098.

Dasgupta, D., & Gonzalez, F. (2002). An immunity-based technique to characterize intrusions in computer networks. *IEEE Transactions on Evolutionary Computation*, *6*(3), 281–291.

Djurdjanovic, D., Ni, J., & Lee, J. (2002). Time-frequency based sensor fusion in the assessment and monitoring of machine performance degradation. *Proceedings of 2002 ASME International Mechanical Engineering Congress and Exposition*, Paper No. IMECE2002-32032 (pp. 15–22). New Orleans, LA, USA, 17–22 November 2002.

Djurdjanovic, D., Lee, J., & Ni, J. (2003). Watchdog Agent-An infotronics based prognostics approach for product performance degradation assessment and prediction. *International Journal of Advanced Engineering Informatics*, *17*(3–4), 109–125.

Du, R., Elbastawi, M., & Wu, S. M. (1995). Automated Monitoring of Manufacturing Processes - Part 2: Applications. *ASME Journal of Engineering for Industry*, *117*(2), 133–141.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification* (2nd ed., pp. 117–124). Wiley.

Funk, P., & Jackobson, M. (2005). Experience based diagnostics and condition based maintenance within production systems. In D. Mba (Ed.), *Proceedings of the 18th International Congress and Exhibition on Condition Monitoring and Diagnostic Engineering Management (COMADEM)*. http://www.mrtc.mdh.se/publications/0968.pdf.

Gorinewsky, D., Dittmar, K., Milaraswamy, D., & Nwadiogbu, E. (2002). Model-based diagnostics for an aircraft auxiliary power unit. In *Proceedings of the 2002 IEEE Conference on Control Applications* (Vol. 2002, pp. 215–220). Glasgow, Scotland, 18–20 September 2002.

Harmer, P. K., Williams, P. D., Gunsch, G. H., & Lamont, G. B. (2002). An artificial immune system architecture for computer security applications. *IEEE Transactions on Evolutionary Computation*, *6*(3), 252–279.

Hofmeyr, S. A., & Forrest, S. (2000). Architecture for an artificial immune system. *Evolutionary Computation*, *8*(4), 443–473.

Hong, G., Crossman, J. A., & Murphey, Y. L. (2000). Automotive signal diagnostics using wavelets and machine learning. *IEEE Transactions on Vehicular Technology*, *49*(5), 1650–1662.

Hortos, W. S. (2003). An artificial immune system for securing mobile ad hoc networks against intrusion attacks. *Proceedings of SPIE*, *5103*, 74–91.

Höskuldsson, A., & Hoskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, *2*(3), 211–228.

Hutchinson, I. (2002). *Principles of plasma piagnostics* (2nd ed.). Cambridge University Press.

IEEE Standard 181TM–2011. (2011). IEEE standard for transitions, pulses and related waveforms. *IEEE Instrumentation and Measurement Society*.

Jeong, J. (1990). *Time-frequency signal analysis and synthesis algorithms*. Ph.D. Dissertation, University of Michigan, Ann Arbor, MI.

Jeong, J., & Williams, W. J. (1992). Kernel design for reduced interference time-frequency distributions. *IEEE Transactions on Signal Processing*, *40*(2), 402–412.

Kobayashi, T., & Simon, D. L. (2001). A hybrid neural network-genetic algorithm technique for aircraft engine performance diagnostics. *Pentagon Reports*, Report No. A432393, July 2001.

Lee, J., Ni, J., Djurdjanovic, D., Qiu, H., & Liao, H. (2006). Intelligent prognostics tools and e-maintenance. *Computers in Industry*, *57*(6), 476–489.

Lee, J., Lapira, E., Yang, S., & Kao, A. (2013). Predictive manufacturing system-Trends of next-generation production systems. *IFAC Proceedings Volumes*, *46*(7), 150–156.

Li, S., Djurdjanovic, D., & Ni, J. (2007). Optimal condition-based maintenance decision-making for a cluster tool. In *Proceedings of the 2007 Semiconductor Research Corporation (SRC) Technical Conference (TechCon)* (pp. 1–40), Austin, TX, 10–12 September 2007.

Marko, K. A., Feldkamp, L. A., Puskoriusis, G. V. (1990)Automotive diagnostics using trainable classifiers: statistical testing and paradigm selection. In *Proceedings of 1990 IEEE International Joint Conference on Neural Networks* (Vol. 1, pp. 33–38). San Diego, CA, 17–21 June 1990.

Marple, S. L. (1987). *Digital spectral analysis*. Englewood Cliffs, NJ: Prentice Hall.

Montgomery, D. C. (2013). *Introduction to statistical quality control* (7th ed.). Wiley.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE*, *77*(2), 257–286.

Rao, B. K. N. (1996). *Handbook of condition monitoring*. Elsevier Advanced Technology.

Semiconductor Industry Association (SIA). (2015). International technology roadmap for semiconductors; section 7–factory integration. https://www.semiconductors.org/clientuploads/Research_Technology/ITRS/2015/7_2015%20ITRS%202.0%20Factory%20Integration.pdf.

Sloan, T. W., & Shanthikumar, J. G. (2000). Combined production and maintenance scheduling for a multiple-product, single- machine production system. *Production Operations Management*, *9*(4), 379–399.

Sloan, T. W., & Shanthikumar, J. G. (2002). Using in-line equipment condition and yield information for maintenance scheduling and dispatching in semiconductor wafer fabs. *IIE Transactions*, *34*(2), 191–209.

Ul Haq, A., & Djurdjanovic, D. (2017) Virtual metrology based on dynamics-inspired features and divide-and-conquer models. *Elsevier Journal of Process Control*, Paper No. S-17-00270.

Ul Haq, A., Wang, K., & Djurdjanovic, D. (2016). Feature construction for dense inline data in semiconductor manufacturing processes. In *Proceedings of the 2016 IFAC Workshop on Advanced Maintenance Engineering, Service and Technology*. Biarritz, France, Paper No. AMEST16_0053_MS, 19-21 October 2016.

Ville, J. (1948). Theorie et applications de la notation de signal analytique. *Cables et Transmissions*, *2A*(1), 61–74.

Wang, L., Mehrabi, M. G., & Kannatey-Asibu, E. J. (2001). Tool wear monitoring in machining processes through wavelet analysis. *Transactions of NAMRI/SME*, *29*, 399–406.

Wegerich, S. W. (2003). Nonparametric modeling of vibration signal features for equipment health monitoring. In *Proceedings of 2003 IEEE Aerospace Conference* (Vol. 7, pp. 3113–3121). Big Sky, MT, USA, 8–15 March 2003.

Wegerich, S. W. (2004). Similarity based modeling of time synchronous averaged vibration signals for machinery health monitoring. In *Proceedings of 2004 IEEE Aerospace Conference* (Vol. 6, pp. 3654–3662). Big Sky, MT, USA, 6–13 March 2004.

Widmalm, S. E., Djurdjanovic, D., & McKay, D. C. (2003). The dynamic range of TMJ sounds. *Journal of Oral Rehabilitation*, *30*(5), 495–500.

Wigner, E. P. (1932). On the quantum correction for thermodynamic equilibrium. *Physical Review*, *40*(5), 749–759.

Williams, W. J. (1996). Reduced interference distributions: biological applications and interpretations. *Proceedings of IEEE*, *84*(9), 1264–1280.

Yan, W., Gobel, K. F., & Evers, N. (2005). Algorithms for partial discharge diagnostics applied to aircraft wiring. In *Proceedings of 2005 Aging Aircraft Conference* (by Joint Council on Aging Aircraft). http://www.jcaa.us/.

Yang, X-S. (2010). *Engineering optimization–An introduction with metaheuristic applications*. Wiley.

Yang, L. (2011). Methodology of prognostics evaluation for multiprocess manufacturing systems. *Ph.D. Dissertation, University of Cincinnati*.

Yang, X., Shen, J., & Wang, R. (2002). Artificial immune theory based network intrusion detection system and the algorithms design. In *Proceedings of the 2002 International Conference on Machine Learning and Cybernetics* (pp. 73–77), Beijing, China, 4–5 November 2002.

Zhang, D., III Bailey, A. D., & Djurdjanovic, D. (2016). Bayesian identification of hidden markov models and their use for condition-based monitoring. *IEEE Transactions on Reliability*, *65*(3), 1471–1482.

Zhou, J., Djurdjanovic, D., Simmons-Ivy, J., & Ni, J. (2007). Integrated reconfiguration and age-based preventive maintenance decision making. *IIE Transactions*, *39*(12), 1085–1102.

Chapter 6
Multistage Manufacturing Processes: Innovations in Statistical Modeling and Inference

Hsiang-Ling Hsu, Ching-Kang Ing, Tze Leung Lai and Shu-Hui Yu

Abstract Modeling multistage manufacturing processes for fault detection and diagnosis in modern production systems has emerged as a cutting-edge research area at the interface of the engineering and statistical sciences. We give an overview of the developments in this area and describe some recent innovations in statistical modeling and inference associated with these developments.

6.1 Introduction

Modern production engineering typically involves multiple stages in the production of an item, and each stage may involve multiple stations and equipments that can be used in parallel to produce many items simultaneously. Modeling multistage manufacturing processes (MMPs) for fault detection and diagnosis has emerged as "a new area within the boundary of engineering and statistical" sciences, as noted by Ding et al. (2002a) who point out the following developments and methods from both sciences:

> Dimensional quality, represented by product dimension variability, is one of the most critical challenges in industries which use multistage manufacturing processes ... In general, part fixturing, which determines the positions of parts during manufacturing (assembly or machining), directly affects the dimensional quality of final products. ... Recent advancements in fixture design have resulted in significant improvement of fixturing accuracy and

H.-L. Hsu (✉) · S.-H. Yu
Institute of Statistics, National University of Kaohsiung, Kaohsiung, Taiwan, R.O.C.
e-mail: hsuhl@nuk.edu.tw

S.-H. Yu
e-mail: shuhui@nuk.edu.tw

C.-K. Ing
Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan, R.O.C.
e-mail: cking@stat.nthu.edu.tw

T. L. Lai
Department of Statistics, Stanford University, Stanford, CA, USA
e-mail: lait@stanford.edu

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_6

> repeatability. Nevertheless, design-oriented methodology alone cannot guarantee the desired quality of the product due to the complexity and random nature of uncertainties and disturbances in manufacturing processes. Therefore, an effective method for detecting and diagnosing dimensional (multivariate) faults during production, based on in-line measurements, is highly desirable. . . . Methodologies (from statistical science) include pattern recognition of single (univariate) fixture fault through PCA and the identification of multiple simultaneous faults using estimation followed by statistical testing. . . . These diagnostics require the pattern vectors to be obtained through off-line modeling . . . The modeling of pattern vectors for all potential fixturing faults in MMP is much more challenging (than a fixture fault on a single manufacturing station) due to the complex interrelations that exist between stations . . . A process-level model is required to characterize such propagation and accumulation of variation, and to relate the fixture variation to the dimension quality of the final product.

In Sect. 6.2, we review several developments in statistical modeling and inference that have been applied or have potential applications to characterizing "propagation and accumulation of variation" in MMP. Recent developments in "estimation followed by statistical testing" are addressed in Sect. 6.3, in which we summarize the recent work by Ing et al. (2017) and Lai et al. (2017) in this area motivated by fault diagnosis based on quality assurance test data in semiconductor device fabrication. Section 6.4 gives further discussion on fault detection and diagnosis in monitoring multicomponent manufacturing systems with a large number of components as in semiconductor device fabrication and on how these applications benefit from and in turn also inspires innovations in statistical modeling and inference.

6.2 Overview of Statistical Models and Methods for MMPs

6.2.1 State-Space Model for In-Line Manufacturing Process

Figure 1 of Ding et al. (2002a) and Fig. 3 of Zhou et al. (2004) display the framework of fault diagnosis for a product produced by an MMP that consists of m stations. Let $\boldsymbol{x}_k$ denote the fixturing deviation at station k ($\boldsymbol{x}_i = \boldsymbol{0}$ if there is no deviation). The linear state-space model

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{A}_{k-1}\boldsymbol{x}_{k-1} + \boldsymbol{B}_k\boldsymbol{u}_k + \boldsymbol{w}_k \\ \boldsymbol{y}_k &= \boldsymbol{C}_k\boldsymbol{x}_k + \boldsymbol{\varepsilon}_k \end{aligned} \tag{6.1}$$

is used to represent the propagation of deviation for what they call an in-line manufacturing process. The term "in-line" stems from analogy to queueing networks in which jobs departing from one station join a queue at another station. In the state equation of (1) in which $\boldsymbol{w}_k$ and $\boldsymbol{\varepsilon}_k$ are zero-mean random errors (disturbances), $\boldsymbol{A}_{k-1}\boldsymbol{x}_{k-1}$ represents the deviation transformation from station $k-1$ to station k, $\boldsymbol{u}_k$ represents the fixturing deviation contributed by station k, and $\boldsymbol{B}_k$ is the input matrix that depends on the fixture layout at the station. The measurement equation of (1) relates the observation vector $\boldsymbol{y}_k \in \mathbb{R}^d$ to the unobserved state, but the observation may be available only at some station $k \in \{1, \ldots, m\}$. The actual data for off-line

diagnosis consist of a sample of n observed vectors $\boldsymbol{y}_{k,i}$ $(i \leq n_k)$ that are stacked into a $dnm \times 1$ vector $\boldsymbol{Y}$, noting that $n_1 + \cdots + n_m = n$. Some of the n_k can be 0, as in end-of-line sensing (for which measurements are only taken at station m, hence $n_1 = \cdots = n_{m-1} = 0$) considered by Ding et al. (2002a). Zhou et al. (2004) rewrite (1) for the n_k observations at station k (with $n_k \neq 0$) as a linear mixed model (LMM):

$$\boldsymbol{y}_{k,i} = \sum_{j=1}^{k} \boldsymbol{\gamma}_{kj}\boldsymbol{\mu}_j + \sum_{j=1}^{k} \boldsymbol{\gamma}_{kj}(\boldsymbol{u}_{j,i} - \boldsymbol{\mu}_j) + \sum_{j=1}^{k} \boldsymbol{\beta}_{kj}\boldsymbol{w}_{j,i} + \boldsymbol{\varepsilon}_{k,i},\ 1 \leq i \leq n_k, \quad (6.2)$$

where $\boldsymbol{\mu}_j = \mathrm{E}(\boldsymbol{u}_{j,i})$, $\boldsymbol{\gamma}_{kk} = \boldsymbol{C}_k\boldsymbol{B}_k$, $\boldsymbol{\beta}_{kk} = \boldsymbol{C}_k$, and for $1 \leq j < k$,

$$\boldsymbol{\gamma}_{kj} = \boldsymbol{C}_k\boldsymbol{A}_{k-1}\cdots\boldsymbol{A}_j\boldsymbol{B}_j,\ \ \boldsymbol{\beta}_{kj} = \boldsymbol{C}_k\boldsymbol{A}_{k-1}\cdots\boldsymbol{A}_j.$$

They assume known system matrices $\boldsymbol{A}_k$, $\boldsymbol{B}_k$, and $\boldsymbol{C}_k$ that are "determined by the process/product design." The unknown parameters of interest are the fixed effects $\boldsymbol{\mu}_j$ and the covariance matrices $\boldsymbol{V}_j$ of the random effects $\boldsymbol{u}_{j,i} - \boldsymbol{\mu}_j$ associated with the process faults. Assuming Gaussian errors in the state-space model (1) and therefore a multivariate normal LMM for (2), they use MLE and restricted maximum likelihood (REML) to estimate these parameters and other variance components. They also suggest using minimum quadratic unbiased estimation (MINQUE) to reduce the computational load. Letting $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_m^\top)^\top$, they test the null hypothesis $H_0 : \mu_\ell = 0$ to identify a mean shift fault for the ℓth coordinate of the mean vector $\boldsymbol{\mu}$, and $H_0' : \sigma_\ell^2 \leq h_\ell$ to identify the corresponding variance fault, where h_ℓ is the tolerance for process variations. Asymptotic normality of the MLE $\hat{\mu}_\ell$ and $\hat{\sigma}_\ell^2 - h_\ell$ of the MLEs (or MINQUE test statistics) is used to determine critical values for rejection. This is in the spirit of "estimation followed by statistical testing" that we have cited in Sect. 1 from Ding et al. (2002a). The testing, however, is carried out separately for each ℓ without adjustments for multiple testing, presumably because the large number of components in the vector $\boldsymbol{\mu}$ discourages them from carrying out Bonferroni-type corrections that may result in low power. In Sect. 6.3, we describe some recent developments in post-selection multiple testing to address this issue.

6.2.2 *Engineering-Driven Factor Analysis*

Liu et al. (2008) rewrite (2) as $\boldsymbol{Y} = \mathrm{E}\boldsymbol{Y} + \boldsymbol{\Gamma}\boldsymbol{U} + \boldsymbol{v}$, where $\boldsymbol{v}$ stacks $\sum_{j=1}^{k} \boldsymbol{\beta}_{kj}\boldsymbol{w}_{j,i} + \boldsymbol{\varepsilon}_{k,i}$ into a $dnm \times 1$ vector that corresponds to stacking the $\boldsymbol{y}_{k,i}$ into $\boldsymbol{Y}$, $\boldsymbol{U}$ likewise stacks the random effects $\boldsymbol{u}_{j,i} - \boldsymbol{\mu}_j$, and the matrix $\boldsymbol{\Gamma}$ basically aligns the $\boldsymbol{\gamma}_{kj}$ to yield the sums $\sum_{j=1}^{k} \boldsymbol{\gamma}_{kj}(\boldsymbol{u}_{j,i} - \boldsymbol{\mu}_j)$ when it acts as a linear transformation of $\boldsymbol{U}$. Instead of specifying the $\boldsymbol{\gamma}_{kj}$ (and therefore $\boldsymbol{\Gamma}$) explicitly through the process/product design, which may be challenging for complex MMPs, Liu et al. (2008) note that the case of unspecified $\boldsymbol{\Gamma}$ falls in the preview of factor analysis. In factor analysis, the estimated matrix $\hat{\boldsymbol{\Gamma}}$ whose entries are called factor loadings is unique only up to

orthogonal transformations. The usual practice is to multiply $\hat{\boldsymbol{\Gamma}}$ by a suitably chosen orthogonal matrix so that the factor loadings have some optimal statistical features; an example is the popular varimax rotation that maximizes the sum of variances of the squared loadings. Liu et al. (2008) point out, however, that the fault diagnosis problem is about variation sources in key product characteristics, which are grouped to form spatial pattern vectors (SPVs) that are illustrated in their Figs. 1 and 2. Instead of using orthogonal rotations, their "engineering-driven factor analysis" uses oblique rotations to estimate the SPVs of variation sources. This approach uses (a) information criterion or minimum description length in information theory to select the number of factors, followed by (b) engineering knowledge representation to specify spatial patterns and an indicator matrix consisting of 1's and 0's that transforms the description of an MMP into directional information of the SPVs of variation sources in key product characteristics, and (c) evaluation of the oblique rotation matrix by solving an optimization problem that achieves maximum agreement between the estimated SPVs and the spatial patterns specified by engineering knowledge; see Sect. 3 of Liu et al. (2008).

6.2.3 *Stream of Variation: Modeling and Design to Reduce Variation*

Stream of Variation (SoV) methodologies, originally developed for automotive body assembly, were extended to modeling and optimization problems related to the identification and reduction of sources of variation in MMPs in the 2000s. As noted by Shi (2007, pp. 1–4), in the late 1980s, the in-line optical coordination measurement machine (OCMM) was introduced at the end of an automotive body assembly line to measure critical features of the auto-body assembly, using about 100 laser sensors each of which targeted a critical feature. "The tremendous amount of in-line quality data provided significant opportunities for more effective process control," but "with hundreds of quality attributes being measured," using SPC techniques available at that time would invariably detect some out-of-control conditions (which might be false alarms) and how to react to these conditions was a challenging problem "because of the complexity of the process and time-consuming efforts in root-cause identification." Djurdjanovic and Ni (2004) give a review of the progress toward addressing this problem in multistation machining systems and a new proposal, saying: " The selection of measurements in multistation machining systems is currently a slow and error-prone process based on expert human knowledge." They also propose "systematic procedures for synthesizing measurement schemes that carry the most information about the root causes of dimensional machining errors." Their systematic procedures are based on the models described in Sect. 6.2.1 that enable "the use of the achievements of linear control theory and multivariate statistics in *formally* and *systematically* solving the problems related to optimal selection of measurement in multistation machining systems." In particular, Djurdjanovic and Ni (2001)

have already used the rank of the regression matrix of a linearized SoV model connecting the measured workpiece errors and their root causes to define measurement scheme diagnosability, and Ding et al. (2002b) have developed this concept further for MMPs, introducing the notations of within-station, between-station, and overall diagnosability. Djurdjanovic and Ni (2004) also study the problem of SoV-based measurement scheme synthesis (fusion), first using the heuristics for successive measurement removal and then applying genetic algorithms for combinatorial optimization to combine information from the data streams.

Diagnosability is basically related to the selection of measurements/sensors in the SoV design. Chapter 12 of Shi (2007) gives an overview of optimal sensor placement and distribution in multistation processes. It begins with and introduction to coordinate measuring machines (CMMs) and OCMMs:

> A CMM usually consists of a spatial frame that provides the coordinate reference, a mechanical arm that can move along guided tracks, and a probe that retrieves coordinate information when its tip touches the surface of a manufactured workpiece. One disadvantage of CMMs is their low throughput performing the measurement job sequentially, . . . Recent innovations in sensor technology have enabled manufacturers to distribute quality-assurance metrology sensors in multistation manufacturing processes. . . . An OCMM replaces the mechanical arm and the touch probe in a CMM with an optical sensor unit that consists of a laser source and two CCD (charge-coupled device) image sensors. The laser source sheds a beam on the surface of a workpiece, and the CCD sensors detect the reflective laser beam. . . . It is more affordable to deploy multiple optical sensor units and build more OCMM stations, performing parallel measurement jobs of multiple product characteristics.

It then goes on to point out the importance of design of a sensor system for root-cause diagnosis since "a poorly designed sensor system is likely to generate an extensive amount of irrelevant or even conflicting information" that may not even meet the diagnosability conditions. The sensor placement and distribution problem involves how to distribute the sensors to the stations and the location of the sensors at the individual stations. The distribution part can be rephrased as sensor allocation for multistage product inspection to minimize overall cost, including inspection costs, scrap or repair, and warranty costs (which include those caused by false alarm or detection delay). However, "research on sensor distribution for multistation systems, which considers the effectiveness of variation diagnosis, is very limited." The sensor placement problem can be approached via single-station sensing optimality; for example, Wang and Nagarkar (1999) use a D-optimal criterion for sensor placement in a single station and use Powell's direct search method to solve for the optimum.

6.2.4 Integrated Quality and Reliability Analysis for MMPs

SoV methodologies include also component reliability and product quality. The failure of a component leading to its downtime is called "component catastrophic failure," and component reliability information includes not only the catastrophic failure rate but also component degradation such as the wear rate. For product quality, the

event that the manufactured products are out of specifications is of most concern. The multivariate quality statistics of the incoming and outgoing workpieces at the intermediate stations of an MMP, if available through system design, are also useful for fault diagnosis. Chen et al. (2004), Chen and Jin (2005), and Chap. 16 of Shi (2007) describe an integrated approach to system analysis of quality and reliability (QR) for MMPs. In particular, Shi (2007) says:

> In an MMP, each station consists of multiple components. To simplify the problem, these components are assumed to be in series: the catastrophic failure of any component may lead to system catastrophic failure. In an MMP, the final product quality is affected by the accumulation or stack-up of all variations generated at previous stations. Considering the QR-Co-Effect at each station, the variation propagation in product quality leads to the propagation of the interaction between the manufacturing-system component variability and the product quality, which is called the *QR-Chain effect*.

Shi (2007, pp. 393–401) discusses some building blocks for a QR-Chain model and how the model can be used to evaluate system reliability. The QR-chain model makes the following assumptions on the model components:

- Component degradation is modeled by a discrete-time realization of a Gauss–Markov process with constant variance and mean linearly dependent (with non-negative slope) on the component degradation state.
- The number of operation cycles is treated as a discrete-time index.
- The conditional probability that a system component fails during the next operation cycle given that it is working in the current cycle is assumed to be proportional to a linear combination of the squared deviations of the product quality characteristics from the target.

To elucidate the first two assumptions, suppose there are M product quality characteristics $y_j(t)$, $j = 1, \ldots, M$, with target value 0. Let $\boldsymbol{\xi}(t) \in \mathbb{R}^p$ represent the degradation state of the p system components, and let $t_1, t_2, \ldots$ be the times of the operation cycles. Then, the QR-chain model is given by the linear Gaussian state-space model

$$\begin{aligned} \boldsymbol{\xi}(t_{k+1}) &= \boldsymbol{P}_k \boldsymbol{\xi}(t_k) + \boldsymbol{G}_k \boldsymbol{\varepsilon}_k \\ y_j(t_k) &= \eta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}(t_k) + \boldsymbol{\beta}_j^\top \boldsymbol{z}_k + \boldsymbol{\xi}^\top(t_k) \boldsymbol{\Gamma}_j \boldsymbol{z}_k, \ j = 1, \ldots, M, \end{aligned} \tag{6.3}$$

in which $\boldsymbol{P}_k$ and $\boldsymbol{G}_k$ are known matrices, $\boldsymbol{\varepsilon}_k$ are i.i.d. normal with mean $\boldsymbol{0}$ prior to degradation and mean $\boldsymbol{\mu}$ after degradation, $\boldsymbol{z}_k$ is a vector of noise variables, and $\boldsymbol{\Gamma}_j$ is a matrix characterizing the interaction effects between $\boldsymbol{\xi}(t_k)$ and $\boldsymbol{z}_k$.

6.2.5 *Sequential Fault Detection and Diagnosis for MMPs*

In their survey on statistical process control (SPC) for MMPs, Shi and Zhou (2009) point out that it is critical not only to detect process changes but also to determine the root causes of the changes and that "most conventional SPC techniques treat the

multistage system as a whole and lack the capability to discriminate among changes at different stages." Indeed, the history of quality control began with acceptance sampling and Shewhart's control charts that use relatively simple univariate quality characteristics, and their gradual adoption by manufacturing and other industries. Multivariate quality control characteristics and more efficient statistical process control (SPC) schemes such as CUSUM and EWMA represented the next long phase of sustained development, making use of advances in multivariate analysis and sequential analysis in the statistics literature. The past decade witnessed the emergence of a new direction in quality control because of the availability of "big data" for fault detection and diagnosis and because of contemporaneous developments in the statistics literature on high-dimensional data analysis. As noted by Choi et al. (2006) and Wang and Jiang (2009) for multivariate and high-dimensional applications, only a sparse subset of quality characteristics or other variables of interest undergoes abnormal changes that lead to deviations from the state of statistical control. Forward stepwise variable selection, Lasso, adaptive Lasso, least angle regression, and their variants feature prominently in the control charts proposed by Wang and Jiang (2009), Zou and Qiu (2009), Capizzi and Masarotto (2011), and Jiang et al. (2012). These works focus primarily on monitoring changes in the mean vector of the quality characteristics when a large covariance matrix has to be estimated.

For multicomponent systems, fault diagnosis after detection is of critical importance in determining appropriate corrective actions to restore the system to its normal state. A Bayesian approach to fault diagnosis was developed by Tan and Shi (2012), who modified the Bayes procedures introduced by George and McCulloch (1993, 1997) to identify promising subsets of predictors in linear regression models. Tan and Shi's procedure is for diagnosis of mean shifts, identifying which means have shifted and the directions of shifts in multivariate SPC. They use Markov chain Monte Carlo (or more precisely, Gibbs sampling) to implement the Bayesian approach. Earlier, Li and Tsung (2009) applied multiple testing ideas to fault diagnosis in an adjusted Shewhart or CUSUM chart, but unlike the references cited in the preceding paragraph, their procedure does not consider applications to high-dimensional data associated with a large number of components.

In his recent survey of sequential fault detection and diagnosis in complex systems and quality control, target detection and classification from radar, navigation system and network monitoring, Nikiforov (2016) describes the change detection and diagnosis problem as "the generalization of the (classical) quickest change-point detection problem to the case of M post-change hypotheses." He assumes that for a series of independent observations $X_1, X_2, \ldots$, there exists a stopping time ν such that the X_t have density function f_0 for $t < \nu$ and f_j for $t \geq \nu$, $j = 1, \ldots, M$. Let P_ν^j denote such probability measure and E_ν^j the corresponding expectation; the case $\nu = \infty$ is denoted by P_∞. A detection–diagnosis rule is a pair $(T, \hat{j})$, in which T is a stopping time signaling the occurrence of a change-point and $\hat{j}$ is a terminal decision rule identifying the type of change. Extending Lorden's (1971) seminal work for the

case $M = 1$, Nikiforov (2016) introduces the constraint on the average run length (ARL) for false alarm and diagnosis:

$$\mathrm{E}_\infty(T) \geq \gamma \text{ and } \mathrm{E}_1^j(T_{n(h)}) \geq \gamma \text{ for } 1 \leq h \neq j \leq M, \tag{6.4}$$

in which $(T_1, \hat{j}_1)$, $(T_2 - T_1, \hat{j}_2), \ldots$ are i.i.d. copies of $(T, \hat{j})$ and $n(h) = \inf\{i \geq 1 : \hat{j}_i = h\}$. Under this constraint, he derives an asymptotic lower bound for the worst-worst-case detection–diagnosis delay

$$\bar{\mathrm{E}}(T) = \max_{1 \leq j \leq M} \sup_{\nu \geq 1} \{\operatorname{ess\,sup} \mathrm{E}_\nu^j[(T - \nu + 1)^+ | X_1, \ldots, X_{\nu-1}]\}. \tag{6.5}$$

He also develops asymptotically minimax procedures that attain an asymptotic lower bound for (5) subject to the constraint (4) as $\gamma \to \infty$ and uses maximal invariant statistics to get around nuisance parameters in certain problems.

For the quick detection problem that corresponds to $M = 1$, Lai (1995, 1998) has introduced an alternative optimality theory for sequential change-point detection via a comprehensive theory of sequential hypothesis testing using sequential generalized likelihood ratio (GLR) statistics. Whereas Lorden (1971) imposes the classical constraint $\mathrm{E}_\infty(T) \geq \gamma$ on the detection procedures under consideration, Lai replaces this constraint by a constraint on the false alarm rate per unit time, which is also called "maximal local probability of false alarm," defined by

$$\sup_{\nu \geq 1} P_\infty(\nu \leq T \leq \nu + m)/m \leq \alpha. \tag{6.6}$$

The basic underlying idea is that for a stopping time T that has a geometric distribution, $\mathrm{E}_\infty(T) = 1/P_\infty(T = 1) \sim m/P_\infty(T \leq m)$ as $\mathrm{E}_\infty(T) \to \infty$, uniformly in $m = o(\mathrm{E}_\infty T)$. For a general class of window-limited GLR schemes, Lai (1995, 1998, 2001, 2004) has shown that

$$\mathrm{E}_\infty(T) \sim m/P_\infty(T \leq m) \text{ as } \mathrm{E}_\theta(T) \sim \gamma \to \infty$$

if $m/\log\gamma \to \infty$ but $\log m = o(\log\gamma)$, not only when the X_t are independent but also when $\{X_t\}$ is an ergodic Markov chain on a general state space satisfying certain assumptions under P_∞. Moreover, for these rules in such settings, we also have $P_\infty(T \leq m) \sim \sup_{\nu \geq 1} P_\infty(\nu \leq T < \nu + m)$. Lai (1995) proposes the constraint (6) as an alternative to the ARL constraint $\mathrm{E}_\infty(T) \geq \gamma$ for the implementation of these rules for change-point detection in stochastic systems. Besides noting the difficulties in evaluating the ARL in complex systems for which Monte Carlo methods are needed, Lai (1995) also points out that a long expected duration to false alarm in the ARL constraint does not necessarily imply that the probability of having a false alarm prior to some specified time m is small and that it is easy to construct T "with a large mean γ and also having a high probability that $T = 1$." Chakraborti et al. (2001) also note that "the ARL loses much of its attractiveness as a typical summary

if the distribution is skewed (as is often the case)," but explain why the ARL has remained to be a popular measure of chart performance: "Two control charts are often compared on the basis of out-of-control ARL, such that their in-control ARLs are roughly the same. This parallels comparing two statistical tests on the basis of power against some alternative hypothesis when they are roughly of the same size." Lai (1995) uses this hypothesis testing analogy to find better alternatives to the ARL constraint on false alarm and to the out-of-control ARL as a measure of detection delay.

Lai (2000) has extended this approach to the case of general M, thereby providing an alternative to Nikiforov's constraint (4) on the ARL for false detection and diagnosis. He first gives an equivalent representation of (4). Since the X_t are i.i.d. under P_1^j, it follows from Wald's equation that $\mathrm{E}_1^j(T_{n(h)}) = \mathrm{E}_1^j(T)/P_1^j(\hat{j} = h)$ and therefore (4) can be rewritten as

$$\mathrm{E}_\infty(T) \geq \gamma \text{ and } \max_{1\leq j\leq M} \max_{1\leq h\neq j\leq M} P_1^j(\hat{j} = h)/\mathrm{E}_1^j(T) \leq 1/\gamma. \tag{6.7}$$

In addition, instead of the conventional ARL constraint in (4) and (7), he considers the maximal local probability constraints on false detection and false diagnosis:

$$\begin{aligned} &\sup_{\nu\geq 1} P_\infty(\nu \leq T < \nu + m_\alpha) \leq \alpha m_\alpha, \\ &\max_{1\leq j\leq M} \sup_{\nu\geq 1} P_\nu^j(\nu \leq T < \nu + m_\alpha,\ \hat{j} \neq j) \leq \alpha m_\alpha. \end{aligned} \tag{6.8}$$

His approach to sequential detection and diagnosis starts with the theory of sequential multiple hypothesis testing and then converts asymptotically optimal sequential testing procedures to corresponding detection–diagnosis rules.

The post-change hypotheses $H_1, \ldots, H_M$ considered by Nikiforov (2016) and Lai (2000) represent disjoint subsets of the parameter space so that the actual parameter belongs to only one of them. This framework does not cover the setting of multiple data streams studied by Tartakovsky and Veeravalli (2008), Mei (2010) and Xie and Siegmund (2013), who consider monitoring multiple streams of data in applications like cyber network security systems, where only partial locations would detect or be affected by the abrupt intrusion. In these multisensor problems, it is of great importance to integrate information from all data streams and also identify the anomaly from noisy observations. Suppose one observes independent $X_{j,t}$ for the jth data streams at time $t = 1, 2, \ldots$, with the number of streams being M. After an unknown change-point ν, the distributions of the observations from some proportion of the streams get changed. Hence, for the jth data stream, the density function of $X_{j,t}$ is f_0^j for $t < \nu$ and is f_1^j for $t \geq \nu$. This leads to the null hypothesis $H_j : f_1^j = f_0^j$ that the jth data stream is not among those data streams which change their distributions at the change-point ν. Section 6.4 will discuss some recent developments for this problem.

6.3 Post-selection Multiple Testing and Fault Detection–Diagnosis

This section first reviews our recent work on root-cause identification following fault detection in quality assurance (QA) testing in semiconductor device fabrication, particularly the pivotal role played by a novel approach to post-selection multiple testing in this work. It then describes further developments of this approach for sequential fault detection and diagnosis for high-dimensional quality characteristics.

6.3.1 Fault Diagnosis from QA Test Data in Semiconductor Fabrication

Semiconductor devices are fundamental to the electronics industry, and fabrication is the process that converts semiconducting materials into devices or electronic products based on the integrated circuits created in the process. Silicon is almost always used as the semiconducting material, but other semiconducting compounds may also be used for specific applications. The fabrication process involves multiple stages, the first of which is growth of a large piece of crystalline semiconducting material called an ingot. Ingots are then sawed into wafers whose thickness ranges from 0.5 to 1 mm. Subsequent stages include thermal and local oxidation, photolithography, etching, dopant diffusion, ion implantation, and chemical–mechanical planarization processes. After these stages, each wafer contains hundreds of dies (or chips). The dies are separated by scribing and cleaving and then packaged for protection (Grout 2006). Modern semiconductor factories (known as "fabs" or fabrication facilities) are organized into "workcells" so that all necessary equipment for completing a given stage of the process is placed in the same room to reduce the chance of wafer mishandling. The desiderata of an efficient semiconductor manufacturing process are high yield and high throughput, thereby reducing the cost of production, besides high quality. The throughput refers to the number of chips produced per unit time, and yield is the proportion of functionally operational chips per wafer. Quality is achieved by quality assurance (QA) testing after wafer packaging, in addition to wafer testing prior to packaging.

A "barebone" version of wafer testing consists of wafer sorting, which is testing individual dies on the wafer with a test equipment called a wafer prober, and a wafer final test, which is functional testing at the completion of its production. When the wafer prober is in contact with a die, the automated test equipment (ATE) software applies tests on it involving current and voltage measurements in a short time. Failed dies are not packaged, and this saves packaging cost. If a wafer has a large proportion of failed dies, the whole wafer is discarded. There are many enhancements of this barebone version besides new hardware/software developments in performing wafer tests; see Grout (2006) and May and Spanos (2006) for a basic introduction and the Proceedings of annual IEEE SW (Semiconductor Wafer) Test Workshops for

ongoing developments. In particular, May and Spanos (2006) describe (a) applications of multivariate statistical process control (SPC) with model-reference adaptation, (b) applicability of intelligent supervisory control (using "intelligent modeling techniques such as neural networks"), and (c) process and equipment diagnosis using expert systems, neural networks, and algorithms for automation.

Instead of testing for each wafer prior to its packaging, QA testing uses a sample of the packaged wafers to test if the product actually meets the customers' specifications even though the production design already aims at meeting them. Although "micro-testing" prior to packaging should have eliminated markedly defective chips and wafers, the high-throughput and high-yield manufacturing process cannot afford to check these specifications for individual wafers and defers "macro-testing" to QA before delivery to the customers. Section 4 of Ing et al. (2017) shows some wafer QA test datasets and gives boxplots of the differences of the quality characteristics Y_i from the target value a, from a sample $(\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, n$, of inspected wafers and the vector of tools $\boldsymbol{X}_i$ used to manufacture the ith wafer. Labeling all possible tools over successive stages of the multistage wafer manufacturing process by $j = 1, \ldots, p$, x_{ij} is defined as 1 or 0 according to whether tool j is used or not for the ith wafer.

Letting H_0 denote the state of statistical control, we can write $Y_i = a + \varepsilon_i$ under H_0, where $a = \mathrm{E}_0 Y_i$ and ε_i are independent with mean 0 and variance σ^2. The t-statistic $\sqrt{n}(\bar{Y} - a)/s$ can be used to test whether production is out of the state of statistical control, where $s^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})/(n-1)$. For root-cause identification following fault detection, linear regression $Y_i - a = \boldsymbol{\beta}^\top \boldsymbol{x}_i + \varepsilon_i$ is arguably the simplest model as it corresponds to finding the smallest linear subspace of $x_{i1}, \ldots, x_{ip}$ in which $\mathrm{E}Y_i - a$ lies; this formulation is more general than finding the set of j's for which $\beta_j \neq 0$ because it allows multicolinearity among $x_{i1}, \ldots, x_{ip}$. In this model, fault diagnosis is basically a problem of multiple testing on the parameter vector $\boldsymbol{\beta}$ in the regression model

$$Y_i - \bar{Y} = \boldsymbol{\beta}^\top (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) + \varepsilon_i - \bar{\varepsilon}. \tag{6.9}$$

Although $\boldsymbol{x}_i$ typically has a large number of components in applications to multistage wafer manufacturing processes, the unknown parameter vector $\boldsymbol{\beta}$ has certain sparsity properties that make it estimable. The problem falls in the purview of (a) variable selection and parameter estimation, together with (b) post-selection multiple testing, in high-dimensional sparse linear models. To solve this problem, Ing et al. (2017) have developed a new approach to multiple testing following a greedy forward stepwise variable selection procedure, called the *orthogonal greedy algorithm* (OGA), in high-dimensional sparse linear models. Particularly noteworthy for this new approach is that it can maintain the family-wise error rate (FWER) or the overall type I error of mis-identifying a properly functioning tool as defective, whereas previous methods fail to do so because of inadequate adjustment for post-selection testing or a "spillover" of a relevant regressor on an irrelevant one due to their correlations; see Sect. 3.1 and Sect. 5 of Ing et al. (2017). As will be discussed in Sect. 6.4, their major innovation is a test-based variable selection method after the termination of OGA so that error rate guarantees are still applicable to multiple testing after

variable selection, which is of basic importance in MMPs that involve a large number of tools/equipment over multiple stages. Note that Zhou et al. (2004) also formulate fault diagnosis as testing multiple hypotheses concerning the fixed and random effects in the linear mixed model (2) but do not control the FWER in testing these multiple hypotheses, as we have pointed out in Sect. 6.2.1.

6.3.2 High-Dimensional Process Monitoring and Diagnosis

The preceding methodology for post-selection multiple testing has recently been extended by Lai et al. (2017) to the problem of sequential fault detection and diagnosis for high-dimensional quality characteristics, providing a new and considerably more efficient approach to the SPC problem in the big data era than those reviewed in the first paragraph of Sect. 6.2.5. This new approach uses the same principle, described in the fourth paragraph of Sect. 6.2.5, to derive an optimal sequential change-point detection rule from a corresponding sequential test of some composite null hypothesis. For the case of high-dimensional quality characteristics, one has a large number of null hypotheses, one for each quality characteristic. In particular, the fault detection–diagnosis problem in Sect. 6.3.1 involves the null hypotheses

$$H_0 : \mathrm{E}_0(Y_i) = a,\ H_j : \beta_j = 0 \quad \text{for } L \le j \le p, \tag{6.10}$$

in which H_0 is associated with the state of statistical control and β_j is the jth component of the parameter vector $\boldsymbol{\beta}$ in the regression model (9).

Whereas Ing et al. (2017) have developed a new approach to testing the multiple null hypotheses based on a sample $\{(\boldsymbol{X}_i, Y_i),\ 1 \le i \le n\}$ of fixed size n, the first step of Lai et al. (2017) is to extend that approach to group sequential tests of the $p + 1$ null hypotheses in (10). As we have already pointed out in Sect. 6.3.1 and will explain further in Sect. 6.4, regression with more input variables than the sample size requires variable selection to come up with a manageable set $\hat{J}(t)$ as stage t with sample size n_t. Because the set $\hat{J}(t)$ typically changes slowly with the sample size, group sequential methods that only update variable selection and the associated regression parameter estimates when the sample size reaches $n_1, n_2, \ldots, n_T$ have to be used in lieu of fully sequential methods that update whenever a new observation is added to the sample, where T is the number of groups in the group sequential procedure. The second step of Lai et al. (2017) proceeds as in the fourth paragraph of Sect. 6.2.5, transforming the group sequential multiple hypothesis testing procedure to a window-limited group sequential GLR fault detection–diagnosis rule.

6.4 From High-Dimensional Statistical Innovations to MMP and Back

As pointed out in the Abstract and Sect. 4 of Ing et al. (2017), our work in post-selection multiple testing was "motivated by applications to root-cause identification of faults" in semiconductor MMPs "that involve a large number of tools or equipment at each stage." In fact, the semiconductor company that introduced the fault diagnosis problem to us had tried various high-dimensional regression methods to fit the regression model (9) and found OGA introduced by Ing and Lai (2011) to give results that it found most applicable to its data. This illustrates "from high-dimensional statistical innovations to MMP" monitoring and diagnosis in the title of this section. Other examples of this theme are SPC described in the first paragraph of Sect. 6.2.5, which have led to the recent work of Lai et al. (2017) described in Sect. 3.2.

The other part of the title—"and back" from MMP to statistical innovations—is also illustrated by Ing et al. (2017), who were led by the MMP application to develop a relatively complete theory of post-selection multiple testing in linear regression models. This work has subsequently led to a more general theory that is applicable to nonlinear and generalized linear models developed by Lai and Tsang (2017). Such a theory not only broadens the applications to fault diagnosis of MMPs but is also envisioned to help resolve the "irreproducibility/replication crisis" of contemporary science. Because big data from complex experiments in modern science typically require variable/hypothesis selection based on some sparsity principle to make the problem feasible, there is contemporaneous awareness of irreproducible research associated with invalid p-values in post-selection multiple testing; see the editorial articles in *The Economist* (2013, Oct. 19), *American Psychological Association* (Oct. 2015), and *Nature* (Feb. 2016) on "unreliable research," a "reproducible crisis," and "challenges in irreproducible research," respectively.

Extension of these ideas from samples of fixed size to sequential samples involves another level of multiplicity, namely repeated testing besides testing multiple hypotheses. Big data not only result in data-dependent hypothesis/variable selection decisions as discussed above but also have computational issues that become intractable if carried out sequentially over time. The work of Lai et al. (2017) aims at resolving both fundamental issues and is therefore relevant to both parts of the title of this section. A variant of that work is currently in progress and is related to sequential change-point detection and diagnosis for multiple (and in particular, numerous) data streams. As noted in the last paragraph of Sect. 6.2.5, such data streams arise not only from MMPs and multicomponent systems but also in monitoring large networks. The references cited in Sect. 6.2.5 consider the case of a fixed (and relatively small) number M of data streams. We are currently working on the case $M \to \infty$ as $\alpha \to 0$, where α is the maximal local probability constraint on false detection and false diagnosis, as in (8).

Acknowledgements Hsu's research was partially supported by the Ministry of Science and Technology of Taiwan under grant MOST 105-2118-M-390-004. Ing's research was supported by the Science Vanguard Research Program, Ministry of Science and Technology, Taiwan. Lai's research was supported by National Science Foundation grant DMS-1407828. Yu's research was partially supported by the Ministry of Science and Technology of Taiwan under grant MOST 105-2118-M-390-001.

References

Capizzi, G., & Masarotto, G. (2011). A least angle regression control chart for multidimensional data. *Technometrics*, *53*, 285–296.

Chakraborti, S., Van der Laan, P., & Bakir, S. T. (2001). Nonparametric control charts: an overview and some results. *Journal of Quality Technology*, *33*(3), 304–315.

Chen, Y., Jin, J., & Shi, J. (2004). Integration of dimensional quality and locator reliability in design and evaluation of multi-station body-in-white assembly processes. *IIE Transactions*, *36*(9), 827–839.

Chen, Y., & Jin, J. (2005). Quality-reliability chain modeling for system-reliability analysis of complex manufacturing processes. *IEEE Transactions on Reliability*, *54*(3), 475–488.

Choi, S. W., Martin, E. B., Morris, A. J., & Lee, I.-B. (2006). Adaptive multivariate statistical process control for monitoring time-varying processes. *Industrial & Engineering Chemistry Research*, *45*, 3108–3118.

Ding, Y., Ceglarek, D., & Shi, J. (2002a). Fault diagnosis of multistage manufacturing processes by using state space approach. *Journal of Manufacturing Science and Engineering*, *124*, 313–322.

Ding, Y., Ceglarek, D., & Shi, J. (2002b). Diagnosability analysis of multi-station manufacturing processes. *ASME Journal of Dynamic Systems, Measurement, and Control*, *124*, 1–13.

Djurdjanovic, D., & Ni, J. (2001). Stream of variation based analysis and synthesis of measurement schemes in multi-station machining systems. In *Proceedings of the International Mechanical Engineering Congress and Exposition*. New York.

Djurdjanovic, D., & Ni, J. (2004). Measurement scheme synthesis in multi-station machining systems. *Journal of Manufacturing Science and Engineering*, *126*(1), 178–188.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*, 881–889.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*, 339–373.

Grout, I. A. (2006). *Integrated Circuit Test Engineering: Modern Techniques*. New York: Springer.

Ing, C. -K., & Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models, *Statistica Sinica*, *21*, 1473–1513.

Ing, C. -K., Lai, T. L., Shen, M., Tsang, K. W., & Yu, S. -H. (2017). Multiple testing in regression models with applications to fault diagnosis in big data era. *Technometrics*. https://doi.org/10.1080/00401706.2016.1236755.

Jiang, W., Wang, K., & Tsung, F. (2012). A variable-selection-based multivariate EWMA chart for process monitoring and diagnosis. *Journal of Quality Technology*, *44*, 209–230.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems (with discussion and rejoinder). *Journal of the Royal Statistical Society: Series B*, *57*, 613–658.

Lai, T. L. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, *44*(7), 2917–2929.

Lai, T. L. (2000). Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems. *IEEE Transactions on Information Theory*, *46*(2), 595–608.

Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, *11*, 303–408.

Lai, T. L. (2004). Likelihood ratio identities and their applications to sequential analysis. *Sequential Analysis*, *23*, 467–497.

Lai, T. L., Shen, M., & Tsang, K. W. (2017). A new approach to high-dimensional process monitoring and diagnosis. Technical Report, Department of Statistics, Stanford University.

Lai, T. L., & Tsang, K. W. (2017). Post-selection multiple testing and a new approach to test-based variable selection. Technical Report, Department of Statistics, Stanford University.

Li, Y., & Tsung, F. (2009). False discovery rate-adjusted charting schemes for multistage process monitoring and fault identification. *Technometrics*, **51**, 186–205.

Liu, J., Shi, J., & Hu, S. J. (2008). Engineering-driven factor analysis for variation source identification in multistage manufacturing processes. *Journal of Manufacturing Science and Engineering*, *130*(4), 041009.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 1897–1908.

May, G. S., & Spanos, C. J. (2006). *Fundamentals of Semiconductor Manufacturing and Process Control*. Hoboken NJ: Wiley.

Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika*, *97*(2), 419–433.

Nikiforov, I. V. (2016). Sequential detection/isolation of abrupt changes (with discussion and rejoinder). *Sequential Analysis*, *35*(3), 268–301.

Shi, J. (2007). *Stream of Variation Modeling and Analysis for Multistage Manufacturing Processes*. Boca Raton, FL: CRC Press/Taylor & Francis.

Shi, J., & Zhou, S. (2009). Quality control and improvement for multistage systems: A survey. *IIE Transactions*, *41*(9), 744–753.

Tartakovsky, A. G., & Veeravalli, V. V. (2008). Asymptotically optimal quickest change detection in distributed sensor systems. *Sequential Analysis*, *27*(4), 441–475.

Tan, M. H., & Shi, J. (2012). A Bayesian approach for interpreting mean shifts in multivariate quality control. *Technometrics*, *54*, 294–307.

Wang, K., & Jiang, W. (2009). High-dimensional process monitoring and fault isolation via variable selection. *Journal of Quality Technology*, *41*, 247–258.

Wang, Y., & Nagarkar, S. R. (1999). Locator and sensor placement for automated coordinate checking fixture. *Transactions of the ASME, Journal of Manufacturing Science and Engineering*, *121*, 709–719.

Xie, Y., & Siegmund, D. (2013). Sequential multi-sensor change-point detection. *The Annals of Statistics*, *41*, 670–692.

Zhou, S., Chen, Y., & Shi, J. (2004). Statistical estimation and testing for variation root-cause identification of multistage manufacturing processes. *IEEE Transactions on Automation Science and Engineering*, *1*(1), 73–83.

Zou, C., & Qiu, P. (2009). Multivariate statistical process control using Lasso. *Journal of the American Statistical Association*, *104*, 1586–1596.

Part IV
Reliability and Health Management

Chapter 7
Recent Research in Dynamic Screening System for Sequential Process Monitoring

Peihua Qiu and Lu You

Abstract Dynamic Screening problems arise from a variety of applications where we need to sequentially monitor the performance of individuals to detect any malfunction as early as possible. These applications have stimulated much recent research in the literature, and a new methodology called dynamic screening system (DySS) has been developed. By comparing the longitudinal performance of a given individual with that of well-functioning individuals and by sequentially monitoring their difference, DySS can detect their significant difference early so that the potential damage to the given individual can be avoided or reduced. This paper aims to introduce recent research on DySS in different cases, including cases with univariate or multivariate performance variables and cases with independent or correlated observations.

7.1 Introduction

Dynamic screening (DS) problems encompass a wide range of applications where some performance variables (e.g., variables measuring the quality of a product, or risk factors of a disease) need to be sequentially monitored for early detection of faults and/or diseases. The DS problems are important because early detection of faults and/or diseases can warrant timely interventions so that adverse consequences (e.g., airplane crashes, occurrence of stroke, or other deadly diseases) can be prevented or detected at early stages.

To solve the DS problems, one simple method is to construct pointwise confidence intervals of the mean performance variables from observed data of some well-functioning individuals. Then, the longitudinal performance of a given individual can be detected as abnormal if its observations of the performance variables are beyond the confidence intervals. In the longitudinal data analysis (LDA)

P. Qiu (✉) · L. You
Department of Biostatistics, University of Florida, Gainesville, FL 32610, USA
e-mail: pqiu@ufl.edu

L. You
e-mail: you3939@ufl.edu

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_7

literature, there has been some discussion about construction of such confidence intervals (e.g., Ma et al. 2012; Yao et al. 2005; Zhao and Wu 2008). However, this confidence interval approach cannot monitor the given individual in a sequential way, and thus the history data cannot be used efficiently by this approach. Another potential approach to handle the DS problems is related to statistical process control (SPC). By a SPC control chart, we can sequentially monitor the longitudinal performance of an individual (cf., Hawkins and Olwell 1998; Montgomery 2009; Qiu 2014). But, this approach usually does not compare an individual with other individuals regarding their longitudinal performance, and it usually assumes that the observation distribution is unchanged over time when the longitudinal performance of an individual is in-control (IC) or satisfactory, which is often invalid in the DS problems. As an example, the distribution of our blood pressure readings would change when we get older even in cases when we are healthy and do not have any serious cardiovascular diseases. Therefore, both the confidence interval approach and the traditional SPC charts cannot solve the DS problems effectively.

Motivated by the SHARe Framingham Heart study, Qiu and Xiang (2014) suggested a so-called dynamic screening system (DySS) for solving the DS problem in univariate cases. The DySS method combines the strengths of LDA and SPC approaches by comparing the longitudinal performance of a given individual with that of some well-functioning individuals and by sequentially monitoring their difference. This method is designed mainly for cases when observations at different time points are independent. In recent several years, several alternative DySS methods have been proposed for cases when the observations at different time points are correlated and when observations are multivariate. In the next two sections, we will introduce these different versions of the DySS method in details. Some remarks about certain future research problems on this topic will conclude the article in the last section.

7.2 DySS Methods When Observations Are Independent

In this section, we introduce some recent DySS methods for cases when process observations collected at different time points are assumed independent. Univariate cases are discussed in Sect. 7.2.1, multivariate cases are discussed in Sect. 7.2.2, and an improved version is discussed in Sect. 7.2.3.

7.2.1 Univariate Cases

Qiu and Xiang (2014) suggested the first DySS method for univariate cases. This method was mainly discussed in cases when process observations collected at different time points were assumed independent, although correlated data cases were also briefly discussed. The method consists of the following three steps:

1. Estimate the regular longitudinal pattern of the performance variable y from an observed longitudinal dataset of a group of m well-functioning individuals. This dataset is called *IC dataset* hereafter.
2. For a new individual to monitoring, standardize his/her observations using the estimated regular longitudinal pattern obtained in step 1.
3. Monitor the standardized observations of the new individual and give a signal as soon as all available data suggest a significant shift in his/her longitudinal pattern from the estimated regular pattern.

These three steps will be briefly described below.

Assume that the longitudinal observations of the m well-functioning individuals included in the IC dataset follow the model

$$y(t_{ij}) = \mu(t_{ij}) + \sigma(t_{ij})\varepsilon(t_{ij}), \text{ for } j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, m, \tag{7.1}$$

where $t_{ij} \in [0, T]$ are observation times, $y(t_{ij})$ is the jth observation of the ith individual, $\mu(\cdot)$ and $\sigma^2(\cdot)$ are the mean and variance functions of the performance variable $y(\cdot)$, and $\varepsilon(\cdot)$ is the standardized noise with mean 0 and variance 1. Based on the local pth-order polynomial kernel smoothing, Qiu and Xiang (2014) suggested a four-step procedure for estimating $\mu(\cdot)$ and $\sigma^2(\cdot)$ in model (7.1). Their estimators are denoted as $\widehat{\mu}(\cdot)$ and $\widehat{\sigma}^2(\cdot)$, respectively.

For a given individual to monitor, assume that his/her observations are obtained at times $t_j^* \in [0, T]$, for $j = 1, 2, \ldots$. When the performance of that individual is IC, his/her observations should follow model (7.1). So, we define the standardized observations of that individual as

$$\widehat{\varepsilon}(t_j^*) = \frac{y(t_j^*) - \widehat{\mu}(t_j^*)}{\widehat{\sigma}(t_j^*)}, \text{ for } j \geq 1. \tag{7.2}$$

When the performance of the given individual is IC, the standardized observations $\{\widehat{\varepsilon}(t_j^*), j \geq 1\}$ should be independent of each other with the same mean 0 and the same variance 1. If the longitudinal performance of that individual becomes out-of-control (OC), e.g., his/her mean response starts to deviate from the IC mean function $\mu(\cdot)$, then this will be reflected in the distribution of the standardized observations.

Assume that we are interested in detecting an upward mean shift in the original performance variable y for the given individual, then we can apply a conventional control chart for detecting upward mean shifts to the standardized observations. In Qiu and Xiang (2014), an upward cumulative sum (CUSUM) chart was selected. This chart has the charting statistic defined as

$$C_j^+ = \max(0, C_{j-1}^+ + \widehat{\varepsilon}(t_j^*) - k), \text{ for } j \geq 1, \tag{7.3}$$

where $C_0^+ = 0$ and $k > 0$ is an allowance constant. Then, the chart gives a signal of an upward mean shift when

$$C_j^+ > h_C, \tag{7.4}$$

where $h_C > 0$ is a control limit. For detecting a downward or arbitrary shift, a downward or two-sided CUSUM chart can be used. For such CUSUM charts and other alternative control charts that can also be considered here, read Chaps. 3–6 in the book Qiu (2014).

The performance of a control chart, such as the one defined by (7.3)–(7.4), is usually measured by the IC and OC average run lengths (ARLs). However, these measures are appropriate only in cases when the observation times are equally spaced, which is often invalid in the DS applications. To overcome that difficulty, Qiu and Xiang (2014) suggested using the average time to signal (ATS) measure, described as follows. Let ω be a basic time unit, which is the largest time unit that all observation times are its integer multiples. Then, we define

$$n_j^* = t_j^*/\omega, \text{ for } j = 1, 2, \ldots,$$

where $n_0^* = t_0^* = 0$. For an individual whose longitudinal performance is IC, assume that a signal is given at the sth observation time. Then, the expected value of n_s^*, (i.e., $\mathrm{E}(n_s^*)$) is called the IC ATS, denoted as ATS_0. Similarly, for an individual whose longitudinal performance starts to deviate from the regular longitudinal pattern at the time point τ, the value $\mathrm{E}(n_s^*|n_s^* \geq \tau) - \tau$ is called OC ATS, denoted as ATS_1. Then, for the control chart (7.3)–(7.4), the value of ATS_0 can be specified beforehand, and the chart performs better for detecting a shift of a given size if its ATS_1 value is smaller. For the chart (7.3)–(7.4), the value k is often pre-specified, a large k value is good for detecting large shifts, and a small k value is good for detecting small shifts. Commonly used k values include 0.1, 0.2, 0.5, and 1.0. Once k is pre-specified, the value of h_C can be chosen such that a given value of ATS_0 is reached.

7.2.2 *Multivariate Cases*

Qiu and Xiang (2015) proposed a multivariate DySS method. In multivariate cases, we have multiple performance variables that are included in the q-dimensional vector $\mathbf{y}$. In such cases, the model corresponding to the univariate model (7.1) becomes

$$\mathbf{y}(t_{ij}) = \boldsymbol{\mu}(t_{ij}) + \Sigma^{1/2}(t_{ij}, t_{ij})\boldsymbol{\varepsilon}(t_{ij}), \text{ for } j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, m,$$

where $\mathbf{y}(t_{ij}) = (y_1(t_{ij}), y_2(t_{ij}), \ldots, y_q(t_{ij}))'$ is the q-dimensional observation at time t_{ij}, $\boldsymbol{\mu}(t_{ij}) = (\mu_1(t_{ij}), \mu_2(t_{ij}), \ldots, \mu_q(t_{ij}))'$ and $\Sigma(t_{ij}, t_{ij})$ are the mean and covariance matrix of $\mathbf{y}(t_{ij})$, and $\boldsymbol{\varepsilon}(t_{ij}) = (\varepsilon_1(t_{ij}), \varepsilon_2(t_{ij}), \ldots, \varepsilon_q(t_{ij}))'$ is the q-dimensional error term with mean $\mathbf{0}$ and variance $I_{q\times q}$. By the estimation procedure proposed in Xiang et al. (2013), we can obtain estimators of $\boldsymbol{\mu}(t)$ and $\Sigma(t, t)$ that are denoted as $\widehat{\boldsymbol{\mu}}(t)$ and $\widehat{\Sigma}(t, t)$, respectively.

For a new individual to monitor, assume that its observations are obtained at $t_j^* \in [0, T]$, for $j \geq 1$. Then, similar to (7.2), his/her standardized observations are defined as

$$\widehat{\boldsymbol{\varepsilon}}(t_j^*) = \Sigma^{-1/2}(t_j^*, t_j^*)\big[\mathbf{y}(t_j^*) - \widehat{\boldsymbol{\mu}}(t_j^*)\big].$$

Any mean shifts in $\mathbf{y}(t_j^*)$ will be reflected in $\widehat{\boldsymbol{\varepsilon}}(t_j^*)$ and we can use a multivariate control chart to monitor $\{\widehat{\boldsymbol{\varepsilon}}(t_j^*)\}$ in order to detect distributional shifts in $\{\mathbf{y}(t_j^*)\}$.

Qiu and Xiang (2015) adopted the multivariate exponentially weighted moving average (MEWMA) chart for monitoring $\{\widehat{\boldsymbol{\varepsilon}}(t_j^*)\}$. The charting statistic of this chart is

$$\mathbf{E}_j = \lambda\widehat{\boldsymbol{\varepsilon}}(t_j^*) + (1-\lambda)\mathbf{E}_{j-1}, \text{ for } j \geq 1,$$

where $\mathbf{E}_0 = \mathbf{0}$ and $\lambda \in (0, 1]$ is a weighting parameter. It gives a signal when

$$\mathbf{E}_j' \Sigma_{\mathbf{E}_j}^{-1} \mathbf{E}_j > h_E,$$

where $h_E > 0$ is a control limit, and $\Sigma_{\mathbf{E}_j}$ is the covariance matrix of $\mathbf{E}_j$. Since $\text{Var}(\boldsymbol{\varepsilon}(t_j^*))$ is asymptotically $I_{q\times q}$ and observations at different time points are assumed independent, $\Sigma_{\mathbf{E}_j}$ is approximately $\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2j}]I_{q\times q}$, or $\frac{\lambda}{2-\lambda}I_{q\times q}$ when j is large. So, the above expression can be replaced by

$$\frac{2-\lambda}{\lambda[1-(1-\lambda)^{2j}]}\mathbf{E}_j'\mathbf{E}_j > h_E,$$

or $\frac{2-\lambda}{\lambda}\mathbf{E}_j'\mathbf{E}_j > h_E$ for large values of j. When the dimensionality q is large, Qiu and Xiang (2015) suggested using the multivariate control charts that was based on variable selection and discussed in several papers, including Capizzi and Masarotto (2011), Wang and Jiang (2009), Zou and Qiu (2009).

7.2.3 An Improved Version

As mentioned earlier, the observation times are often unequally spaced in the DS problems. In the DySS methods described above, we have accommodated the unequally spaced observation times in the performance evaluation metrics ATS_0 and ATS_1. However, the construction of the control charts (cf., (7.3)–(7.4)) has not accommodated the unequally spaced observation times yet. To overcome this limitation, Qiu et al. (2017) proposed a control chart that takes into account the unequally spaced observation times in its construction, which is introduced below.

To detect mean shifts in the standardized observations $\{\widehat{\varepsilon}(t_j^*),\ j \geq 1\}$, let us consider the following hypothesis testing problem: for a given $j \geq 1$,

$$H_0 : \mu_{\widehat{\varepsilon}(t_j^*)} = 0 \text{ versus } H_a : \mu_{\widehat{\varepsilon}(t_j^*)} = g(t_j^*) \neq 0.$$

At the current time point j, let us consider the following local constant kernel estimation procedure:

$$\underset{a\in R}{\operatorname{argmin}} \sum_{\ell=1}^{j} \left[\widehat{\varepsilon}(t_\ell^*) - a\right]^2 (1-\lambda)^{t_j^* - t_\ell^*}, \tag{7.5}$$

where $\lambda \in (0, 1]$ is a weighting parameter. The solution to a is the local constant kernel estimator of $g(t_j^*)$, which has the expression

$$\widehat{g}_\lambda(t_j^*) = \frac{\sum_{\ell=1}^{j} w_\ell(t_j^*)\widehat{\varepsilon}(t_\ell^*)}{\sum_{\ell=1}^{j} w_\ell(t_j^*)},$$

where $w_\ell(t_j^*) = (1-\lambda)^{t_j^* - t_\ell^*}$. In (7.5), we estimate $g(t_j^*)$ using all observations collected at or before the current time t_j^*, they receive different weights at different time points, and the weights exponentially decay when the related observation times move away from t_j^*. From the weight formula $w_\ell(t_j^*) = (1-\lambda)^{t_j^* - t_\ell^*}$, it can be seen that unequally spaced observation times have been taken into account.

By considering a weighted generalized likelihood ratio test (WGLR), if we define

$$Q_{H_a}(t_j^*; \lambda) = \sum_{\ell=1}^{j} \left[\widehat{\varepsilon}(t_\ell^*) - \widehat{g}(t_j^*)\right]^2 w_\ell(t_j^*)$$
$$Q_{H_0}(t_j^*; \lambda) = \sum_{\ell=1}^{j} \left[\widehat{\varepsilon}(t_\ell^*)\right]^2 w_\ell(t_j^*),$$

then the WGLR test statistic for testing hypotheses in (7.5) is

$$W_\lambda(t_j^*) = Q_{H_0}(t_j^*; \lambda) - Q_{H_a}(t_j^*; \lambda) = \sum_{\ell=1}^{j} \left[2\widehat{\varepsilon}(t_\ell^*) - \widehat{g}(t_\ell^*)\right] \widehat{g}(t_\ell^*) w_\ell(t_j^*).$$

A signal could be triggered at t_j^* if $W_\lambda(t_j^*)$ is large. By noticing the fact that the sequence $\{(W_\lambda(t_j^*), \widehat{g}(t_j^*)), j = 1, 2, \ldots\}$ forms a two-dimensional Markov chain given the design points, the test statistic $W_\lambda(t_j^*)$ can be computed recursively in the following way:

$$W_\lambda(t_j^*) = w_{j-1}(t_j^*) W_\lambda(t_{j-1}^*) + \left[2\widehat{\varepsilon}(t_j^*) - \widehat{g}(t_j^*)\right]\widehat{g}(t_j^*),$$
$$\widehat{g}(t_j^*) = \left[\alpha_{j-1}\widehat{g}(t_{j-1}^*) + \widehat{\varepsilon}(t_j^*)\right]/\alpha_j,$$

where $\alpha_j = \sum_{\ell=1}^{j} w_\ell(t_j^*) = w_{j-1}(t_j^*)\alpha_{j-1} + 1$. Since the distribution of $W_\lambda(t_j^*)$ is changing over time, it usually requires a quite long time for it to reach a steady state. Thus, the following standardized statistic would be preferred here:

$$W_\lambda^*(t_j^*) = [W_\lambda(t_j^*) - E_\lambda(t_j^*)]/\sqrt{V_\lambda(t_j^*)},$$

where $E_\lambda(t_j^*)$ and $V_\lambda(t_j^*)$ are, respectively, the mean and variance of $W_\lambda(t_j^*)$. A recursive algorithm for calculating $E_\lambda(t_j^*)$ and $V_\lambda(t_j^*)$ can also be found in Qiu et al. (2017). Then the chart gives a signal when

$$W_\lambda^*(t_j^*) > h_W,$$

where $h_W > 0$ is a control limit. Proper selection of the parameter λ and the computation of h_W were discussed in Qiu et al. (2017).

7.3 DySS Methods When Observations Are Correlated

The DySS methods described in the previous section are for cases when process observations are independent of each other. In cases when process observations are correlated, they can still be used if their control limits are chosen properly from an IC dataset using numerical approaches such as the bootstrap algorithms. However, they may not be as effective as we would expect because the data correlation is not taken into account in their construction. In this section, we introduce some recent DySS methods proposed specifically for cases when process observations are correlated.

In model (7.1), assume that the covariance function of the longitudinal response $y(t)$ is $V(s,t) = \text{Cov}(y(s), y(t))$, for $s, t \in [0, T]$. By the four-step model estimation procedure in Qiu and Xiang (2014), we can obtain an estimator of $V(s,t)$ from an IC dataset, denoted as $\widehat{V}(s,t)$. For a new individual to monitor, we assume that his/her observations are obtained at times $\{t_j^*, j = 1, 2, \ldots\}$, as in Sect. 7.2. Instead of monitoring the original observations $\{y(t_j^*), j = 1, 2, \ldots\}$, Li and Qiu (2016) suggested monitoring their decorrelated values as follows. Let t_j^* be the current time point. The covariance matrix of $\mathbf{y}_j = (y(t_1^*), y(t_2^*), \ldots, y(t_j^*))'$ can then be estimated by

$$\widehat{\Sigma}_{j,j} = \begin{pmatrix} \widehat{V}(t_1^*, t_1^*) & \cdots & \widehat{V}(t_1^*, t_j^*) \\ \vdots & \ddots & \vdots \\ \widehat{V}(t_j^*, t_1^*) & \cdots & \widehat{V}(t_j^*, t_j^*) \end{pmatrix}.$$

By the Cholesky decomposition, we have

$$\Phi_j \widehat{\Sigma}_{j,j} \Phi_j' = D_j^2,$$

where Φ_j is a $j \times j$ lower triangular matrix with all diagonal elements being 1, and $D_j = \text{diag}\{d_1, \ldots, d_j\}$ is a diagonal matrix with all diagonal elements positive. Let $\widehat{\boldsymbol{\varepsilon}}_j = (\widehat{\varepsilon}(t_1^*), \ldots, \widehat{\varepsilon}(t_j^*))'$ and $\widehat{\varepsilon}(t_\ell^*) = y(t_\ell^*) - \widehat{\mu}(t_\ell^*)$, for $\ell = 1, 2, \ldots, j$. Then, if we define $\mathbf{e}_j^* = D_j^{-1} \Phi_j \widehat{\boldsymbol{\varepsilon}}_j$, we have $\text{Var}(\mathbf{e}_j^*) = I_{j \times j}$. The last element of $\mathbf{e}_j^*$ is denoted as $e^*(t_j^*)$. Then, values in the sequence $\{e^*(t_1^*), e^*(t_2^*), \ldots\}$ are uncorrelated with each other, and they have the common mean 0 and the common variance 1. Li and Qiu

(2016) suggested monitoring this sequence using a CUSUM chart. For instance, to detect an upward mean shift in the original observations, we can use the CUSUM chart

$$C_j^+ = \max(0, C_{j-1}^+ + e^*(t_j^*) - k),$$

where $C_0^+ = 0$ and $k > 0$ is an allowance constant, and the chart gives a signal when

$$C_j^+ > h,$$

where $h > 0$ is a control limit. In Li and Qiu (2016), it has been shown that the data decorrelation described above can be achieved by a recursive computation, which can speed up the computation significantly.

The above data-decorrelation procedure has several limitations, including (i) extensive computation when j is large because computation of a large inverse matrix is involved at each time point, (ii) requirement of a relatively large data storage, and (iii) attenuation of a possible process mean shift as a price to pay for obtaining uncorrelated observations. To partially overcome these limitations, You and Qiu (2017) proposed a modified version of the data-decorrelation procedure. The main idea of the modified version is that instead of decorrelating all history data, we only decorrelate a small portion of the history data observed after the previous time that the related CUSUM chart restarts its charting statistic. Thus, the unnecessary decorrelation for the majority portion of the history data is avoided in this algorithm. To this end, You and Qiu (2017) used the concept of sprint length that was originally defined in Chatterjee and Qiu (2009) as follows:

$$T_j = \begin{cases} 0, \text{ if } C_j^+ = 0, \\ k, \text{ if } C_j^+ \neq 0, \ldots, C_{j-k+1}^+ \neq 0, C_{j-k}^+ = 0. \end{cases}$$

Then, we only need to decorrelate the current residual $\widehat{\varepsilon}(t_j^*)$ with residuals within the sprint length T_j of the current time point t_j^*. It has been shown that the computation of this modified version is much faster than that of the original method.

A multivariate extension of the data-decorrelation method discussed in Li and Qiu (2016) was discussed in Li and Qiu (2017). Again, this multivariate data-decorrelation method can be simplified using the sprint length idea described above, which has not been discussed yet in the literature.

7.4 Conclusions

We have introduced some recent methodologies for solving the DS problems in different cases, including the ones with univariate or multivariate performance variables and the ones with independent or correlated process observations. As the DS problems have broad applications in industries, public health, medical studies, and many other areas, the introduced DySS methods should have a great potential to provide a major statistical tool for properly handling these applications.

The research topic on DySS is still new, and there are many open research problems. For instance, the classical performance measures for control charts, including ARL and ATS (see the related discussions in Sect. 7.2), accommodate the signal times well. But, they cannot reflect the overall false-positive and false-negative performance of the DySS methods. On the other hand, the regular false-positive rate (FPR) and false-negative rate (FNR) cannot accommodate the signal times well. So, a new performance metric is needed for the DySS methods. Also, there could be different covariates involved in the DS problems in practice. The existing DySS methods discussed in this paper have not accommodated such covariates properly yet.

Acknowledgements This research is supported in part by an NSF grant.

References

Capizzi, G., & Masarotto, G. (2011). A least angle regression control chart for multidimensional data. *Technometrics*, *53*(3), 285–296.

Chatterjee, S., & Qiu, P. (2009). Distribution-free cumulative sum control charts using bootstrap-based control limits. *The Annals of Applied Statistics*, *3*, 349–369.

Hawkins, D. M., & Olwell, D. H. (1998). *Cumulative sum charts and charting for quality improvement*. New York: Sprinter.

Li, J., & Qiu, P. (2016). Nonparametric dynamic screening system for monitoring correlated longitudinal data. *IIE Transactions*, *48*(8), 772–786.

Li, J., & Qiu, P. (2017). Construction Of An Efficient Multivariate Dynamic Screening System. *Quality and Reliability Engineering International*, *33*, 1969–1981.

Ma, S., Yang, L., & Carroll, R. J. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statistica Sinica*, *22*, 95–122.

Montgomery, D. C. (2009). *Introduction to statistical quality control*. New York: Wiley.

Qiu, P. (2014). *Introduction to statistical process control*. Boca Raton, FL: Chapman Hall/CRC.

Qiu, P., & Xiang, D. (2014). Univariate dynamic screening system: An approach for identifying individuals with irregular longitudinal behavior. *Technometrics*, *56*(2), 248–260.

Qiu, P., & Xiang, D. (2015). Surveillance of cardiovascular diseases using a multivariate dynamic screening system. *Statistics in Medicine*, *34*(14), 2204–2221.

Qiu, P., Zi, X., & Zou, C. (2017). Nonparametric Dynamic Curve Monitoring. *Technomatrics*, https://doi.org/10.1080/00401706.2017.1361340.

Wang, K., & Jiang, W. (2009). High-dimensional process monitoring and fault isolation via variable selection. *Journal of Quality Technology*, *41*(3), 247–258.

Xiang, D., Qiu, P., & Pu, X. (2013). Nonparametric regression analysis of multivariate longitudinal data. *Statistica Sinica*, *23*, 769–789.

Yao, F., Müller, H. G., & Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, *100*(470), 577–590.

You, L., & Qiu, P. (2017). Fast Computing For Dynamic Screening Systems When Analyzing Correlated Data. Unpublished manuscript.

Zhao, Z., & Wu, W. B. (2008). Confidence bands in nonparametric time series regression. *The Annals of Statistics*, *36*, 1854–1878.

Zou, C., & Qiu, P. (2009). Multivariate statistical process control using lasso. *Journal of the American Statistical Association*, *104*(488), 1586–1596.

Chapter 8
Degradation Analysis with Measurement Errors

Chien-Yu Peng and Hsueh-Fang Ai

Abstract The lifetime information for highly reliable products is usually assessed by a degradation model. When there are measurement errors in monotonic degradation paths, non-monotonic model assumption can lead to contradictions between physical/chemical mechanisms and statistical explanations. To settle the contradiction, this study presents an independent increment degradation-based process that simultaneously considers the unit-to-unit variability, the within-unit variability, and the measurement error in the degradation data. Several case studies show the flexibility and applicability of the proposed models. This paper also uses a separation-of-variables transformation with a quasi-Monte Carlo method to estimate the model parameters. A degradation diagnostic is provided to evaluate the validity of model assumptions.

8.1 Introduction

High-quality products are more frequently being designed with higher reliability and developed in a relatively short period of time. Manufacturers must obtain the product reliability quickly and efficiently with severe time constraints for internal reliability tests. One difficulty with traditional life tests is the lack of sufficient failure-time data to efficiently make inferences about a product's lifetime. Under this situation, if there are quality characteristics (QCs), whose degradation of physical characteristics over time (referred to degradation paths) is related to product reliability, an alternative option is the use of sufficient degradation data to accurately estimate the product's lifetime distribution. For a comprehensive discussion on degradation models, see Nelson (1990), Meeker and Escobar (1998), and the references therein. Other applications of degradation models are in case studies such as the error rates

C.-Y. Peng (✉) · H.-F. Ai
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
e-mail: chienyu@stat.sinica.edu.tw

H.-F. Ai
e-mail: hsuehfang.ai@gmail.com

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_8

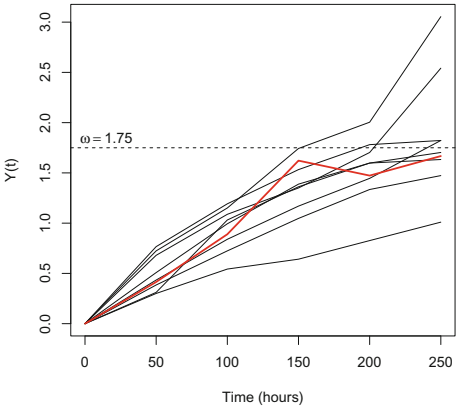


Fig. 8.1 Degradation paths of LED data

of magneto-optic data storage disks, the wear of brake pads, the shelf life of pharmaceuticals, the luminous flux of light bulbs, the resistance change of metal alloys, the power loss of solar cells or power supplies, the voltage of a battery, the propagation of crack size, the power output of integrated circuit devices, the corrosion in a chemical container, the strength of an adhesive bond.

The transformed LED (light-emitting diode) data for an experiment described by Hamada, Wilson, Reese, and Martz (2008, Exercise 8.1) is used as a motivating example. LEDs are widely used in many areas (e.g., traffic signals and full-color displays) because of their high reliability, high brightness, and low power consumption. The QC of an LED device is its light intensity. The LED device is considered to have failed as the light intensity reaches a predefined critical degradation level $\omega = 1.75$. The primary objective of this experiment is to assess the lifetime information for LEDs, such as the mean-time-to-failure ($MTTF$) or the q-quantile. Figure 8.1 shows a plot of the light intensity over 250 h for nine tested units. Light intensity values were recorded every 50 h. Table 8.1 shows the degradation data in 50 h increments for each degradation path. The accuracy and precision of the product's lifetime estimation greatly depends on modeling the degradation paths. Describing the failure-causing mechanism based on the additive accumulation of damage is particularly germane for the light intensity of LED devices. The cumulative damage can be approximated as an independent increment process, and each random shock can be seen as an independent increment in the degradation. See Singpurwalla (1995); van Noortwijk (2009) for more details. The widely used Wiener, gamma, and inverse Gaussian (IG) processes are the special cases of Lévy processes (Barndorff-Nielsen et al. 2001). This approximation presents a physical interpretation of these stochastic processes and provides an applicability to address realistic problems.

Generally speaking, the Gaussian process is usually used to characterize a nonmonotonic degradation path. For instance, Whitmore (1995) proposed a Wiener diffusion process subject to measurement error to model the declining gain of a transistor. Doksum and Normand (1995) presented two Wiener degradation-based processes to connect biomarker processes, event times, and covariates of interest. Peng (2015b)

Table 8.1 Degradation data in 50-hour increment for LED data

Unit	Inspection time interval (h)				
Number	0–50	50–100	100–150	150–200	200–250
1	0.3005	0.2427	0.0985	0.1849	0.1834
2	0.4147	0.4748	0.7324	**−0.1487**	0.1942
3	0.5082	0.4855	0.3944	0.2097	0.1054
4	0.4340	0.4055	0.3301	0.2760	0.3771
5	0.6800	0.4080	0.2620	0.3530	0.8387
6	0.3848	0.3393	0.3231	0.2876	0.1384
7	0.3114	0.7200	0.3304	0.2360	0.0354
8	0.7664	0.4259	0.3404	0.2484	0.0416
9	0.7241	0.4268	0.5920	0.2614	1.0501

provided a comprehensive study of classification problems by using a Gaussian mixture degradation-based process. Further applications using the Gaussian process have been widely investigated by Doksum and Hóyland (1992), Whitmore and Schenkelberg (1997), Padgett and Tomlinson (2004), Peng and Tseng (2013), and the references given therein. However, when the degradation path is strictly monotonic (e.g., increasing or decreasing), the gamma or IG process is commonly used to fit strictly monotonic degradation paths. For example, Bagdonavičius and Nikulin (2000) employed a gamma process with time-dependent explanatory variables as a degradation model. Lawless and Crowder (2004) proposed a gamma process with random effects and covariates to model the crack growth data. Further applications based on the gamma process can be found in Singpurwalla (1995), Singpurwalla (1997), Park and Padgett (2005), van Noortwijk (2009), Tsai et al. (2012), Peng and Cheng (2016), and the references therein. When neither the Wiener nor the gamma degradation-based processes adequately fit strictly monotonic degradation paths (see Wang and Xu (2010), Ye and Chen (2014)), the IG process is an alternative degradation model that can be used to represent the strictly monotonic degradation paths. Peng (2015a) proposed an IG degradation-based process with inverse normal-gamma random effects and derived the corresponding lifetime distribution and its properties.

Figure 8.1 shows that the Gaussian process is a suitable model to describe the non-monotonic LED degradation path. However, in the literature, Fukuda (1991), Chuang et al. (1997), and Yanagisawa and Kojima (2005) theoretically and empirically showed that the light intensity of an LED is strictly monotonic over time. The assumption of a non-monotonic process can lead to contradictions between physical/chemical mechanisms and statistical explanations. Two common approaches are used to reconcile the scenario in this work. Because there is only one negative increment for the second unit (red solid line in Fig. 8.1), which is shown in bold-faced type in Table 8.1, the abnormal measurement point may be an outlier. If the anomaly is excluded from the degradation path, then the strictly monotonic processes can be used to fit the remaining LED data without the suspected point. Otherwise, an alter-

native method is to consider the measurement errors in a strictly monotonic process for degradation modeling, whether the abnormal measurement point is removed or not.

This study uses an independent increment degradation-based process (for strictly monotonic and non-monotonic paths), which is defined in the following section, as a general degradation model that provides a consistent interpretation between physical/chemical mechanisms and statistical explanations. The general degradation model simultaneously considers three sources of variation (i.e., unit-to-unit variability, within-unit variability, and measurement error) in the degradation data. This study uses a separation-of-variables transformation with a quasi-Monte Carlo method to estimate the model parameters and to develop procedures using a bootstrap method to obtain confidence intervals for reliability assessment. Furthermore, model-selection criterion and degradation diagnostic are provided to evaluate the validity of different model assumptions. We use several case studies to illustrate the advantages (i.e., flexibility and applicability) of the proposed degradation models.

8.2 Independent Increment Degradation-Based Process

Let $Y(t|\boldsymbol{\vartheta})$ and $L(t|\boldsymbol{\vartheta})$ with $t \geq 0$, respectively, denote the observed and true values of the QC of a product at time t given the random effects $\boldsymbol{\vartheta}$, where $\boldsymbol{\vartheta}$ denotes the random effects to represent heterogeneity in the degradation paths of distinct units. Assume that

$$Y(t|\boldsymbol{\vartheta}) = L(t|\boldsymbol{\vartheta}) + \varepsilon, \tag{8.1}$$

where the measurement error ε follows a normal distribution with zero mean and variance σ_ε^2 (denoted by $\mathcal{N}(0, \sigma_\varepsilon^2)$), assumed to be independent of the cross time; an independent increment process $\{L(t|\boldsymbol{\vartheta})|t \geq 0\}$ has the following properties: (i) $\Pr\{L(0|\boldsymbol{\vartheta}) = 0\} = 1$; (ii) $L(t|\boldsymbol{\vartheta})$ has independent increments, i.e., $L(t_2|\boldsymbol{\vartheta}) - L(t_1|\boldsymbol{\vartheta})$ and $L(t_4|\boldsymbol{\vartheta}) - L(t_3|\boldsymbol{\vartheta})$ are independent for $0 \leq t_1 < t_2 \leq t_3 < t_4$. Clearly, the original Wiener, gamma, and IG processes are well-known and special cases of the independent increment processes. In addition, the independent increment process, the random effects $\boldsymbol{\vartheta}$, and the measurement error ε are assumed to be mutually independent. For convenience, we call (8.1) an independent increment degradation-based process, which can simultaneously consider the unit-to-unit variability (i.e., random effects), the within-unit variability (i.e., stochastic process), and the measurement error in the degradation data. Note that the assumption of independent increments is not held for the (unconditional) independent increment degradation-based process and this assumption is only used in the (conditional) stochastic process given the random effects $\boldsymbol{\vartheta}$.

For practical applications, we focus on the Wiener, gamma, and IG processes. These common processes can be generalized further by incorporating random effects to describe the unit-to-unit variability of products. The following stochastic processes with random effects have been used as degradation models in many case studies:

$$L(t|\Theta) = \Theta\Lambda(t) + \sigma_W W(t), \quad \Theta \sim \mathcal{N}(\eta, \sigma_\eta^2), \tag{8.2}$$

$$L(t|\tilde{\beta}) \sim \mathcal{G}(\tilde{\alpha}\Lambda(t), \tilde{\beta}), \quad \tilde{\beta}^{-1} \sim \mathcal{G}(\tilde{r}, \tilde{s}) \tag{8.3}$$

and

$$L(t|\mu, \lambda) \sim \mathcal{IG}(\mu\Lambda(t), \lambda\Lambda(t)^2), \quad \delta \equiv \mu^{-1}|\lambda \sim \mathcal{N}(\xi, \sigma_\mu^2/\lambda), \quad \lambda \sim \mathcal{G}(\alpha, \beta), \tag{8.4}$$

where Θ, $\tilde{\beta}$, and $(\mu, \lambda)'$ are the random effects for the Wiener, gamma, and IG processes, respectively; $\Lambda(\cdot)$ is a given, strictly increasing function in time t with $\Lambda(0) = 0$; σ_W is a diffusion coefficient; $W(t)$ is the standard Wiener process (denoted by $\mathcal{N}(0, t)$); the probability density functions (PDFs) of gamma distribution with shape α and scale β (denoted by $\mathcal{G}(\alpha, \beta)$) and IG distribution with mean μ and shape λ (denoted by $\mathcal{IG}(\mu, \lambda)$) are, respectively, given by

$$f(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \exp(-x/\beta), \ x, \alpha, \beta > 0,$$

and

$$f(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \ x, \lambda > 0, \ \mu \in \mathbb{R}.$$

The time-scale transformation function $\Lambda(t)$ represents the degradation measurement of a physical/chemical characteristic over time such as tensile strength, hydrogenation, galvanic action, and fatigue growth. Hence, the time-scale transformation function $\Lambda(t)$ depends on the particular failure-causing mechanism that would directly affect the subsequent performance of the product. The natural conjugate distributions for a specific degradation-based process are used as the random effects $\boldsymbol{\vartheta}$. The assumption of the conjugate distribution not only leads the likelihood function of degradation models to be mathematically tractable, but also provides a satisfactory model fitting for degradation data. This feature can be seen as a computational convenience which shortens the computation time of the likelihood function involving a multiple integral. In addition, similar to the assumptions in Lu and Meeker (1993) and Peng (2015a), assume that $\Pr\{\Theta \leq 0\}$ for Wiener degradation-based processes and $\Pr\{\delta \leq 0\}$ for IG degradation-based processes are negligible to avoid obtaining negative degradation slopes. Incorporating the random effects into the time-scale transformation function $\Lambda(t)$ or other distribution assumptions for the random effects will be studied elsewhere.

The Wiener degradation-based processes with different variation sources are defined as follows:

$$M_1^W: \begin{cases} Y(t|\Theta) = L(t|\Theta) + \varepsilon, \\ L(t|\Theta) = \Theta\Lambda(t) + \sigma_W W(t), \\ \Theta \sim \mathscr{N}(\eta, \sigma_\eta^2), \end{cases} \quad M_2^W: \begin{cases} Y(t) = L(t) + \varepsilon, \\ L(t) = \eta\Lambda(t) + \sigma_W W(t), \end{cases}$$

$$M_3^W: \begin{cases} Y(t|\Theta) = L(t|\Theta) + \varepsilon, \\ L(t|\Theta) = \Theta\Lambda(t), \\ \Theta \sim \mathscr{N}(\eta, \sigma_\eta^2), \end{cases} \quad M_4^W: \begin{cases} Y(t|\Theta) = L(t|\Theta), \\ L(t|\Theta) = \Theta\Lambda(t) + \sigma_W W(t), \\ \Theta \sim \mathscr{N}(\eta, \sigma_\eta^2), \end{cases}$$

$$M_5^W: \begin{cases} Y(t) = L(t) + \varepsilon, \\ L(t) = \eta\Lambda(t), \end{cases} \quad M_6^W: \begin{cases} Y(t) = L(t), \\ L(t) = \eta\Lambda(t) + \sigma_W W(t). \end{cases}$$

The model M_3^W is a classical mixed-effect model for the Wiener degradation-based process. For comparing the other degradation models, the traditional regression model M_5^W is used as a benchmark. The extensions, M_1^W–M_6^W, have been used as the degradation models in Doksum and Hóyland (1992), Doksum and Normand (1995), Whitmore (1995), Peng and Tseng (2009), Cheng and Peng (2012), Si et al. (2012), and Peng (2015b).

The gamma degradation-based processes with different variation sources are defined as follows:

$$M_1^G: \begin{cases} Y(t|\tilde{\beta}) = L(t|\tilde{\beta}), \\ L(t|\tilde{\beta}) \sim \mathscr{G}(\tilde{\alpha}\Lambda(t), \tilde{\beta}), \tilde{\beta}^{-1} \sim \mathscr{G}(\tilde{r}, \tilde{s}), \end{cases} \quad M_2^G: \begin{cases} Y(t) = L(t), \\ L(t) \sim \mathscr{G}(\tilde{\alpha}\Lambda(t), \tilde{\beta}), \end{cases}$$

$$M_3^G: \begin{cases} Y(t|\tilde{\beta}) = L(t|\tilde{\beta}) + \varepsilon, \\ L(t|\tilde{\beta}) \sim \mathscr{G}(\tilde{\alpha}\Lambda(t), \tilde{\beta}), \tilde{\beta}^{-1} \sim \mathscr{G}(\tilde{r}, \tilde{s}), \end{cases} \quad M_4^G: \begin{cases} Y(t) = L(t) + \varepsilon, \\ L(t) \sim \mathscr{G}(\tilde{\alpha}\Lambda(t), \tilde{\beta}). \end{cases}$$

The gamma degradation-based processes, M_1^G–M_4^G, include the models proposed by Bagdonavičius and Nikulin (2000), Lawless and Crowder (2004), Kallen and van Noortwijk (2005), Zhou et al. (2011), Tsai et al. (2012), and Lu et al. (2013) as special cases.

The IG degradation-based processes with different variation sources are defined as follows:

$$M_1^{IG}: \begin{cases} Y(t|\mu, \lambda) = L(t|\mu, \lambda), \\ L(t|\mu, \lambda) \sim \mathscr{IG}(\mu\Lambda(t), \lambda\Lambda(t)^2), \\ \mu^{-1}|\lambda \sim \mathscr{N}(\xi, \sigma_\mu^2/\lambda), \lambda \sim \mathscr{G}(\alpha, \beta), \end{cases} \quad M_2^{IG}: \begin{cases} Y(t|\lambda) = L(t|\lambda), \\ L(t|\lambda) \sim \mathscr{IG}(\mu\Lambda(t), \lambda\Lambda(t)^2), \\ \lambda \sim \mathscr{G}(\alpha, \beta), \end{cases}$$

$$M_3^{IG}: \begin{cases} Y(t|\mu) = L(t|\mu), \\ L(t|\mu) \sim \mathscr{IG}(\mu\Lambda(t), \lambda\Lambda(t)^2), \\ \mu^{-1}|\lambda \sim \mathscr{N}(\xi, \sigma_\mu^2/\lambda), \end{cases} \quad M_4^{IG}: \begin{cases} Y(t) = L(t), \\ L(t) \sim \mathscr{IG}(\mu\Lambda(t), \lambda\Lambda(t)^2), \end{cases}$$

$$M_5^{IG}: \begin{cases} Y(t|\mu, \lambda) = L(t|\mu, \lambda) + \varepsilon, \\ L(t|\mu, \lambda) \sim \mathscr{IG}(\mu\Lambda(t), \lambda\Lambda(t)^2), \\ \mu^{-1}|\lambda \sim \mathscr{N}(\xi, \sigma_\mu^2/\lambda), \lambda \sim \mathscr{G}(\alpha, \beta), \end{cases} \quad M_6^{IG}: \begin{cases} Y(t|\lambda) = L(t|\lambda) + \varepsilon, \\ L(t|\lambda) \sim \mathscr{IG}(\mu\Lambda(t), \lambda\Lambda(t)^2), \\ \lambda \sim \mathscr{G}(\alpha, \beta), \end{cases}$$

$$M_7^{IG}: \begin{cases} Y(t|\mu) = L(t|\mu) + \varepsilon, \\ L(t|\mu) \sim \mathscr{IG}(\mu\Lambda(t), \lambda\Lambda(t)^2), \\ \mu^{-1}|\lambda \sim \mathscr{N}(\xi, \sigma_\mu^2/\lambda), \end{cases} \qquad M_8^{IG}: \begin{cases} Y(t) = L(t) + \varepsilon, \\ L(t) \sim \mathscr{IG}(\mu\Lambda(t), \lambda\Lambda(t)^2). \end{cases}$$

The models proposed by Wang and Xu (2010), Ye and Chen (2014), and Peng (2015a) are the special cases of the IG degradation-based processes M_1^{IG}–M_8^{IG}.

The proposed degradation models M_3^G, M_5^{IG}–M_8^{IG} are new, and they have not been studied in the literature.

8.3 Lifetime Distribution

Let ω denote the critical level for the degradation path. The lifetime, T, of a product can be defined as the first-passage-time (FPT) when the true degradation path $L(t|\boldsymbol{\vartheta})$ crosses the critical level ω, i.e.,

$$T|\boldsymbol{\vartheta} = \inf\{t|L(t|\boldsymbol{\vartheta}) \geq \omega\}.$$

Hence, the PDF, f_T, of T is given by

$$f_T(t) = \int_{\boldsymbol{\Theta}} f_{T|\boldsymbol{\vartheta}}(t|\boldsymbol{\vartheta}) f(\boldsymbol{\vartheta}) \mathrm{d}\boldsymbol{\vartheta},$$

where $f_{T|\boldsymbol{\vartheta}}(\cdot|\boldsymbol{\vartheta})$ denotes the conditional PDF of T given $\boldsymbol{\vartheta}$; $\boldsymbol{\Theta}$ denotes the support of $\boldsymbol{\vartheta}$; and $f(\boldsymbol{\vartheta})$ is a joint PDF of the random effects.

Note that the product's q-quantile, $t(q)$, can be computed by solving $F_T(t(q)) = q$, where $F_T(\cdot)$ stands for the cumulative density function (CDF) of T. For the commonly used Wiener, gamma, and IG degradation-based processes, the corresponding lifetime distributions can be obtained in the following examples.

Example 1 For the Wiener degradation-based processes, the lifetime distribution T_4^W (T_6^W) is the same as T_1^W (T_2^W) with different parameter estimates. More precisely, for degradation model M_1^W, we have $\vartheta = \Theta$. Using Theorem 3.1 of Di Nardo et al. (2001), the conditional PDF of T_1^W given Θ, $f_{T_1^W|\Theta}(t|\Theta)$, satisfies the non-singular second-kind Volterra integral equation

$$f_{T_1^W|\Theta}(t|\Theta) = \Psi(t|0,0) - \int_0^t f_{T_1^W|\Theta}(x|\Theta)\Psi(t|x,\omega)\mathrm{d}x,$$

where

$$\Psi(t|x,y) = \left\{\Theta\dot{\Lambda}(t) + \frac{\omega - \Theta\Lambda(t) - y + \Theta\Lambda(x)}{t-x}\right\} f(t|x,y),$$

$$f(t|x,y) = \frac{1}{\sqrt{2\pi\sigma_W^2(t-x)}} \exp\left\{-\frac{(\omega - \Theta\Lambda(t) - y + \Theta\Lambda(x))^2}{2\sigma_W^2(t-x)}\right\}$$

and $\dot{\Lambda}(t) = \mathrm{d}\Lambda(t)/\mathrm{d}t$. Furthermore, the PDF of T_1^W can be shown as

$$f_{T_1^W}(t) = \frac{\eta\sigma_W^2(\dot{\Lambda}(t)t - \Lambda(t)) + \omega(\sigma_\eta^2\dot{\Lambda}(t)\Lambda(t) + \sigma_W^2)}{\sqrt{2\pi(\sigma_\eta^2\Lambda(t)^2 + \sigma_W^2 t)^3}} \exp\left\{-\frac{(\omega - \eta\Lambda(t))^2}{2(\sigma_\eta^2\Lambda(t)^2 + \sigma_W^2 t)}\right\}$$
$$- \int_0^1 \int_0^t f_{T_1^W|\Theta}(x|\eta + \sigma_\eta\Phi^{-1}(z))\Psi(t|x,\omega)\mathrm{d}x\mathrm{d}z,$$

where $\Phi(\cdot)$ denotes the CDF of $\mathcal{N}(0,1)$. Note that under specific conditions, Si et al. (2012) dropped the integral term from the above equation to approximate the PDF of the lifetime distribution.

Example 2 For the gamma degradation-based processes, the lifetime distribution T_3^G (T_4^G) is the same as T_1^G (T_2^G) with different parameter estimates. For degradation model M_3^G, we have $\vartheta = \tilde{\beta}$. Following the same procedure of Tsai et al. (2012), the CDF of the lifetime distribution T_3^G, $F_{T_3^G}(t)$, can be written in the form

$$F_{T_3^G}(t) = \mathbb{B}\left(\frac{1}{\tilde{s}\omega + 1}; \tilde{r}, \tilde{\alpha}\Lambda(t)\right), \tag{8.5}$$

where $\mathbb{B}(x; a, b)$ denotes the regularized (incomplete) beta function.

Example 3 For the IG degradation-based processes, the lifetime distribution T_5^{IG} (T_6^{IG}, T_7^{IG}, T_8^{IG}) is the same as T_1^{IG} (T_2^{IG}, T_3^{IG}, T_4^{IG}) with different parameter estimates. For degradation model M_5^{IG}, we have $\boldsymbol{\vartheta} = (\mu, \lambda)'$. From the equation (6) of Peng (2015a), the CDF of the lifetime distribution T_5^{IG}, $F_{T_5^{IG}}(t)$, is given by

$$F_{T_5^{IG}}(t) = \sqrt{\frac{\beta}{2\pi}} \frac{\Gamma(\alpha + 1/2)\Lambda(t)}{\Gamma(\alpha)} \int_\omega^\infty y^{-3/2}(\sigma_\mu^2 y + 1)^{-1/2} \left(1 + \frac{\beta(\xi y - \Lambda(t))^2}{2y(\sigma_\mu^2 y + 1)}\right)^{-(\alpha+1/2)} \mathrm{d}y. \tag{8.6}$$

To assess the lifetime information of products, the parameters in the independent increment degradation-based process are needed to be estimated in practical applications. Hence, a quasi-Monte Carlo-type method is used to estimate these unknown parameters in the following section. Afterward, the confidence intervals (CIs) of a product's lifetime information can be easily constructed by using the bias-corrected percentile bootstrap method.

8.4 Parameter Estimation and Confidence Intervals

8.4.1 Likelihood Function

Let $\boldsymbol{\theta}$ be a column vector of all the parameters to be estimated in the independent increment degradation-based process. Let $Y_i(t_{i,j}|\boldsymbol{\vartheta})$ and $L_i(t_{i,j}|\boldsymbol{\vartheta})$, respectively, denote the observed and true degradation path of the ith unit at time $t_{i,j}$ given the random effects $\boldsymbol{\vartheta}$ with $t_{i,0} = 0$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, where n and m_i denote the number of sample size and measurements of the ith unit, respectively. This means that every degradation path can be observed at different inspection times. For simplicity, let $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,m_i})'$, $Y_{i,j} = Y_i(t_{i,j}|\boldsymbol{\vartheta}) - Y_i(t_{i,j-1}|\boldsymbol{\vartheta})$, $\boldsymbol{L}_i = (L_{i,1}, \ldots, L_{i,m_i})'$, $L_{i,j} = L_i(t_{i,j}|\boldsymbol{\vartheta}) - L_i(t_{i,j-1}|\boldsymbol{\vartheta})$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,m_i})'$, where $\varepsilon_{i,1}, \ldots, \varepsilon_{i,m_i}$ are the measurement errors for the ith unit at times $t_{i,1}, \ldots, t_{i,m_i}$ and are i.i.d. normal distribution with zero mean and variance σ_ε^2. The observed increments $\boldsymbol{Y}_i$ can be written as

$$\boldsymbol{Y}_i = \boldsymbol{L}_i + \tilde{\boldsymbol{\varepsilon}}_i, \tag{8.7}$$

where $\tilde{\boldsymbol{\varepsilon}}_i = (\tilde{\varepsilon}_{i,1}, \ldots, \tilde{\varepsilon}_{i,m_i})' \equiv C_i \boldsymbol{\varepsilon}_i$ and C_i is a $m_i \times m_i$ lower bidiagonal matrix with $c_{j,j} = 1$ and $c_{k,k-1} = -1$ for $k = 2, \ldots, m_i$. Note that $\tilde{\varepsilon}_{i,1}, \ldots, \tilde{\varepsilon}_{i,m_i}$ are not independent of each other since every $\tilde{\varepsilon}_{i,j}$ depends on $\tilde{\varepsilon}_{i,j-1}$. By using the property of independent increments, the likelihood function of $\boldsymbol{\theta}$ for the ith degradation path can be expressed as

$$l_i(\boldsymbol{\theta}) = \int_{\boldsymbol{\Theta}} \int_{\mathscr{S}_i} \prod_{j=1}^{m_i} f_{L_{i,j}}(y_{i,j} - \tilde{\varepsilon}_{i,j}|\tilde{\boldsymbol{\varepsilon}}_i, \boldsymbol{\vartheta}) f(\tilde{\boldsymbol{\varepsilon}}_i) f(\boldsymbol{\vartheta}) \mathrm{d}\tilde{\boldsymbol{\varepsilon}}_i \mathrm{d}\boldsymbol{\vartheta}, \tag{8.8}$$

where $\mathscr{S}_i = \{\tilde{\boldsymbol{\varepsilon}}_i \in \mathbb{R}^{m_i} | -\infty < \tilde{\varepsilon}_{i,j} < y_{i,j}\}$; $f_{L_{i,j}}(\cdot)$ denotes the PDF of the independent increment process; $f(\tilde{\boldsymbol{\varepsilon}}_i)$ is the PDF of the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\sigma_\varepsilon^2 C_i C_i'$ (i.e., $\tilde{\boldsymbol{\varepsilon}}_i \sim \mathscr{N}_{m_i}(\mathbf{0}, \sigma_\varepsilon^2 C_i C_i')$). Hence, the log-likelihood function of $\boldsymbol{\theta}$ for the general degradation model is

$$\mathscr{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln l_i(\boldsymbol{\theta}). \tag{8.9}$$

Example 4 For degradation model M_5^{IG} with $\Lambda(t) = t^\gamma$, we have $\boldsymbol{\vartheta} = (\mu, \lambda)'$, $\boldsymbol{\theta} = (\xi, \sigma_\mu, \alpha, \beta, \sigma_\varepsilon, \gamma)'$ and

$$f_{L_{i,j}}(y_{i,j}|\boldsymbol{\vartheta}) = \sqrt{\frac{\lambda \Lambda_{i,j}^2}{2\pi y_{i,j}^3}} \exp\left\{-\frac{\lambda}{2y_{i,j}}\left(\frac{y_{i,j}}{\mu} - \Lambda_{i,j}\right)^2\right\},$$

where $\Lambda_{i,j} = \Lambda_i(t_{i,j}) - \Lambda_i(t_{i,j-1})$. The joint PDF $f(\boldsymbol{\vartheta})$ is $f(\boldsymbol{\vartheta}) = f(\mu^{-1}|\lambda)f(\lambda)$. Hence, the log-likelihood function of $\boldsymbol{\theta}$ for degradation model M_5^{IG} is given by $\mathscr{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln l_i(\boldsymbol{\theta})$, where

$$l_i(\boldsymbol{\theta}) = \int_{-\infty}^{y_{i,1}} \cdots \int_{-\infty}^{y_{i,m_i}} \frac{\Gamma(\alpha + m_i/2) a_i^{-(\alpha+m_i/2)} \prod_{j=1}^{m_i} \Lambda_{i,j}}{\Gamma(\alpha)(2\pi/\beta)^{m_i/2} \prod_{j=1}^{m_i}(y_{i,j} - \tilde{\varepsilon}_{i,j})^{3/2} \sqrt{1 + \sigma_\mu^2 \sum_{j=1}^{m_i}(y_{i,j} - \tilde{\varepsilon}_{i,j})}} f(\tilde{\boldsymbol{\varepsilon}}_i) \mathrm{d}\tilde{\boldsymbol{\varepsilon}}_i. \tag{8.10}$$

and

$$a_i = 1 + \frac{\beta(\xi^2 \sum_{j=1}^{m_i}(y_{i,j} - \tilde{\varepsilon}_{i,j}) - 2\xi \Lambda(t_{i,m_i}) - \Lambda(t_{i,m_i})^2 \sigma_\mu^2)}{2(\sigma_\mu^2 \sum_{j=1}^{m_i}(y_{i,j} - \tilde{\varepsilon}_{i,j}) + 1)} + \frac{\beta}{2} \sum_{j=1}^{m_i} \frac{\Lambda_{i,j}^2}{y_{i,j} - \tilde{\varepsilon}_{i,j}}.$$

8.4.2 *Quasi-Monte Carlo-Type Integration*

Generally speaking, there is no way to have a closed-form expression of maximum likelihood estimates (MLEs) for each unknown parameter because of the multiple integral in (8.8). The quasi-Monte Carlo integration method offers an alternative and simpler framework for computing the MLEs of the unknown parameters. Therefore, any numerical integration method can waste significant amounts of computational effort due to the infinite integration limits. A separation-of-variables method developed by Genz (1992) is used to transform the original integral (8.8) into an integral over a unit hypercube. For $j = 1, \ldots, m_i$, first define the following three transformations

$$\begin{aligned} \tilde{\varepsilon}_{i,j} &= \sigma_\varepsilon(z_{i,j} - z_{i,j-1}), \\ z_{i,j} &= \Phi^{-1}(x_{i,j}) \end{aligned}$$

and

$$x_{i,j} = \Phi(y_{i,j}/\sigma_\varepsilon + \Phi^{-1}(x_{i,j-1}))w_{i,j},$$

where $z_{i,0} = w_{i,0} = 0$ and $x_{i,0} = 0.5$. Applying the above transformations to $l_i(\boldsymbol{\theta})$ in (8.8), the likelihood function $l_i(\boldsymbol{\theta})$ can then be written as

$$l_i(\boldsymbol{\theta}) = \int_{\boldsymbol{\Theta}} \int_{[0,1]^{m_i}} g_i(\boldsymbol{w}_i; \boldsymbol{\vartheta}) f(\boldsymbol{\vartheta}) \mathrm{d}\boldsymbol{w}_i \mathrm{d}\boldsymbol{\vartheta}, \tag{8.11}$$

where

$$g_i(\boldsymbol{w}_i;\boldsymbol{\vartheta}) = \prod_{j=1}^{m_i} f_{L_{i,j}}(y_{i,j} - \sigma_\varepsilon z_{i,j}(\boldsymbol{w}_i) + \sigma_\varepsilon z_{i,j-1}(\boldsymbol{w}_i)|\boldsymbol{\vartheta})e_{i,j}(\boldsymbol{w}_i),$$

$\boldsymbol{w}_i = (w_{i,1},\ldots,w_{i,m_i})'$, $z_{i,0}(\boldsymbol{w}_i) = 0$, $e_{i,j}(\boldsymbol{w}_i) = \Phi(y_{i,j}/\sigma_\varepsilon + z_{i,j-1}(\boldsymbol{w}_i))$, and $z_{i,j}(\boldsymbol{w}_i) = \Phi^{-1}(e_{i,j}(\boldsymbol{w}_i)w_{i,j})$ for $j = 1,\ldots,m_i$. These transformations have the effect of flattening the surface of the original function and improving numerical tractability. In particular, the first transformation can avoid the computational complexity of the Cholesky decomposition, although the calculation in (8.8) can be transformed into an easier computational problem if the variables are reordered in the multiple integral. Note that when computing terms like $g_i(\boldsymbol{w}_i;\boldsymbol{\vartheta})$ for $f_{L_{i,j}}, e_{i,j}$ are small, the value $g_i(\boldsymbol{w}_i;\boldsymbol{\vartheta})$ may be outside the range of double precision arithmetic. Instead, one should use expressions like

$$g_i(\boldsymbol{w}_i;\boldsymbol{\vartheta}) = \exp\left(\sum_{j=1}^{m_i} \ln f_{L_{i,j}}(y_{i,j} - \sigma_\varepsilon z_{i,j}(\boldsymbol{w}_i) + \sigma_\varepsilon z_{i,j-1}(\boldsymbol{w}_i)|\boldsymbol{\vartheta}) + \ln e_{i,j}(\boldsymbol{w}_i)\right)$$

to avoid the overflow problem.

We next use a periodized, randomized quasi-Monte Carlo (QMC) rule (Richtmyer 1951) to approximate (8.11) in the following form:

$$l_i(\boldsymbol{\theta}) \approx \bar{l}_i(\boldsymbol{\theta}) = \frac{1}{N_1}\sum_{k_1=1}^{N_1} l_{i,N_2}^{(k_1)}(\boldsymbol{\theta}),$$

with

$$l_{i,N_2}^{(k_1)}(\boldsymbol{\theta}) = \frac{1}{2N_2}\sum_{k_2=1}^{N_2} g_i(|2\langle k_2\sqrt{\mathbf{p}} + \boldsymbol{w}_i^{(k_1)}\rangle - \mathbf{1}_i|;\boldsymbol{\vartheta}^{(k_1)}) + g_i(\mathbf{1}_i - |2\langle k_2\sqrt{\mathbf{p}} + \boldsymbol{w}_i^{(k_1)}\rangle - \mathbf{1}_i|;\boldsymbol{\vartheta}^{(k_1)}) \tag{8.12}$$

where $\mathbf{p} = (2,3,5,\ldots,p_{m_i})'$ in which p_k is the kth prime number; $\mathbf{1}_i$ is a column of 1's having length m_i; $\langle\cdot\rangle$ denotes the remainder mod 1; the sample points $\boldsymbol{w}_i^{(k_1)} = (w_{i,1}^{(k_1)},\ldots,w_{i,m_i}^{(k_1)})'$ and $\boldsymbol{\vartheta}^{(k_1)}$ are randomly generated from a uniform distribution between 0 and 1 and the random effects $\boldsymbol{\vartheta}$, respectively. Note that the antithetic variates method is used replacing $g_i(\boldsymbol{w}_i;\boldsymbol{\vartheta})$ by $(g_i(\boldsymbol{w}_i;\boldsymbol{\vartheta}) + g_i(\mathbf{1}_i - \boldsymbol{w}_i;\boldsymbol{\vartheta}))/2$ for variance reduction. The baker's transformation (i.e., $|2\boldsymbol{w}_i - \mathbf{1}_i|$) of periodization in (8.12) provides $O(N_2^{-2+\varepsilon})$ integration errors shown by Hickernell (2002) for randomized lattice rules, where N_2^ε stands for $\ln(N_2)$ to some power. The standard error of randomly shifted QMC integration for each degradation path can be calculated by

$$\sigma_i^{QMC}(\boldsymbol{\theta}) = \left(\frac{1}{N_1(N_1-1)} \sum_{k_1=1}^{N_1} (l_{i,N_2}^{(k_1)}(\boldsymbol{\theta}) - \bar{l}_i(\boldsymbol{\theta}))^2 \right)^{1/2}.$$

The relative error can be estimated by the coefficient of variation $\sigma_i^{QMC}(\boldsymbol{\theta})/\bar{l}_i(\boldsymbol{\theta})$. More details about lattice rules in general and the approach used here can be referred to Sloan and Joe (1994), Hickernell (1998), Genz and Bretz (2009), and the references therein.

Finally, the MLE, $\hat{\boldsymbol{\theta}}$, of all unknown parameters can be found numerically by maximizing the approximation of the log-likelihood function given above.

8.4.3 Bootstrap Confidence Intervals

The CDF estimate of the lifetime for the independent increment degradation-based process can be easily constructed by substituting the MLE ($\hat{\boldsymbol{\theta}}$) into the corresponding formula provided in Sect. 8.2. However, the CI estimation by the usual asymptotic normal likelihood methods is not easy to carry out because of the intractable Fisher's information matrix. Under this scenario, an attractive alternative is to use the bias-corrected percentile bootstrap method (see Efron and Tibshirani (1993); Meeker and Escobar (1998)). For illustrative purposes, we use the bias-corrected percentile bootstrap method to compute the $100(1-\alpha^*)\%$ CI for CDF of the degradation model M_5^{IG}. The bootstrap algorithm is implemented with the following steps.

(i) Use the observed data (i.e., n sample paths) and the previous estimation procedure to compute the MLEs $\hat{\boldsymbol{\theta}}_5^{IG}$ of the degradation model M_5^{IG}.

(ii) Given the threshold ω, substitute the estimates $\hat{\boldsymbol{\theta}}_5^{IG}$ into (8.6) giving $\hat{F}_{T_5^{IG}}(t) = F_{T_5^{IG}}(t; \hat{\boldsymbol{\theta}}_5^{IG})$.

(iii) Generate a large number of bootstrap samples B (e.g., $B = 2000$) that mimic the original sample and compute the corresponding bootstrap estimates $\hat{F}^*_{T_5^{IG}}(t)$ according to the following steps.

  (a) Generate, from $\hat{\boldsymbol{\theta}}_5^{IG}$, n simulated realizations of the random path parameters μ_i^*, λ_i^* and the measurement errors $\boldsymbol{\varepsilon}_i^*$ for $i = 1, \ldots, n$. i.e.,

$$\begin{aligned}
&\frac{1}{\mu_i^*}\Big|\lambda_i^* \sim \mathcal{N}(\hat{\xi}, \hat{\sigma}_\mu^2/\lambda_i^*),\\
&\lambda_i^* \sim \mathcal{G}(\hat{\alpha}, \hat{\beta}),\\
&\boldsymbol{\varepsilon}_i^* \sim \mathcal{N}_{m_i}(\mathbf{0}_i, \hat{\sigma}_\varepsilon^2 \boldsymbol{I}_i),
\end{aligned}$$

where $\delta_i^* = 1/\mu_i^*$ and λ_i^* have the joint PDF $f(\delta_i^*, \lambda_i^*) = f(\delta_i^*|\lambda_i^*) f(\delta_i^*)$ and $\mathbf{0}_i$ and $\boldsymbol{I}_i$ denote a column of 0's with length m_i and an identity matrix of order m_i, respectively.

(b) Using the property of independent increments, (8.7), and the same sampling scheme as in the original test, generate n simulated observed sample paths

$$Y_i^*(t_{i,j}|\mu_i^*, \lambda_i^*) = L_i^*(t_{i,j}|\mu_i^*, \lambda_i^*) + \varepsilon_{i,j}^*$$

from

$$L_i^*(t_{i,j}|\mu_i^*, \lambda_i^*) \sim \mathscr{IG}(\mu_i^* \hat{\Lambda}(t_{i,j}), \lambda_i^* \hat{\Lambda}(t_{i,j})^2)$$

up to the test stopping time t_{i,m_i}, where $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$.

(c) Use the n simulated sample paths and the previous estimation procedure to estimate parameters of the degradation model M_5^{IG}, giving the bootstrap estimates $\hat{\boldsymbol{\theta}}_5^{*IG}$.

(d) Given the threshold ω, substitute the estimates $\hat{\boldsymbol{\theta}}_5^{*IG}$ into (8.6) giving $\hat{F}^*_{T_5^{IG}}(t) (= F^*_{T_5^{IG}}(t; \hat{\boldsymbol{\theta}}_5^{*IG}))$ at desired values of time t.

(iv) For each desired value of t, the bootstrap CI for $F_{T_5^{IG}}(t)$ of the degradation model M_5^{IG} is computed using the following steps.

(a) Sort the B bootstrap estimates $\hat{F}^*_{T_5^{IG},1}(t), \ldots, \hat{F}^*_{T_5^{IG},B}(t)$ in increasing order giving $\hat{F}^*_{T_5^{IG},(b)}(t), b = 1, \ldots, B$.

(b) The lower and upper bounds of approximate $100(1-\alpha^*)\%$ CI for $F_{T_5^{IG}}(t)$ are

$$\left[\underline{F_{T_5^{IG}}(t)},\ \overline{F_{T_5^{IG}}(t)}\right] = \left[\hat{F}^*_{T_5^{IG},(l)}(t),\ \hat{F}^*_{T_5^{IG},(u)}(t)\right],$$

where

$$l = B \times \Phi(2\Phi^{-1}(p^*) + \Phi^{-1}(\alpha^*/2)),$$
$$u = B \times \Phi(2\Phi^{-1}(p^*) + \Phi^{-1}(1-\alpha^*/2)),$$

and p^* is the proportion of the B values of $\hat{F}^*_{T_5^{IG}}(t)$ that are less than $\hat{F}_{T_5^{IG}}(t)$. Note that l and u are chosen as the next lowest and next highest integers, respectively.

8.5 Degradation Model with Explanatory Variables

When explanatory variables such as accelerating variables (e.g., humidity, voltage, temperature) are used in accelerated degradation tests (ADTs), the independent

increment degradation-based processes incorporated with explanatory variables can be used to make inferences of the product's lifetime information more accurate. The time-independent explanatory variables are only considered in this study. Let $x_z = \exp(\boldsymbol{X}'\boldsymbol{Z})$ stand for the explanatory variables for simplicity, where $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively, denote the vectors of explanatory variables and regression coefficients. Without loss of generality, $\boldsymbol{X}'\boldsymbol{Z}$ without any constant term is used to avoid the non-identifiable problem. An extension of the independent increment degradation-based processes without explanatory variables in (8.2)–(8.4) to those with explanatory variables is considered as follows:

$$L(t|\Theta) = \Theta x_z \Lambda(t) + \sigma_W W(t), \quad \Theta \sim \mathcal{N}(\eta, \sigma_\eta^2), \tag{8.13}$$

$$L(t|\tilde{\beta}) \sim \mathcal{G}(\tilde{\alpha} x_z \Lambda(t), \tilde{\beta}), \quad \tilde{\beta}^{-1} \sim \mathcal{G}(\tilde{r}, \tilde{s}) \tag{8.14}$$

and

$$L(t|\mu, \lambda) \sim \mathcal{IG}(\mu x_z \Lambda(t), \lambda \Lambda(t)^2), \quad \delta \equiv \mu^{-1}|\lambda \sim \mathcal{N}(\xi, \sigma_\mu^2/\lambda), \quad \lambda \sim \mathcal{G}(\alpha, \beta). \tag{8.15}$$

The parameters Θ in (8.13), $\tilde{\alpha}$ in (8.14), and μ in (8.15) are assumed to be dependent on explanatory variables, and the other parameters are considered to be independent of explanatory variables. This means that the explanatory variables can impact the mean of degradation paths at different accelerated conditions. Clearly, setting $x_z = 1$ means that there are no explanatory variables involved in the ADT model.

Example 5 The product's lifetime distribution and the log-likelihood function for the degradation model M_5^{IG} with explanatory variables can be easily obtained by both replacing ξ with ξ/x_z and σ_μ with σ_μ/x_z in (8.6) and (8.10).

8.6 Case Applications

The independent increment degradation-based processes, M_1^W–M_6^W, M_1^G–M_4^G and M_1^{IG}–M_8^{IG}, are used to fit the degradation data. The Akaike information criterion (i.e., AIC $= -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2r$) is adopted for the degradation model selection, where $\mathcal{L}(\hat{\boldsymbol{\theta}})$ is the sample log-likelihood of the corresponding degradation model and r represents the number of unknown parameters in $\boldsymbol{\theta}$. Smaller AIC values indicate that the degradation model fits the data better. For the models M_3^G, M_4^G, and M_5^{IG}–M_8^{IG}, the numbers N_1 and N_2 are indicated in the table heading of parameter estimation to allow the maximization of the log-likelihood function to converge stably. The likelihood-ratio (LR) test is performed to assess whether the measurement error term is necessary for a specific dataset. Furthermore, the product's lifetime information (e.g., $MTTF$ or q-quantile) and the corresponding CIs based on the selected degradation model can be obtained by using the procedures in Sect. 3, where the number of bootstrap samples is 2000. The Anderson–Darling test is used to diagnose whether the selected degradation model is suitable to fit the degradation data.

8.6.1 LED Data Revisited

The motivating example demonstrates the advantages of independent increment degradation-based processes. For illustrative purposes, the widely used case $\Lambda(t) = t^{\gamma}$ is used for modeling the degradation paths. Table 8.2 gives the results for MLEs, sample log-likelihoods, and AICs of the LED (complete) data. In Table 8.2, some AIC values are different to the others because the corresponding models with large AIC values are unsuitable for fitting the LED (complete) data. For the Wiener degradation-based process, models M_4^W and M_6^W without measurement errors are substantially better than models M_1^W–M_3^W and M_5^W with measurement errors by using the AIC for model selection. The monotonic processes with measurement errors (i.e., M_3^G, M_4^G, M_5^{IG}–M_8^{IG}) are substantially better than the Wiener (non-monotonic) degradation-based processes (i.e., M_1^W–M_6^W) by using the AIC for model selection. The statistical analysis is in agreement with the material theory and empirical experiments to determine the monotonic degradation path for the LED. The IG degradation-based process, M_8^{IG}, is also suitable, and the sources of the variation are both the measurement errors and the within-unit variability.

An alternative approach, as mentioned in the introduction, excludes an abnormal measurement point in the LED data as incomplete data. The results for the MLEs, the sample log-likelihoods, and the AICs for the LED (incomplete) data are listed in Table 8.3. Again, models M_4^W and M_6^W without measurement errors are substantially better than models M_1^W–M_3^W and M_5^W with measurement errors, whether the anomaly is excluded or not. Therefore, models M_3^G and M_5^{IG}–M_8^{IG} with measurement errors are substantially better than models M_1^G and M_1^{IG}–M_4^{IG} without measurement errors by using the AIC for model selection. Comparing model M_3^G (M_4^G) with M_1^G (M_2^G) using the LR test, there is no sufficient evidence to reject the hypothesis $H_0 : \sigma_\varepsilon = 0$ at significance level 0.05. This reveals that the measurement error term in the model M_3^G (M_4^G) is insignificant. Comparing model M_5^{IG} (M_6^{IG}–M_8^{IG}) with M_1^{IG} (M_2^{IG}–M_4^{IG}) using the LR test, there is sufficient evidence to reject the hypothesis $H_0 : \sigma_\varepsilon = 0$ at significance level 0.05, and it shows that the measurement error term in model M_5^{IG} (M_6^{IG}–M_8^{IG}) is necessary for the LED (incomplete) data. The IG degradation-based process, M_8^{IG}, is a suitable model for modeling the LED (incomplete) data. This means that there are still measurement errors in the LED (incomplete) data, even though the observed degradation paths are monotonic.

A comparison of the Wiener degradation-based processes in Tables 8.2 and 8.3 shows that removing the anomaly does not increase the corresponding log-likelihood. This means that substantial data information can be lost when the anomaly is removed. According to the AIC for model selection, the negative increment in the second degradation path may be normal. The nonlinear Wiener process (with fixed effects) without measurement errors, M_6^W in Table 8.2, could fit the LED (complete) data. When AIC is used for gamma/IG degradation-based processes (M_3^G, M_4^G, M_6^{IG}, and M_8^{IG}), removing the anomaly increases the corresponding log-likelihood and reduces the estimated measurement error $\hat{\sigma}_\varepsilon^2$. This means that the anomaly can be excluded from the degradation data. The nonlinear gamma process (with fixed

Table 8.2 MLEs, sample log-likelihoods, and AICs for the LED (complete) data ($N_1 = 47$, $N_2 = 5 \times 10^6$)

Wiener degradation-based process ($\Lambda(t) = t^\gamma$)										
Model	$\hat{\eta} \times 10^2$	$\hat{\sigma}_\eta \times 10^3$	$\hat{\sigma}_W \times 10^2$	$\hat{\sigma}_\varepsilon \times 10^1$	$\hat{\gamma} \times 10^1$				$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_1^W	2.41	4.61	2.09	1.02	7.87				6.24	−2.48
M_2^W	2.37	—	2.84	0.54	7.90				5.69	−3.37
M_3^W	2.58	5.95	—	1.60	7.76				5.40	−2.80
M_4^W	2.27	3.36	2.91	—	7.99				5.82	−3.64
M_5^W	2.74	—	—	3.43	7.64				−15.74	37.48
M_6^W	2.32	—	3.02	—	7.94				5.63	−5.26
Gamma degradation-based process ($\Lambda(t) = t^\gamma$)										
Model	$\hat{\tilde{\alpha}} \times 10^1$	$\hat{\tilde{\beta}} \times 10^1$	$\hat{\tilde{r}} \times 10^{-1}$	$\hat{\tilde{s}} \times 10^1$	$\hat{\sigma}_\varepsilon \times 10^2$	$\hat{\gamma} \times 10^1$			$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_3^G	3.97	—	2.23	6.16	7.32	7.47			8.63	−7.27
M_4^G	3.00	1.01	—	—	6.93	7.46			7.80	−7.61
Inverse Gaussian degradation-based process ($\Lambda(t) = t^\gamma$)										
Model	$\hat{\mu} \times 10^2$	$\hat{\xi} \times 10^{-1}$	$\hat{\sigma}_\mu \times 10^1$	$\hat{\lambda} \times 10^2$	$\hat{\alpha} \times 10^{-1}$	$\hat{\beta} \times 10^4$	$\hat{\sigma}_\varepsilon \times 10^2$	$\hat{\gamma} \times 10^1$	$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_5^{IG}	—	3.12	7.82	—	27.4	0.54	7.73	7.29	9.00	−6.00
M_6^{IG}	3.35	—	—	—	5.33	2.03	7.36	7.29	8.23	−6.46
M_7^{IG}	—	3.15	7.77	1.45	—	—	7.72	7.31	9.00	−8.00
M_8^{IG}	**3.34**	—	—	**1.06**	—	—	**7.35**	**7.29**	**8.23**	**−8.46**

Table 8.3 MLEs, sample log-likelihoods, and AICs for the LED (incomplete) data ($N_1 = 47$, $N_2 = 5 \times 10^6$)

Wiener degradation-based process ($\Lambda(t) = t^\gamma$)										
Model	$\hat{\eta} \times 10^2$	$\hat{\sigma}_\eta \times 10^3$	$\hat{\sigma}_W \times 10^2$	$\hat{\sigma}_\varepsilon \times 10^1$	$\hat{\gamma} \times 10^1$				$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_1^W	2.37	4.25	2.35	0.85	7.90				5.71	−1.42
M_2^W	2.34	—	3.00	0.09	7.93				5.36	−2.73
M_3^W	2.57	5.93	—	1.61	7.76				4.78	−1.55
M_4^W	2.28	3.43	2.89	—	7.98				5.57	−3.14
M_5^W	2.73	—	—	3.47	7.65				−15.84	37.69
M_6^W	2.34	—	3.01	—	7.93				5.36	−4.73
Gamma degradation-based process ($\Lambda(t) = t^\gamma$)										
Model	$\hat{\tilde{\alpha}} \times 10^1$	$\hat{\tilde{\beta}} \times 10^1$	$\hat{\tilde{r}} \times 10^{-1}$	$\hat{\tilde{s}} \times 10^1$	$\hat{\sigma}_\varepsilon \times 10^2$	$\hat{\gamma} \times 10^1$			$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_1^G	2.76	—	6.06	1.23	—	7.07			7.41	−6.83
M_2^G	2.62	1.42	—	—	—	7.08			7.32	−8.64
M_3^G	3.60	—	2.12	5.62	5.38	7.39			8.72	−7.45
M_4^G	2.88	1.11	—	—	4.64	7.36			8.18	−8.36
Inverse Gaussian degradation-based process ($\Lambda(t) = t^\gamma$)										
Model	$\hat{\mu} \times 10^2$	$\hat{\xi} \times 10^{-1}$	$\hat{\sigma}_\mu \times 10^1$	$\hat{\lambda} \times 10^2$	$\hat{\alpha}$	$\hat{\beta} \times 10^2$	$\hat{\sigma}_\varepsilon \times 10^2$	$\hat{\gamma} \times 10^1$	$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_1^{IG}	—	2.70	4.98	—	1.30	1.01	—	7.09	5.28	−0.56
M_2^{IG}	3.92	—	—	—	1.43	0.86	—	7.03	5.06	−2.11
M_3^{IG}	—	1.54	1.75×10^{-5}	1.71	—	—	—	6.08	0.21	7.58
M_4^{IG}	6.48	—	—	1.71	—	—	—	6.08	0.21	5.58
M_5^{IG}	—	3.03	6.79	—	22.45	6.05×10^{-3}	6.41	7.25	8.96	−5.93
M_6^{IG}	3.49	—	—	—	12.99	8.79×10^{-2}	5.88	7.22	8.42	−6.84
M_7^{IG}	—	3.05	6.82	1.34	—	—	6.47	7.26	8.96	−7.93
M_8^{IG}	**3.49**	—	—	**1.08**	—	—	**6.00**	**7.22**	**8.41**	**−8.81**

Table 8.4 MLEs, 95% CIs of $MTTF_8^{IG}$ and $t_8^{IG}(q)$, where $q = 0.05, 0.1, 0.5, 0.9$, and 0.95 for the LED (incomplete) data ($N_1 = 47$, $N_2 = 10^5$)

	Degradation model (M_8^{IG})	
Lifetime	MLE	95% bootstrap CI
$MTTF_8^{IG}$	240.24	[187.89, 320.44]
$t_8^{IG}(0.05)$	118.47	[68.18, 200.54]
$t_8^{IG}(0.10)$	142.26	[89.34, 214.43]
$t_8^{IG}(0.50)$	236.20	[183.94, 318.37]
$t_8^{IG}(0.90)$	343.30	[273.18, 415.64]
$t_8^{IG}(0.95)$	375.74	[292.28, 454.45]

effects) with measurement errors, M_4^G in Table 8.2, and the nonlinear gamma process (with fixed effects) without measurement errors, M_2^G in Table 8.3, are also, respectively, suitable for complete and incomplete LED data. For IG degradation-based processes, the nonlinear IG process (with fixed effects) with measurement errors, M_8^{IG}, is consistently chosen by using AIC in Tables 8.2 and 8.3. Overall, the IG degradation-based process, M_8^{IG} in Table 8.3, is substantially better than the others in terms of the AIC for the LED (complete or incomplete) data.

Using the estimated $\hat{\boldsymbol{\theta}}_8^{IG}$ (under the degradation model M_8^{IG} with $\Lambda(t) = t^\gamma$), and given the pre-fixed critical level (i.e., $\omega = 1.75$), Table 8.4 lists values for $\widehat{MTTF}_8^{IG}$, $\hat{t}_8^{IG}(q)$ (where $q = 0.05, 0.1, 0.5, 0.9$, and 0.95), and the corresponding 95% bootstrap CIs. Figure 8.2 shows its CDF estimate (solid line) and the corresponding pointwise 95% bootstrap CIs (dashed lines) for model M_8^{IG} with $\Lambda(t) = t^\gamma$. The pseudo-failure-time (PFT) estimation is also estimated by $\Lambda^{-1}(\omega/\tilde{\mu}_i)$ for $i = 1, \ldots, n$, where $\tilde{\mu}_i$ is the least squares estimate by fitting the ith degradation path. The PFTs are the times when the fitted curves reach the critical level ω, and they are shown with black dots in Fig. 8.2. For goodness of fit, the p-value of the Anderson–Darling test

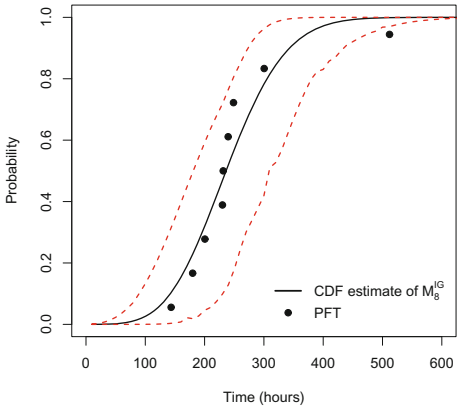


Fig. 8.2 Estimated time-to-failure distribution of the LED (incomplete) data, using IG degradation-based process M_8^{IG} with $\omega = 1.75$

for the model M_8^{IG} with $\Lambda(t) = t^\gamma$ is 0.58. This indicates that the ING degradation model, M_8^{IG}, is suitable for fitting the LED (incomplete) data.

8.6.2 Laser Data

Since the laser data in Peng (2015a, Example 1) was analyzed and fitted using the degradation models M_1^W–M_6^W, M_1^G, M_2^G, and M_1^{IG}–M_4^{IG} with $\Lambda(t) = t$ and $x_z = 1$, this study uses the proposed models, M_3^G, M_4^G, and M_5^{IG}–M_8^{IG} with $\Lambda(t) = t$ and $x_z = 1$, to determine whether there are measurement errors in the degradation paths.

The results for the MLEs, the sample log-likelihoods, and the AICs for the laser data are summarized in Table 8.5. A comparison of Table 3 in Peng (2015a) and Table 8.5 shows that the monotonic processes (i.e., M_1^G–M_4^G, M_1^{IG}–M_8^{IG}) are substantially better than the Wiener degradation-based processes by using the AIC for model selection. By using the LR test, the estimated measurement errors $\hat{\sigma}_\varepsilon$ are negligible for the monotonic processes, which means that there are no measurement errors in the laser data. The IG degradation-based process, M_3^{IG}, is suitable, and there is parametric variation in neither λ nor the measurement error ε in the laser data. The remaining reliability assessment and goodness of fit can be found in Peng (2015a) and are omitted here.

8.6.3 Carbon-Film Data

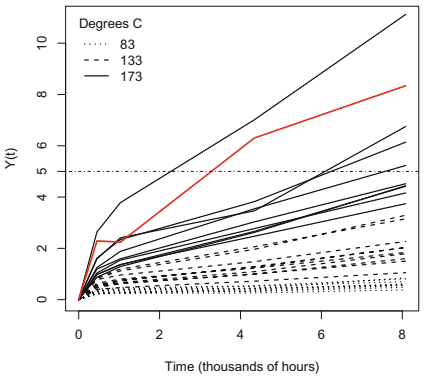
The following carbon-film data is taken from Meeker and Escobar (1998, Example 18.2, Table C.3). See Shiomi and Yanagisawa (1979), Suzuki et al. (1993) for more details. The QC of a resistor is its resistance. The single accelerating variable is temperature, and the Arrhenius reaction law is used as follows:

$$x_z = \exp\left(-\frac{11605 \times E_a}{273.15 + \text{temp}}\right),$$

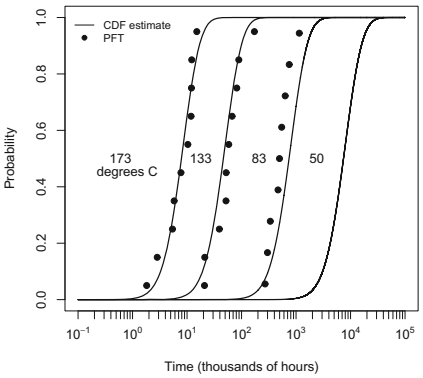
where temp and E_a denote temperature in degrees Celsius and an unknown activation energy, respectively. Figure 8.3a shows the accelerated degradation paths for carbon-film resistors. Table 8.6 records the increments for the accelerated degradation data for each degradation path. The subject devices were tested at three levels of temperature (i.e., 83°C, 133°C, and 173°C). The respective sample sizes of the tested samples at 83°C, 133°C, and 173°C are 9, 10, and 10. The measurement times at the three levels of temperature are the same at 0.452, 1.030, 4.341, and 8.084 (in thousands of hours). The primary objective of these ADTs is to assess the lifetime information for carbon-film resistors, such as the $MTTF$ or the q-quantile at the normally used operating temperature (e.g., 50°C). For an individual resistor, the time-to-failure is assumed to be the time when the true resistance is 5 (i.e., $\omega = 5$) more than the initial

Table 8.5 MLEs, sample log-likelihoods, and AICs for the laser data ($N_1 = 20$, $N_2 = 10^7$)

Gamma degradation-based process ($\Lambda(t) = t$)									
Model	$\hat{\tilde{\alpha}} \times 10^2$	$\hat{\tilde{\beta}} \times 10^2$	$\hat{\tilde{r}} \times 10^{-1}$	$\hat{\tilde{s}} \times 10^1$	$\hat{\sigma}_\varepsilon \times 10^6$			$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_3^G	3.90	—	2.89	6.87	1.87			93.74	−179.48
M_4^G	2.88	7.09	—	—	1.15			69.61	−133.22
Inverse Gaussian degradation-based process ($\Lambda(t) = t$)									
Model	$\hat{\mu} \times 10^3$	$\hat{\xi} \times 10^{-2}$	$\hat{\sigma}_\mu$	$\hat{\lambda} \times 10^5$	$\hat{\alpha} \times 10^{-1}$	$\hat{\beta} \times 10^6$	$\hat{\sigma}_\varepsilon \times 10^6$	$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_5^{IG}	—	5.04	0.79	—	2.34	3.21	1.79	95.61	−181.22
M_6^{IG}	2.04	—	—	—	4.73	1.18	1.18	75.10	−142.21
M_7^{IG}	—	5.09	0.76	7.19	—	—	1.46	95.33	−182.65
M_8^{IG}	2.04	—	—	5.45	—	—	1.31	75.03	−144.07



a. Accelerated degradation paths



b. CDF estimates based on the carbon-film (incomplete) data, using the degradation model M_1^G with $\omega = 5$. The PFTs are indicated by dots.

Fig. 8.3 Carbon-film data

resistance. Similar to the LED example, there is only one negative increment for the 26th unit (red solid line in Fig. 8.3a) in the 173°C test, which is shown in bold-faced type in Table 8.6. The abnormal measurement point may be an outlier and has more influence on the degradation model than others.

For illustrative purposes, the time-scale transformation function $\Lambda(t) = t^\gamma$ is considered in the log-likelihood functions and model selections in this case. Table 8.7 summarizes the results for the MLEs, the sample log-likelihoods, and the AICs for the carbon-film (complete) data. For the Wiener degradation-based processes, models M_1^W, M_3^W, and M_4^W with the random effects are substantially better than the other models by using the AIC for model selection. The proposed monotonic processes (i.e., M_3^G, M_4^G and M_5^{IG}–M_8^{IG}) in this paper have a smaller AIC than the non-monotonic processes (i.e., M_1^W–M_6^W). A comparison of the AICs shows that the degradation model, M_3^G, with $\Lambda(t) = t^\gamma$ is suitable, and $\tilde{\beta}$ appears to vary in the experiment with measurement errors. The degradation model, M_3^G, is clearly better than the other degradation models (i.e., M_5^{IG}–M_8^{IG} and M_1^W–M_6^W).

When the anomaly in the carbon-film (complete) data is excluded, the results for the MLEs, the sample log-likelihoods, and the AICs for the carbon-film (incomplete) data are shown in Table 8.8. Again, models M_1^W, M_3^W, and M_4^W with the random effects are substantially better than the other models, whether the anomaly is excluded or not. Models M_5^{IG}–M_8^{IG} with measurement errors are substantially better than models M_1^{IG}–M_4^{IG} without measurement errors by using the AIC for model selection. Comparing model M_5^{IG} (M_6^{IG}–M_8^{IG}) with M_1^{IG} (M_2^{IG}–M_4^{IG}) using the LR test, there is sufficient evidence to reject the hypothesis $H_0 : \sigma_\varepsilon = 0$ at significance level 0.05, and it reveals that the measurement error term in model M_5^{IG} (M_6^{IG}–M_8^{IG}) is necessary for the carbon-film (incomplete) data. This means that there are still measurement errors in the carbon-film (incomplete) data, even though the observed

Table 8.6 Increments of accelerated degradation data of carbon-film resistors

Unit Number	Temperature (°C)	Inspection time interval (hours)			
		0–452	452–1030	1030–4341	4341–8084
1	8	0.28	0.04	0.06	0.24
2		0.22	0.02	0.02	0.12
3		0.41	0.05	0.08	0.27
4		0.25	0.04	0.03	0.16
5		0.25	0.01	0.16	0.15
6		0.32	0.04	0.09	0.13
7		0.36	0.05	0.11	0.18
8		0.24	0.04	0.06	0.21
9		0.33	0.07	0.04	0.41
10	133	0.40	0.07	0.25	0.33
11		0.88	0.31	0.87	1.09
12		0.53	0.11	0.35	0.61
13		0.47	0.15	0.38	0.50
14		0.57	0.18	0.51	0.77
15		0.55	0.12	0.42	0.70
16		0.78	0.18	0.52	0.79
17		0.83	0.29	0.84	1.33
18		0.64	0.16	0.43	0.61
19		0.55	0.19	0.55	0.74
20	173	0.87	0.42	1.33	1.82
21		1.25	0.63	1.66	1.69
22		2.64	1.14	3.23	4.11
23		0.98	0.38	1.30	1.76
24		1.62	0.72	1.48	2.32
25		1.59	0.82	1.05	3.29
26		2.29	**−0.05**	4.06	2.04
27		0.98	0.39	1.10	1.27
28		1.04	0.50	1.23	1.39
29		1.19	0.40	1.44	1.49

degradation paths are monotonic for the IG degradation-based processes. Therefore, for the gamma degradation-based processes, the difference in the log-likelihood (or using the LR test) between models M_1^G (M_2^G) and M_3^G (M_4^G) is insignificant, which implies that there is no measurement error in the carbon-film (incomplete) data. The gamma degradation-based process M_1^G is a suitable model for fitting the carbon-film (incomplete) data. The removal of the abnormal measurement point seems to screen out the measurement errors. This result is similar to the influential points in regression analysis, in which the deletion of observations significantly affects the fit of the regression model and the subsequent conclusions.

Table 8.7 MLEs, sample log-likelihoods, and AICs for the carbon-film (complete) data ($N_1 = 47$, $N_2 = 2 \times 10^7$)

Wiener degradation-based process ($\Lambda(t) = t^{\gamma}$)											
Model	$\hat{\eta} \times 10^{-4}$	$\hat{\sigma}_{\eta} \times 10^{-3}$	$\hat{\sigma}_W \times 10^{1}$	$\hat{\sigma}_{\varepsilon} \times 10^{1}$	$\hat{\gamma} \times 10^{1}$	$\hat{Ea} \times 10^{1}$				$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_1^W	1.37	4.19	2.07	1.19	4.86	3.39				−40.59	93.19
M_2^W	1.34	—	3.88	1.19×10^{-4}	4.68	3.37				−71.87	153.74
M_3^W	1.71	5.48	—	2.42	5.15	3.50				−39.35	88.69
M_4^W	1.10	3.34	2.59	—	4.68	3.30				−42.03	94.07
M_5^W	4.27	—	—	8.82	5.13	3.85				−150.09	308.19
M_6^W	1.34	—	3.88	—	4.68	3.37				−71.87	151.74
Gamma degradation-based process ($\Lambda(t) = t^{\gamma}$)											
Model	$\hat{\tilde{\alpha}} \times 10^{-4}$	$\hat{\tilde{\beta}} \times 10^{1}$	$\hat{\tilde{r}} \times 10^{-1}$	$\hat{\tilde{s}} \times 10^{1}$	$\hat{\sigma}_{\varepsilon} \times 10^{2}$	$\hat{\gamma} \times 10^{1}$	$\hat{Ea} \times 10^{1}$			$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_3^G	5.45	—	1.66	4.47	3.35	4.71	3.19			−8.29	28.58
M_4^G	4.05	2.11	—	—	2.98	4.70	3.20			−16.39	42.78
Inverse Gaussian degradation-based process ($\Lambda(t) = t^{\gamma}$)											
Model	$\hat{\mu} \times 10^{-4}$	$\hat{\xi} \times 10^{4}$	$\hat{\sigma}_{\mu}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_{\varepsilon} \times 10^{2}$	$\hat{\gamma} \times 10^{1}$	$\hat{Ea} \times 10^{1}$	$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_5^{IG}	—	1.87	1.06	—	10.02	1.88	9.98	4.32	3.06	−10.16	34.33
M_6^{IG}	1.33	—	—	—	4.69	2.03	7.51	3.94	3.33	−12.93	37.87
M_7^{IG}	—	2.32	1.14	18.34	—	—	10.07	4.30	2.99	−10.27	32.53
M_8^{IG}	1.22	—	—	8.30	—	—	7.55	3.85	3.29	−13.42	36.84

Table 8.8 MLEs, sample log-likelihoods, and AICs for the carbon-film (incomplete) data ($N_1 = 47$, $N_2 = 2 \times 10^7$)

Wiener degradation-based process ($\Lambda(t) = t^\gamma$)											
Model	$\hat{\eta} \times 10^{-4}$	$\hat{\sigma}_\eta \times 10^{-3}$	$\hat{\sigma}_W \times 10^{1}$	$\hat{\sigma}_\varepsilon \times 10^{1}$	$\hat{\gamma} \times 10^{1}$	$\hat{Ea} \times 10^{1}$				$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_1^W	0.91	2.86	2.09	1.56×10^{-4}	4.68	3.23				−22.43	56.86
M_2^W	1.44	—	3.70	3.78×10^{-4}	4.68	3.40				−66.32	142.64
M_3^W	1.73	5.56	—	2.31	5.04	3.50				−35.51	81.03
M_4^W	0.91	2.86	2.09	—	4.68	3.23				−22.43	54.86
M_5^W	4.21	—	—	8.86	5.16	3.84				−149.25	306.49
M_6^W	1.44	—	3.70	—	4.68	3.40				−66.32	140.64
Gamma degradation-based process ($\Lambda(t) = t^\gamma$)											
Model	$\hat{\tilde{\alpha}} \times 10^{-4}$	$\hat{\tilde{\beta}} \times 10^{1}$	$\hat{\tilde{r}} \times 10^{-1}$	$\hat{\tilde{s}} \times 10^{1}$	$\hat{\sigma}_\varepsilon \times 10^{2}$	$\hat{\gamma} \times 10^{1}$	$\hat{Ea} \times 10^{1}$			$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_1^G	7.72	—	1.53	6.03	—	4.67	3.23			5.27	−0.54
M_2^G	5.04	1.75	—	—	—	4.65	3.21			−6.73	21.46
M_3^G	7.89	—	1.54	5.98	0.48	4.67	3.24			5.38	1.24
M_4^G	5.24	1.80	—	—	1.29	4.66	3.24			−6.31	22.63
Inverse Gaussian degradation-based process ($\Lambda(t) = t^\gamma$)											
Model	$\hat{\mu} \times 10^{-4}$	$\hat{\xi} \times 10^{5}$	$\hat{\sigma}_\mu$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_\varepsilon \times 10^{2}$	$\hat{\gamma} \times 10^{1}$	$\hat{Ea} \times 10^{1}$	$\mathscr{L}(\hat{\boldsymbol{\theta}})$	AIC
M_1^{IG}	—	9.16	5.25×10^{-9}	—	0.73	10.97	—	4.10	3.27	−23.79	59.57
M_2^{IG}	1.09	—	—	—	0.73	10.97	—	4.10	3.27	−23.79	57.57
M_3^{IG}	—	6.76	7.23×10^{-9}	2.95	—	—	—	2.64	3.27	−53.91	117.82
M_4^{IG}	1.48	—	—	2.95	—	—	—	2.64	3.27	−53.91	115.82
M_5^{IG}	—	19.35	1.29	—	10.75	2.19	9.99	4.35	3.05	−6.31	26.62
M_6^{IG}	1.34	—	—	—	4.63	2.15	7.22	3.92	3.33	−10.34	32.68
M_7^{IG}	—	22.95	1.36	22.27	—	—	9.98	4.33	3.00	−6.42	24.84
M_8^{IG}	1.22	—	—	8.71	—	—	7.26	3.84	3.29	−10.92	31.83

Table 8.9 MLEs, 95% bootstrap CIs of $MTTF_1^G$, and $t_1^G(q)$ where $q = 0.05$, 0.1, 0.5, 0.9, and 0.95 for the carbon-film (incomplete) data

	Degradation model (M_1^G) with $\omega = 5$ at 50 °C	
Lifetime	MLE	95% bootstrap CIs
$MTTF_1^G$	8661.3	[3320.4, 23995.2]
$t_1^G(0.05)$	2322.9	[733.0, 6210.7]
$t_1^G(0.10)$	3074.8	[1075.2, 8202.3]
$t_1^G(0.50)$	7427.2	[2784.3, 20063.1]
$t_1^G(0.90)$	15785.3	[6064.7, 46220.0]
$t_1^G(0.95)$	19192.1	[7223.3, 58309.1]

A comparison of the Wiener degradation-based processes in Tables 8.7 and 8.8 shows that removing the anomaly increases the corresponding log-likelihood and produces more information about the data. According to the AIC for model selection, the negative increment in the 173°C test may be an outlier and should be excluded from the degradation path. The nonlinear Wiener process without measurement errors, M_4^W in Table 8.8, could fit the carbon-film (incomplete) data. When AIC is used for gamma/IG degradation-based processes (M_3^G–M_4^G/M_5^{IG}–M_8^{IG}), removing the anomaly increases the corresponding log-likelihood and reduces the estimated measurement error, $\hat{\sigma}_\varepsilon^2$. This means that the anomaly should be excluded. The nonlinear gamma process (with random effects) with measurement errors, M_3^G in Table 8.7, and the nonlinear gamma process (with random effects) without measurement errors, M_1^G in Table 8.8, could also be, respectively, used for the complete and incomplete ADT data. For IG degradation-based processes, the nonlinear IG process (with random effects) with measurement errors, M_7^{IG}, is consistently chosen using the AIC in Tables 8.7 and 8.8. Overall, the gamma degradation-based process, M_1^{IG} in Table 8.8, is substantially better than the other models, in terms of the AIC for model selection for the carbon-film (incomplete) data.

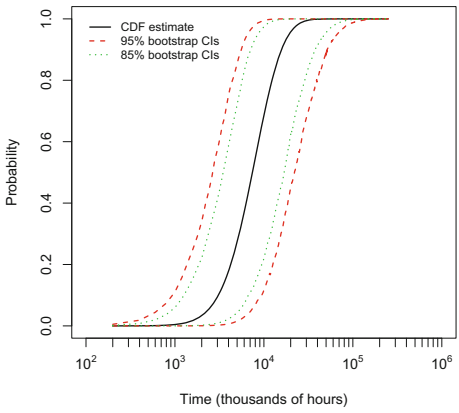


Fig. 8.4 CDF estimate of the time-to-failure distribution at 50°C with pointwise 85% and 95% bootstrap CIs, based on the carbon-film (incomplete) data, using the degradation model M_1^G with $\omega = 5$

Using the estimated $\hat{\boldsymbol{\theta}}_1^G$ for model M_1^G with $\Lambda(t) = t^\gamma$, and given the specified threshold level (i.e., $\omega = 5$) at 50°C, Table 8.9 summarizes values for $\widehat{MTTF}_1^G$, $\hat{t}_1^G(q)$ (where $q = 0.05, 0.1, 0.5, 0.9$, and 0.95), and the corresponding 95% bootstrap CIs. The wide CIs may be due to the large amount of extrapolation that is required to estimate the lifetime information for $\omega = 5$ at 50°C. Figure 8.3b plots the CDF estimates (solid lines) and the PFTs (black dots) for model M_1^G with $\Lambda(t) = t^\gamma$ at three levels of temperature. Figure 8.4 shows the CDF estimate and the corresponding pointwise 85 and 95% bootstrap bias-corrected percentile CIs at 50°C. For goodness of fit, the p-values of the Anderson–Darling test for the 83°C, 133°C, and 173°C tests are 0.08, 0.57, and 0.81, respectively. This demonstrates that the proposed ADT model is suitable for modeling the carbon-film (incomplete) data.

8.7 Concluding Remarks

Unusual observations can reflect an incorrectly specified model, in which case the observations may be rectified or deleted entirely. The independent increment degradation-based process that is proposed in this paper arises naturally when degradation paths that simultaneously consider the unit-to-unit variability, the within-unit variability, and the measurement error are necessary. The proposed degradation models M_3^G, M_5^{IG}–M_8^{IG} are new and share the similar properties of the corresponding models without measurement error (i.e., M_1^G, M_1^{IG}–M_4^{IG}). The assumption of the natural conjugate distribution makes this process eminently suitable for degradation modeling and allows the proposed degradation model to be computationally tractable. A separation-of-variables transformation with a quasi-Monte Carlo method is provided to estimate the model parameters and to develop procedures based on a bootstrap method to obtain CIs for reliability assessment. The LR test is performed to assess whether the measurement error term is necessary for a specific dataset. In addition, the AIC for model selection and the Anderson–Darling test for goodness of fit are provided to evaluate the validity of different model assumptions.

Although the estimation procedure and numerical results are obtained based on the independent increment degradation-based process, these assumptions are needed to verify before using the proposed models. An extension of this work to a more general model setting (such as Bayesian approach) will be studied in future research.

Acknowledgements This work was supported by the Ministry of Science and Technology (Grant No: MOST-104-2118-M-001-007) of Taiwan, Republic of China. The authors would like to thank Ms. Ya-Shan Cheng for her assistance in the computations.

References

Bagdonavičius, V., & Nikulin, M. S. (2000). Estimation in degradation models with explanatory variables. *Lifetime Data Analysis*, *7*, 85–103.

Barndorff-Nielsen, O. E., Mikosch, T., & Resnick, S. I. (2001). *Lévy Processes: Theory and Applications*. Boston: Birkhäuse.

Cheng, Y. S., & Peng, C. Y. (2012). Integrated degradation models in R using iDEMO. *Journal of Statistical Software*, *49*, 1–22.

Chuang, S. L., Ishibashi, A., Kijima, S., Nakayama, N., Ukita, M., & Taniguchi, S. (1997). Kinetic model for degradation of light-emitting diodes. *IEEE Journal of Quantum Electronics*, *33*, 970–979.

Di Nardo, E., Nobile, A. G., Pirozzi, E., & Ricciardi, L. M. (2001). A computational approach to first-passage-time problems for Gauss-Markov processes. *Advances in Applied Probability*, *33*, 453–482.

Doksum, K. A., & Hóyland, A. (1992). Model for variable-stress accelerated life testing experiments based on Wiener processes and the inverse Gaussian distribution. *Technometrics*, *34*, 74–82.

Doksum, K. A., & Normand, S. L. T. (1995). Gaussian models for degradation processes-part I: methods for the analysis of biomarker data. *Lifetime Data Analysis*, *1*, 131–144.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC.

Fukuda, M. (1991). *Reliability and Degradation of Semiconductor Lasers and LEDs*. Boston: Artech House.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *1*, 141–150.

Genz, A., & Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Berlin: Springer.

Hamada, M. S., Wilson, A. G., Reese, C. S., & Martz, H. F. (2008). *Bayesian Reliability*. New York: Springer.

Hickernell, F. J. (1998). A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, *67*, 299–322.

Hickernell, F. J. (2002). Obtaining $O(N^{-2+\varepsilon})$ convergence for lattice quadrature Rrules. In K. T. Fang, F. J. Hickernell, & H. Niederreiter (Eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2000* (pp. 274–289). Berlin: Springer.

Kallen, M. J., & van Noortwijk, J. M. (2005). Optimal maintenance decisions under imperfect inspection. *Reliability Engineering and System Safety*, *90*, 177–185.

Lawless, J., & Crowder, M. (2004). Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Analysis*, *10*, 213–227.

Lu, C. J., & Meeker, W. Q. (1993). Using degradation measures to estimate a time-to-failure distribution. *Technometrics*, *35*, 161–174.

Lu, D., Pandey, M. D., & Xie, W. C. (2013). An efficient method for the estimation of parameters of stochastic gamma process from noisy degradation measurements. *Journal of Risk and Reliability*, *227*, 425–433.

Meeker, W. Q., & Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. New York: Wiley.

Nelson, W. (1990). *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*. New York: Wiley.

Padgett, W. J., & Tomlinson, M. A. (2004). Inference from accelerated degradation and failure data based on Gaussian process models. *Lifetime Data Analysis*, *10*, 191–206.

Park, C., & Padgett, W. J. (2005). Accelerated degradation models for failure based on geometric Brownian motion and gamma process. *Lifetime Data Analysis*, *11*, 511–527.

Peng, C. Y., & Tseng, S. T. (2013). Statistical lifetime inference with skew-Wiener linear degradation models. *IEEE Transactions on Reliability*, *62*, 338–350.

Peng, C. Y. (2015a). Inverse Gaussian processes with random effects and explanatory variables for degradation data. *Technometrics*, *57*, 100–111.

Peng, C. Y. (2015b). Optimal classification policy and comparisons for highly reliable products. *Sankhyā B*, *77*, 321–358.

Peng, C. Y. & Cheng, Y. S. (2016), Threshold degradation in R using iDEMO. In M. Dehmer, Y. Shi, & F. Emmert-Streib (Eds.) *Computational Network Analysis with R: Applications in Biology, Medicine and Chemistry* (pp. 83–124). Germany: Wiley-VCH Verlag GmbH & Co. https://doi.org/10.1002/9783527694365.ch4.

Peng, C. Y., & Tseng, S. T. (2009). Misspecification analysis of linear degradation models. *IEEE Transactions on Reliability*, *58*, 444–455.

Richtmyer, R. D. (1951). The evaluation of definite integrals, and a quasi-Monte-Carlo method based on the properties of algebraic numbers. Technical Report LA-1342, Los Alamos Scientific Laboratory.

Shiomi, H., & Yanagisawa, T. (1979). On distribution parameter during accelerated life test for a carbon film resistor. *Bulletin of the Electrotechnical Laboratory*, *43*, 330–345.

Si, X. S., Wang, W. B., Hu, C. H., Zhou, D. H., & Pecht, M. G. (2012). Remaining useful life estimation based on a nonlinear diffusion degradation process. *IEEE Transactions on Reliability*, *61*, 50–67.

Singpurwalla, N. D. (1995). Survival in dynamic environments. *Statistical Science*, *10*, 86–103.

Singpurwalla, N. D. (1997), Gamma processes and their generalizations: an overview. In R. Cook, M. Mendel, & H. Vrijling, (Eds.), *Engineering Probabilistic Design and Maintenance for Flood Protection* (pp. 67–73). Dordrecht: Kluwer Academic.

Sloan, I. H., & Joe, S. (1994). *Lattice Methods for Multiple Integration*. Oxford: Oxford University Press.

Suzuki, K., Maki, K., & Yokogawa, S. (1993). An analysis of degradation data of a carbon film and the properties of the estimators. In K. Matusita, M. L. Puri, & T. Hayakawa, (Eds.), *Proceedings of the Third Pacific Area Statistical Conference* (pp. 501–511). Zeist: The Netherlands.

Tsai, C. C., Tseng, S. T., & Balakrishnan, N. (2012). Optimal design for gamma degradation processes with random effects. *IEEE Transactions on Reliability*, *61*, 604–613.

van Noortwijk, J. M. (2009). A survey of the application of gamma processes in maintenance. *Reliability Engineering and System Safety*, *94*, 2–21.

Wang, X., & Xu, D. (2010). An inverse Gaussian process model for degradation data. *Technometrics*, *52*, 188–197.

Whitmore, G. A. (1995). Estimating degradation by a Wiener diffusion process subject to measurement error. *Lifetime Data Analysis*, *1*, 307–319.

Whitmore, G. A., & Schenkelberg, F. (1997). Modeling accelerated degradation data using Wiener diffusion with a time scale transformation. *Lifetime Data Analysis*, *3*, 27–45.

Yanagisawa, T., & Kojima, T. (2005). Long-term accelerated current operation of white light-emitting diodes. *Journal of Luminescence*, *114*, 39–42.

Ye, Z. S., & Chen, N. (2014). The inverse Gaussian process as a degradation model. *Technometrics*, *56*, 302–311.

Zhou, Y., Sun, Y., Mathew, J., Wolff, R., & Ma, L. (2011). Latent degradation indicators estimation and prediction: a Monte Carlo approach. *Mechanical Systems and Signal Processing*, *25*, 222–236.

Part V
Recent Advances in Statistical Methods

Chapter 9
A Least Squares Method for Detecting Multiple Change Points in a Univariate Time Series

Kyu S. Hahn, Won Son, Hyungwon Choi and Johan Lim

Abstract Detecting and interpreting influential turning points in time series data is a routine research question in many disciplines of applied social science research. Here we propose a method for identifying important turning points in a univariate time series. The most rudimentary methods are inadequate when the researcher lacks preexisting expectations or hypotheses concerning where such turning points ought to exist. Other alternatives are computationally intensive and dependent on strict model assumptions. Our method is fused LASSO regression, a variant of regularized least squares method, providing a convenient alternative for estimation and inference of multiple change points under mild assumptions. We provide two examples to illustrate the method in social science applications. First, we assessed the validity of our method by reanalyzing the Greenback prices data used in (Willard et al. in Am Econ Rev 86:1001–1017, 1996). We next used the method to identify major change points in President Clinton's approval ratings.

K. S. Hahn (✉)
Department of Communication, Seoul National University,
Seoul, South Korea
e-mail: kyuhahn@snu.ac.kr

W. Son
Bank of Korea, Seoul, South Korea
e-mail: wons@bok.or.kr

H. Choi
Saw Swee Hock School of Public Health, National University
of Singapore, Singapore, Singapore
e-mail: hwchoi@nus.edu.sg

J. Lim
Department of Statistics, Seoul National University, Seoul, South Korea
e-mail: johanlim@snu.ac.kr

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_9

9.1 Introduction

Inquiries about the changes in the state of matters are core research questions in physical and social sciences alike. Social scientists recognize the notion of a turning point because many important social theories are inspired by inquiries into structural shifts in the historical trend. Intuitively, a turning point occurs when a series of observations which had been moving in one direction reverses or changes its course for some duration of time. In other words, turning points give rise to changes in overall direction or regime in a *determining* fashion. Conceptually speaking, turning points are best characterized as short, consequential shifts that redirect a process. Mathematically, a turning point is a maximum or minimum point in some continuous function, the point at which the slope of the function changes sign. This is often called *change points* in the classical probability theory.

In sociology, the concept of turning point has permeated the literature for a long time, with main application in studies of life course. In the life course literature, as Elder (1985) argued, some events are conceived as real important turning points in life as they redirect paths. According to this view, such turning points would interrupt regular patterns and provide major insights in the life course literature. For example, in analyzing criminal careers across individual lives, Sampson and Laub (2005) argued that marriage was often a key turning point in the process of desistance from crime and viewed marriage as a potential causal force in desistance that operates as a dynamic, time-varying process through time.

Similar arguments have been echoed frequently in other disciplines. In political science, turning points have been sought in studies of political realignment. For example, Lasser (1985) defined realignments as fundamental shifts in the structure of the party system, marked by changes in voting behavior and in the basic party attachments of the voting citizens. In particular, the author defined critical realignments as extraordinary upheavals in the flow of American electoral and policy history that occur under conditions of abnormal and general crisis. The notion of turning points has also formed the conceptual basis for developing theories concerning critical elections. (e.g., Key 1955; Burnham 1970; Clubb et al. 1981). First enunciated by Key (1955), the theory of realigning elections suggested that certain critical elections created sudden, massive shifts in the electorate, where a new governing coalition installed which group would represent the majority for decades until the next critical election. In light of this view, Burnham (1970) argued that critical elections were marked by short, sharp reorganizations of the mass coalitional base of the major parties which occurred at periodic intervals on the national level (page 10 of Burnham 1970).

In applied economics, studies of business cycles and other economic regularities have led to widespread analysis of turning points (e.g., Chaffin and Talley 1989; Zellner et al. 1991). For example, Zellner et al. (1991) employed two variants of an autoregressive, leading indicator model to forecast turning points in the growth rates of annual real output across 18 countries between 1974 and 1986. The authors also employed Bayesian predictive densities to compute probabilities of downturns and

upturns. It is worth noting that, in the history of science, revolution has also been a central concept (see Kuhn 1970; Cohen 1985).

Conceptually, as described by Abbott (1997), a social process is organized into *trajectories*. As the author described, here trajectories are considered inertial (and enduring) variation without change where consistent causal regimes persist. In light of this view, for example, a life course can be parsed into trajectories and transitions Elder (1985). Trajectories are interdependent sequences of events in different areas of life whereas transitions are radical shifts Elder (1985). In the current analysis, we are interested in empirically finding these transitions. In short, a social process can be viewed as a sequence of trajectories linked to one another via turning points.

It should be noted that what constitutes a turning point is not necessarily the change of sign. Instead, turning points involve the *separation* of smooth tracks by fairly abrupt and diversionary moments Abbott (1997). In other words, as Abbott (1997) explained, what characterizes the trajectories is their *inertial* quality. Turning points endure large amounts of minor variation without any appreciable changes in overall direction or regime. On the other hand, a true turning point can be distinguished from a mere random episode or minor ripples because its separates long, enduring segments in a series of observations.

Given this conceptual definition, how would one empirically identify a major turning point? In practice, it is not easy to set an operational definition for this intrinsically important concept. Where specific hypotheses are posited, the most direct approach is to define a dummy variable that takes on 0 before an expected change point and takes on 1 afterward. In this setting, the dummy variable can be used to describe a mean shift. Accordingly, one could begin by assuming that the turning point came in one point rather than some other point, fit a statistical model with this period effect, and look at the proportion of the variability of outcome explained by the dummy variable. However, if uncertainty prevails about precisely when such turning points should be expected to occur, choosing potential candidates for turning points becomes arbitrary.

There is also no shortage on the subject of detecting changes in the model parameters of a stochastic system in the statistical literature. Dating back more than a few decades, Hinkley (1970) proposed a method for sequential hypothesis testing where the authors attempted to determine potential change points based on a theoretically derived approximate distribution of test statistics. Siegmund (1986) focused on detecting change points in a stochastic process, which required a rigorous approximation of tail probabilities.

In light of our definition of a turning point, however, the existing approaches suffer from various shortcomings. Many existing approaches are online methods that search for turning points in real time without having observed the ensuing data points (e.g., see Lai 2001 and references therein). As Abbott (1997) pointed out, however, that turning point analysis is sensible only after when a new trajectory or system state is clearly and entirely established. Accordingly, one must observe the entire series in order to distinguish true turning points from random changes. To further elaborate on this point, an undifferentiated smooth curve cannot be regarded as having a turning point, although it might clearly involve long-term change. On the other hand, even

if the slope of the function changes its sign in quick succession, their magnitudes ought to be quite substantial to be considered as anything more than minor ripples in a generally monotonic trend. In short, the so-called online methods are not suitable for operationalizing tuning points as defined in most applied works.

Another problematic feature of existing approaches is that they assume that there is only a single turning point in a given stochastic process. Accordingly, in order to detect multiple turning points, one must apply the same procedure Yang and Kuo (2001) *recursively*. This is because simultaneously searching multiple turning points in a given time process can posit difficulties such as determination of the total number of change points. Often many existing methods compare the fit of alternative models differentially specifying the number and the location of turning points. In this setting, the researcher must first make an assumption about the number of turning points in the given process. In applied research, however, the number of turning points would rarely be known to the researcher in advance. Accordingly, most of the existing methods poorly operationalize turning points.

Finally, the lack of formal testing procedures often limits the ability of researchers to account for the uncertainty associated with the process. In practice, the distinction between trajectories and turning points will be less obvious. Most trajectories will have few turnings, and the researcher could mistakenly ignore real structural changes or mistake random variation for structural changes. On the other hand, if a model is obtained after diagnostic tests are used to locate change points, conventional t-statistics and p-values no longer reflect the underlying prior uncertainty about the timing of structural shifts Western and Kleykamp (2004). For instance, Isaac and Griffin (1989) attempted to detect structural instability with plots of regression coefficients from a time window moving along the series. Currently, therefore, not many existing methods provide a tool for locating change points that can also be easily combined with a procedure for assessing uncertainties concerning their locations.

The Bayesian methods overcome many of these shortcomings. With increased accessibility to Markov chain Monte Carlo methods as a computational aid in Bayesian inference, many researchers have applied the so-called retrospective or off-line approach to the change point problem. For example, Carlin et al. (1992) first proposed a hierarchical model that completely characterized the distribution of a *single* change point. Green (1995) devised a novel class of MCMC methods (or samplers) for detecting *multiple* change points in a series of observations, providing a Bayesian solution to one of the key shortcomings of the existing methods. Chib (1998) proposed a framework for hypothesis testing regarding multiple change points further demonstrating the utility of the Bayesian framework for detecting turning points in a time process. Western and Kleykamp (2004) illustrated the utility of the Bayesian framework for social science applications in their analysis of real wage growth in 18 OECD countries from 1965 to 1992.

Nevertheless, the Bayesian methods also have some well-known weaknesses. The MCMC methods for Bayesian inference are computationally intensive even with the improved computer technology. Also, it is often hard to justify the convergence of the MCMC sampler, which has to traverse an excessively large parameter space of varying dimensions in the case of multiple change points. Perhaps most importantly,

these methods rely on statistical concepts that remain unfamiliar or unacceptable to many applied researchers.

For operationalizing turning points, the recently developed fused lasso (FL) method offers many advantages over the existing methods (see Mammen and van de Geer 1997; Tibshirani et al. 2005; Kim et al. 2009; Rinaldo 2009; Tibshirani 2014; Li et al. 2017). To begin with, the FL method is built on the ideas underlying least squares regression—a technique virtually all sociologists have become familiar with. More specifically, the FL method is based on least squares regression with ℓ_1-norm regularization, where regularization refers to the fact that the regression coefficients are shrunken toward zero simultaneously. To be more specific, by the regularization principle, for a given threshold λ, the shrunken mean estimates $\bar{\mu}_j$'s are

$$\bar{\mu}_j = \text{sgn}(\mu_j)\max(|\mu_j| - \lambda, 0), \tag{9.1}$$

where $\text{sgn}(x)$ is the sign function of x and μ_j is the mean vector y_j. Thus, Eq. (9.1) has the effect of shrinking μ_j to 0 if its absolute value is either smaller or greater than the given threshold λ.

Regularization has attracted the interest of users because of its applicability to automatic variable selection. Most notably, the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996) is based on the regularization principle and has been one of the most significant contributions to the variable selection problem in the past decade (also see Knight and Fu 2000; Leng et al. 2006; Zou 2006).

Applying this notion of regularization, we are interested in examining the change in the difference between any two neighboring observations, $\Delta_{j+1} - \Delta_j$, where $\Delta_j = \mu_j - \mu_{j-1}$. The difference in Δ_j of two neighboring observations equals each other unless they are tuning points, and their difference is shrunken to zero. To be more specific, for a given threshold λ, our method estimates the difference in regression coefficients of two neighboring observations, $\Delta_{j+1} - \Delta_j$, as

$$\overline{\Delta_{j+1} - \Delta_j} = \text{sgn}(\Delta_{j+1} - \Delta_j)\max(|\Delta_{j+1} - \Delta_j| - \lambda, 0), \tag{9.2}$$

where $\Delta_j = \mu_j - \mu_{j-1}$ and μ_j is the mean of y_j. Thus, trivial differences are suppressed toward zero. In short, the above procedure disregards insignificant changes, whereas the survivors of shrinkage can be considered important change points. Accordingly, the end result can inform us of the optimal number of change points, their exact locations, and statistical significance.

The FL method has a number of advantages over the alternative approaches. First, the method detects multiple change points simultaneously, without the need to iteratively search for one change point at a time. For applied work, this is an important advantage over other alternatives since often applied data contain much unknown sampling variation.

Second, the FL method is easy to implement. It relies on the simple least square method with a tuning parameter, and also many efficient algorithms to compute it are available (see Hoefling 2010; Tibshirani and Taylor 2011; Yu et al. 2015;

Arnold and Tibshirani 2016; Lee et al. 2017). LASSO-type regression models have become increasingly popular in many applied disciplines, making a wide range of computer programs available for public use (see Tibshirani and Taylor 2011; Yu et al. 2015; Arnold and Tibshirani 2016). With this increased accessibility to the existing computer software, the proposed estimation procedure can be easily implemented with little extra programming efforts.

Last but not least, it relies on relatively mild assumptions. In applied works, statistical inference is made under the circumstances that are quite discrepant from the ideal large sample world, an assumption typically made in the advanced statistical techniques designed for change point analysis. Accordingly, for example, despite its elegance and practical convenience, the traditional cusum-statistic-based approach (see Lai 2001 and references therein) can be problematic because it requires the fixed probability specification based on strong model assumptions. In contrast, the FL method is a least-square-based method and does not make a strong distributional assumption on the data.

The remainder of this article is organized as follows. First, we detail the formulation of our method and describe its properties. In order to illustrate its utility for applied work, we apply the proposed method to detecting major trend change points in the so-called Greenback prices during the US Civil War Willard et al. (1996). Instead of specifying a list of dates a priori and testing for their importance, Willard et al. (1996) compared the reactions of participants in financial markets to the significance the same events have been assigned by Civil War historians. We also applied the method described here to analyzing President Clinton's job approval ratings.

9.2 Method

There are at least two possible scenarios in which some of the mean parameters can be grouped into common values in a time series setting. First, the first several observations can be generated from a mean value up to a certain time point, while a certain number of ensuing observations are generated from a different mean value. Here the first group of observations can be used to infer their common mean value, and the second group of observations can be used to infer their own common mean value. Alternatively, the underlying mean value may keep increasing by a unit up to a certain time point and turn downwards thereafter; it can also decrease for a specific time period before changing its direction again. In this article, we focus on detecting trend changes, or the second scenario described above.

To best illustrate the core idea underlying the proposed method, suppose that we observe n observations, $y_1, y_2, \ldots, y_n$, from the underlying mean values $\mu_1, \mu_2, \ldots, \mu_n$. Given these n parameters, if there exist no constraints in the relationship between the mean values, the best guess for each mean parameter is the respective observation itself.

A turning point is where the trend significantly changes. A trend in a univariate time series is naturally defined by a change in the means of neighboring observations

$\Delta_j = \mu_j - \mu_{j-1}$, where μ_j is the mean of the j-th observation y_j. Thus, as described earlier, our method suppresses trivial differences toward zero. As stated earlier, in a univariate time series, the unregularized estimate of μ_j is y_j itself. To detect major turning points, we disregard insignificant changes by suppressing small $\Delta_{j+1} - \Delta_j$ to 0. In short, the procedure trivializes insignificant changes, whereas the survivors of shrinkage can be considered important turning points.

In practice, we solve the following ℓ_1-regularized least square problem:

$$\text{Minimize} \sum_{i=1}^{n} (y_i - \mu_i)^2 + \lambda \sum_{j=2}^{n-1} |\Delta_{j+1} - \Delta_j|, \tag{9.3}$$

whose solution shrinks $\Delta_{j+1} - \Delta_j$ toward to 0 yielding the result equivalent to the estimate in (9.2). Here, the tuning parameter λ determines the number of change points in a time series. If λ approaches ∞, all of the estimated $\Delta_{j+1} - \Delta_j$'s equal 0, and no turning points can be detected. On the other hand, if λ approaches 0, our procedure estimates $\Delta_{j+1} - \Delta_j$ as $(y_{j+1} - y_j) - (y_j - y_{j-1})$, which would rarely equal 0. In this case, nearly all data points would be classified as turning points. Therefore, we solve Eq. (9.3) by solving its dual problem:

$$\text{Minimize} \sum_{i=1}^{n} (y_i - \mu_i)^2 \quad \text{subject to} \quad \sum_{j=2}^{n-1} |\Delta_{j+1} - \Delta_j| \le s, \tag{9.4}$$

where $\sum_{j=2}^{n-1} |\Delta_{j+1} - \Delta_j| = |(\mu_3 - \mu_2) - (\mu_2 - \mu_1)| + \cdots + |(\mu_n - \mu_{n-1}) - (\mu_{n-1} - \mu_{n-2})|$.

In this formulation, the choice of the constraint s is crucial because it could significantly influence the results. Although there is no rule set in stone, one could improvise a reasonable approach for finding the optimal s. From the ways in which the parameters are defined, the least squares error is bound to decrease toward the error of an unconstrained least squares solution as we increase s. We thus search for an optimal point s^* such that the error reduction by a unit increase in s changes rapidly before and after s^*, and it becomes stable in a minor scale after s^*. It also requires a threshold, but this is often fairly clear when working with real data. Based on our experiments, about 90% of the total error reduction from the fully constrained model to fully unconstrained model is deemed reasonable. Figure 9.1 presents trade-off curves between the constraint and the residual sum of squares from our two examples illustrating the method outlined in this manuscript. From these plots, one can easily choose the optimal s for the given problem.

In order to assess the significance of potential change points detected by the non-zero mean difference term described earlier, a simple bootstrap method can be applied to calculate their p-values. First let $\varepsilon_i = y_i - \mu_i$ for all $i = 1, 2, \ldots, n$ denote the residuals, and these residuals follow some unknown distribution F. One can iterate the following procedure N times: (1) place prior probability $1/n$ on every residual ε_i

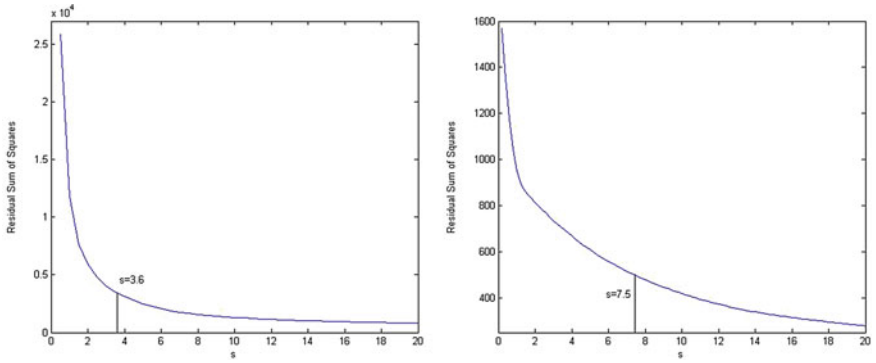


Fig. 9.1 Least squares errors of estimated models with varying s. The left panel is the scree plot for the Greenback price data in Willard et al. (1996). The right panel is that for Clinton's approval rate data

and sample with replacement to obtain $\bar{\varepsilon}^{(1)} = (\bar{\varepsilon}_1^{(1)}, \bar{\varepsilon}_2^{(1)}, \cdots, \bar{\varepsilon}_n^{(1)})$ from the original observations; and (2) using $y_i^{*(1)} = \hat{\mu}_i + \bar{\varepsilon}_i^{(1)}$ as a new observation for i where $\hat{\mu}_i$ is the original fit of the mean parameters, refit the mean parameters. While holding the optimal constraint s constant, reiterating (1) and (2) N times yields a bootstrap distribution of the mean parameters.

Based on this empirically derived distribution, one may obtain the 95% pointwise confidence interval of each mean parameter. In particular, keeping records of the elements in the constraint (i.e., the first order or the second-order differences), a given point is a turning point if the estimated mean differential in the original model fit is greater than that in $(1-\alpha)100\%$ of the bootstrap samples. For example, for the 5% significance level with 1,000 bootstrap samples, time i is not a change point if $|\hat{\mu}_{i+2}^{(k)} - 2\hat{\mu}_{i+1}^{(k)} + \hat{\mu}_i^{(k)}| < |\hat{\mu}_{i+2} - 2\hat{\mu}_{i+1} + \hat{\mu}_i|$ in more than 950 samples.[1]

9.3 Applications

9.3.1 Turning Points in the Greenback Prices During the US Civil War

In our first example, using the prices of the so-called Greenback, we attempt to determine the turning points in the US Civil War as viewed by people at the time. Using data on the gold price of Greenbacks, Willard et al. (1996) compared the

[1] As a side note, another issue with using ℓ_1-norm penalty is that the estimation is biased. One way to account for this potential bias is to identify the change points by shrinkage and refit the model piecewise between change points under quadratic loss. In an engineering application, Chen et al. (2001) adopted a similar idea in developing their basis pursuit technique. In our simple problem of finding the jumps in a constant mean process, this amounts to computing numeric average of data points.

reactions of participants in financial markets to the significance the same events have been assigned by Civil War historians. Instead of specifying a list of dates a priori and testing for their importance, the authors allowed the data to identify the important dates and compared them to historians' accounts.

In 1862, the USA issued an inconvertible currency called the Greenback. As detailed in Willard et al. (1996), as the Union's financial condition deteriorated in 1861, banks suspended the convertibility of their notes into gold suspecting a massive outflow of gold. Likewise, the government suspended the right to convert Treasury notes into specie. In early 1862, Congress authorized the government to issue an inconvertible currency popularly called Greenbacks. Accordingly, the Greenbacks represented promises to pay gold coin. However, the Greenback's value depreciated from par with the gold dollar, and a formal market for trading gold came into existence shortly after the suspension of convertibility.

Many argue that the Greenback's value reflected the expectation of future war costs. People expected that they could convert their Greenbacks to gold dollars one-for-one after the war. Accordingly, the price of a Greenback depended on its expected value in gold dollars. Therefore, the more costly the war, the less likely that this conversion would take place. In short, fluctuations in the Greenback prices can be revealing of how contemporaries perceived the status of the war. Conceptually, Willard et al. (1996) defined a break in the series as a shift in its mean value. According to the authors' definition, the breaks in the price of Greenbacks series marked the turning points of the war.

More specifically, adopting the approach of Banerjee et al. (1992), the authors first estimated the following regression equation using data from the 100-day period between 03/24/1862 and 07/19/1862:

$$\ln p_t = \beta_0 + \sum_{i=1}^{12} \beta_i \ln p_{t-i} + \varepsilon_t,$$

where p_t is the gold price of Greenbacks on day t, the β's are parameters to be estimated, and ε_t is a white noise error term. Subsequently, the authors calculated the F-statistic associated with a test of the hypothesis that the coefficient on an omitted dichotomous variable is zero after choosing the lag length of 12 days. The authors repeated this procedure over and over, each time moving the 100-day window over one day, until the entire period of the war has been covered. The authors sequentially searched for peaks in the series of statistics, first picking the maximum and eliminating the window around that date, then searching for the next peak.

The authors claimed to have detected seven turning points at which financial markets reacted strongly. Of the seven turning points, the authors attributed two to well-known historical events: (1) 09/23/1862 and (2) 07/06/1863.

The authors claimed that 09/23/1862 corresponded to a costly Union victory in the battle at Antietam. As described in Willard et al. (1996), the battle itself cost so many lives that it could lead people to revise upward their estimates of the war's future costs. Also, according to the authors, an equally likely cause of this structural

break is the Emancipation Proclamation, which destroyed any hope for a peaceful settlement to the war.[2]

The authors attributed 07/06/1863 to Gettysburg and Vicksburg. Gettysburg and Vicksburg were clear and significant military victories for the Union. The authors argued that, since news of these two battles reached the east at about the same time, it was impossible to make any statistical distinctions between market reactions to the two separate events.

The authors attributed three other structural breaks or to less prominent historical events. First, 01/08/1863 was the day before the Congressional Ways and Means Committee approved an increase in the supply of Greenbacks $300 million. The authors argued that participants in financial markets viewed this proposal as an admission that the fiscal measures taken to that date were insufficient to meet the Union's needs. From this point of view, the proposal was a sign showing that the government expected the war to be more expensive than anticipated.

07/12/1864 was the largest shift (in absolute value of the percent change) of the entire war. On July 12, as described in Willard et al. (1996), after having approached to within five miles of the White House, Jubal Early's Army, partly in response to the hasty arrival of Union reinforcements, retreated. According to the authors, to the participants of the financial market, this marked the end of any serious threat by the Confederacy. The authors also argued that there might have been financial news as well. After Chase resigned as Treasury Secretary on 06/30/1864, William P. Fessenden was appointed Chase's replacement on July 1. Willard et al. (1996) suspected that it was possible that some character of Fessenden may have caused financial traders to evaluate the Greenback more highly.

The authors also classified 08/24/1864 as another turning point. Although no major military news could be connected to this date, they speculated that the fall of Fort Morgan may be a possible explanation.

Finally, the authors failed to match the remaining two turning points with any well-known historical events: (1) 08/27/1863 and (2) 03/08/1865. Willard et al. (1996) noted that either most military news around these dates was insignificant or did not match the direction of movement in the Greenback prices.

Our results agree with some of the findings in Willard et al. (1996) but also generate some discrepancies. As described earlier, in order to assess the significance level of potential change points, we relied on the bootstrap procedure described earlier. As summarized in Table 9.1, our analysis detected roughly 17 possible turning points.[3]

---

[2] According to the authors, although it did not go into effect until January 1, 1863, the actual structure of the proclamation was a realistic threat since there could no longer be any doubts about Lincoln's willingness to tolerate slavery. This led people to raise the expected cost of the war.

[3] In some cases, our method identified *clusters* of dates as possible turning points. This is because, with daily data, it is difficult to pinpoint a single date as a turning point in the entire series that consists of over 1,200 days due to lack of information in a single data point. In such cases, therefore, we regard the cluster of dates as one turning point affected by the same series of events. Likewise, the bootstrapped probabilities associated with a single date are fused. Accordingly, for any given time point t, it would be more appropriate to simultaneously consider the probabilities associated with the surrounding dates.

Table 9.1 Major change points in Greenback gold prices

	Turning points	Willard et al. (1996)
1	April 24, 1862	
2	May 24, 1862	
3	December 23, 1862	
4	February 23, 1863*	January 8, 1863
5	June 4, 1863	
6	August 15, 1863*	August 27, 1863
	August 19, 1863*	August 27, 1863
7	October 14, 1863	
8	March 2, 1864	
9	May 16, 1864	
	May 17, 1864	
10	July 18, 1864*	July 12, 1864
	July 19, 1864*	July 12, 1864
11	August 8, 1864*	August 24, 1864
	August 12, 1864*	August 24, 1864
12	October 3, 1864	
	October 4, 1864	
13	December 9, 1864	
14	February 16, 1865*	March 8, 1865
	February 17, 1865*	March 8, 1865
15	May 13, 1865	
16	July 29, 1865	
17	November 18, 1865	
	November 25, 1865	
	November 27, 1865	

Figure 9.2 presents the proportion in the bootstrapped samples of the second-order differential that exceeds the fit of the original observation at each time point. The constraint parameter was set at $s = 7.5$ (see Fig. 9.1), and it was the 90% error reduction point as was explained in the previous section.

Our analysis classified 02/23/1863 as one of the most likely turning points. A closer examination of the data reveals that sometime around January or February of 1863, a turning point has occurred. The series remained fairly constant during this period only with some localized fluctuation. Note that Willard et al. (1996) had identified 01/08/1863 as a turning point. Although there is a fair amount discrepancy between the two dates, given the shape of the series around this period, it is not clear whether this should be regarded a turning point. We believe that 02/23/1863 is a more reasonable choice.

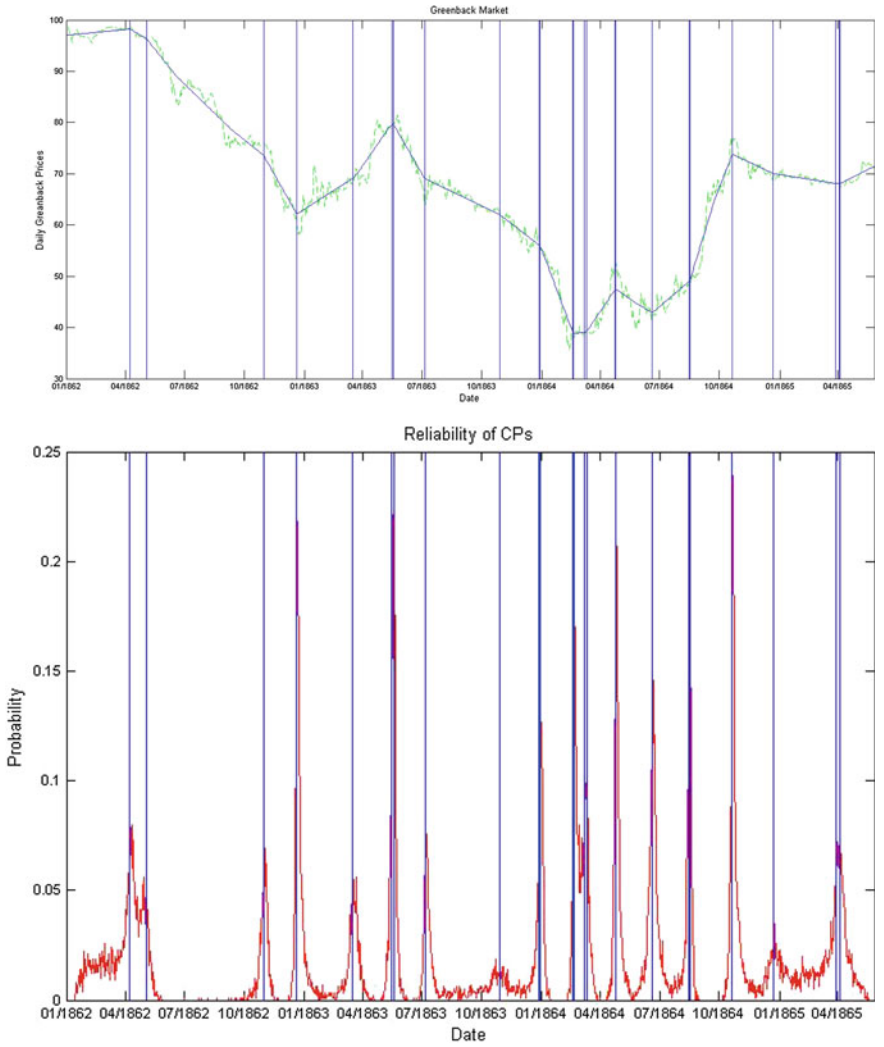


Fig. 9.2 Change points in Greenback gold prices

Our method also identified 08/15/1863 and 08/19/1863 as possible turning points; to roughly estimate their significance, when combined the probabilities associated with the two dates reached approximately 80%. A close look at the data reveals that these two turning points closely correspond to one of the turning points identified by Willard et al. (1996), 08/27/1863, although the authors were unable to match this particular date with a well-known event.

Our method also identified a set of turning points in 1864. First, our analysis classified July 18 and 19 as possible turning points. When combining the bootstrap

probabilities associated with three dates surrounding these two consecutive dates (i.e., from July 14 to 23), they reached approximately 93.4%, clearly indicating that there was a turning point around these dates. These two dates roughly matched one of the turning points identified by Willard et al. (1996), 07/12/1864. Likewise, August 8 and 12 were also classified as possible turning points in 1864. These two dates closely corresponded to another turning point (i.e., August 24) in Willard et al. (1996) analysis.

Our method also identified many other likely turning points overlooked in Willard et al. (1996). First, 04/02/1862 and 05/24/1862 were classified as possible turning points. A close examination of Fig. 9.2 reveals that the slope of the series changes sharply around these two dates. The Greenback prices seem to rise roughly until April or May of 1862 but started declining thereafter for about a year until February of 1863. Accordingly, we believe that any reasonable method ought to uncover turning points somewhere around these two dates. On the other hand, as described earlier, Willard et al. (1996) method only identified 01/08/1863 as a turning point; Fig. 9.2, however, reveals that the Greenback prices had already been declining for a while by this date.

Between 1862 and 1863, there were a couple of other ambiguous dates that were classified as possible turning points in our analysis. First, our method identified 12/23/1862 as a possible turning point. To what extent this date fits the profile of a true turning point seems debatable however; although the rate of decline increases around this date, it is not clear whether it should be regarded as a structural change since the general trend seems to remain unaltered. Similarly, 06/04/1863 was classified as a possible turning point. However, a closer look at the data reveals that it might be appropriate to call this date only a minor ripple, not a structural change. Although the rate of rise does seem to increase slightly around this date, its impact on the existing upward trend seems unaffected. In short, borrowing Willard et al. (1996) terminology, these two dates might be closer to being blurbs than true turning points.

The FL method also identified a few ambiguous dates in 1864. To begin, although 03/02/1864 was classified as a possible turning point in our analysis, Fig. 9.2 reveals that it is likely to be a minor ripple. Likewise, our analysis showed that 10/03/1864, 10/04/1864, and 12/09/1864 was also classified as a possible turning point. Indeed, as shown in Fig. 9.2, the direction of the curve changed around these dates, and arguably they could be regarded turning points. On the other hand, one might consider them as minor ripples since the overall upward trend seems to remain unaltered despite the existence of these local blurbs. The bootstrap probabilities associated with these dates are fairly low by any standard.

Finally, the FL method also identified a few turning points in 1865. First, 07/29/1865 was classified as a possible turning point. Again, however, we believe that it would be more appropriate to consider it a minor ripple. Indeed, the probabilities associated with this date (and the surrounding dates) seem rather low to consider it a major structural change. On the other hand, our method classified November 18, 25, and 27 as possible turning points. This indicates that our method suggests that a major structural change has occurred sometime in November. Indeed, the trend

seems to be changing around November of 1865. On the other hand, since the war ended quickly after these dates, it is unclear whether they can be regarded major turning points as shown by the bootstrap probabilities around this date.

9.3.2 President Clinton's Job Approval

As Lawrence and Bennett (2001) noted, in one of the great political ironies of modern times, Bill Clinton weathered a year-long sexual and obstruction of justice scandal and became only the second president in US history to be impeached, while maintaining some of the most impressive public approval ratings of any modern president.

Scandals have become such a strong influence on the way that Americans view their political leaders that it has even been surmised that scandals are the primary means of conflict within American politics (Williams 2000; Sabato 1991). The most obvious effect of a presidential scandal would be a negative impact on presidential approval because it would affect the president's perceived integrity Greene (2001). For President Clinton, however, no scandalous events seemed to have affected his job approval. Most notably, his high approval during and after impeachment presented a stronger challenge to the conventional models of presidential approval rating.

Many political scientists argue that the simplest explanation for Clinton's continuing popularity is probably the most compelling: to borrow the watchword from his own 1992 campaign, the economy stupid. For example, public opinion scholar Zaller (1998) argued that the public's continued support for Clinton could be accounted for by reference to three fundamental variables: peace, prosperity, and Clinton's moderate policy positions. Similarly, Jacobson (1999) argued that the public would have reacted differently to the scandals if the economy had been in bad shape.

Others have cited another factor: The American public made another distinction that proved to be crucial—a distinction between private and public matters Kagay (1999). Especially concerning the Lewinsky scandal, most of the American public classified Clinton's sex scandal as being in the private zone (see Greene 2001).

In this subsection, to illustrate the FL method, we revisit Clinton's job approval rating. Our objective is to assess whether the president's approval rating was vulnerable to various scandalous events. Most of previous analyses have been based on several polls taken during the period immediately following a particular scandal. In contrast, our approach provides a more comprehensive assessment of the connection between Clinton's job approval and scandalous events.

We use the Gallup Poll ratings of presidential popularity during the Clinton presidency (01/1992 to 12/2000). These time series data have been used for almost all relevant studies of presidential popularity, particularly because the Gallup Poll has a high level of reliability since it regularly asks the identical question: "Do you approve or disapprove of the way —— is handling his job as president?"

Since most relevant studies of presidential popularity used the month as a unit of analysis, when the ratings were collected more than once per month, we chose the first ratings observed in a month (see Gronke and Brehm 2002). Also, as suggested in other

Table 9.2 Major change points in President Clinton's job approval ratings

Date	Events
June, 1993	A missile attack aimed at Iraq's intelligence headquarters
January, 1994	Attorney General Janet Reno announcing the appointment of an Independent Counsel to investigate Whitewater Development Corporation
November, 1994	Republicans' landslide victory in the House
March, 1998	Former Clinton aide Kathleen Willey's appearance on CBS's 60 min, confirming the president made a sexual advance to her in the White House in 1993
December, 1999	Clinton's impeachment by the House of Representatives

studies, we discarded the first five observations in President Clinton's presidency because a president's popularity typically starts off unusually high.

We use the least squares regression with a set of linear constraints by rewriting the constraints in (9.4) as linear ones. That is, we apply shrinkage to the second-order differences simultaneously in order to detect turning points in the *trend*. The choice of this constraint matrix was guided by the observation that Clinton's approval rating generally climbed throughout his presidency, while a few scandalous events placed his increasing popularity under scrutiny.

Figure 9.2 shows the results obtained after applying the shrinkage estimation to the monthly approval ratings. On the other hand, the time points at which the slope alters are potential candidates for important change points. Table 9.2 lists all the turning points identified in our analysis.

As can be seen from Fig. 9.3, the first major change point occurred in June of 1993. On June 26, 1993, President Clinton ordered to launch a missile attack aimed at Iraq's intelligence headquarters in Baghdad in retaliation against an Iraqi plot to assassinate President Bush. A major military action often becomes a critical turning point for presidents' popularity. Numerous studies (i.e., Parker 1995) have explored the impact that international conflicts can have on public opinion, particularly focusing on the rally effect that occurs early on in a conflict (e.g., Levy 1989; Russett 1990).

Our procedure identified January of 1994 as another change point in Clinton's job approval rating. On January 12, 1994, following President Clinton's request, Attorney General Janet Reno announced she was appointing an Independent Counsel to investigate Whitewater Development Corporation. The deposition was a first for a sitting president and first lady. The outbreak of the so-called Whitewater Scandal seemed to have a significant impact on Clinton's job approval ratings.

November of 1994 was also identified as a significant change point. Republicans' landslide victory in the House election might explain this finding. It created the first GOP majority in forty years. As a side note, the 104-th Congress also selected Newt Gingrich as Speaker. As the architect of the Contract with America, Gingrich became Clinton's principal political adversary.

Perhaps most interestingly, our method identified March of 1998 as a major change point in the president's approval rating. This roughly corresponds to the height of the

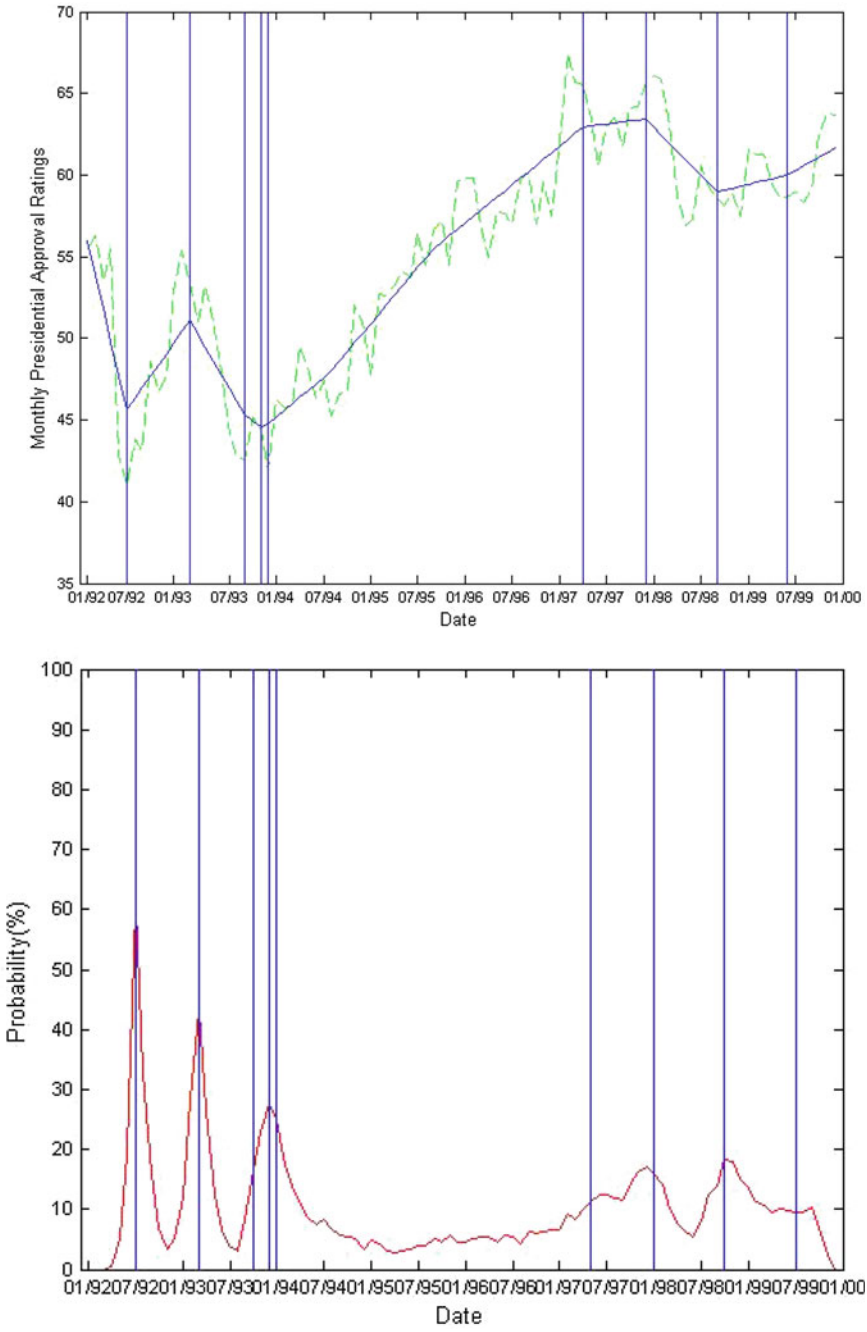


Fig. 9.3 Change points in Clinton's job approval ratings

Lewinsky scandal. Our interpretation is that, contrary to the view widely accepted by many scholars, the Lewinsky scandal was a major turning point in public approval of President Clinton. In January of 1998, Lewinsky's name began surfacing in an Internet gossip column, the Drudge Report, which mentioned rumors that the Newsweek had decided to delay publishing a piece on Lewinsky and the alleged affair. Subsequently, several news organizations reported the alleged sexual relationship between Lewinsky and Clinton. Toward the end of January, the president however declared publicly that he had not had sexual relations with Monica Lewinsky and "never told anybody to lie." On March 15, however, former Clinton aide Kathleen Willey appeared on CBS' 60 min, saying the president made a sexual advance to her in the White House in 1993. Then, there was little doubt that Clinton was guilty as charged.

Finally, December of 1999 was also identified as a major change point. Clinton was impeached by the House of Representatives on a largely party-line vote. Republicans acted despite considerable losses during mid-term elections the month before, which most commentators expected would cool the GOP's ardor for running Clinton out of office.

9.4 Conclusion

Although social scientists are often interested in identifying important turning points in time series data, the change point analysis has not become a popular data analysis technique in social science disciplines. The most rudimentary methods are inadequate when the researcher lacks preexisting expectations or hypotheses concerning where such turning points ought to exist. Other alternatives are computationally intensive and implementation of these methods require expert-level programming efforts. Also, they are mostly based on strong model assumptions that are not representative of acquired data. In addition, many applied social science researchers are unfamiliar with the statistical concepts underlying these methods.

In this paper, we proposed a regression-based technique specifically designed for identifying major change points in a univariate time series. Our least squares method provides a useful alternative to the existing techniques as it is based on a concept that most applied social science researchers have mastered. It is also easy to implement and can be accompanied by a convenient inferential procedure.

To illustrate its utility in social science applications, we provided two applications. First, we reanalyzed Greenback prices data used in Willard et al. (1996). The Greenback values can be used to assess how contemporaries view the status of the US Civil War. Our results were partially consistent with the authors' findings; but, our method also identified a couple of turning points that had gone undetected by the authors' original analysis. We also applied the proposed technique to identifying major change points in President Clinton's approval ratings. Our method identified many of the scandalous events during the Clinton presidency as major change points, even though they had been previously thought as having trivial effects on the president's job approval rating.

References

Abbott, A. (1997). On the concept of turning point. *Comparative Social Research*, *16*, 85–106.

Arnold, T., & Tibshirani, R. (2016). Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, *25*(1), 1–27.

Banerjee, A., Lumsdaine, R. L., & Stock, J. H. (1992). Recursive and sequential tests of the unit-root and trend-break hypotheses: Theory and international evidence. *Journal of Business & Economic Statistics*, *10*(3), 271–287.

Burnham, W. D. (1970). *Critical elections and the mainsprings of American politics*. New York: Norton.

Carlin, B. P., Gelfand, A. E., & Smith, A. F. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics*, *41*, 389–405.

Chaffin, W. W., & Talley, W. K. (1989). Diffusion indexes and a statistical test for predicting turning points in business cycles. *International Journal of Forecasting*, *5*(1), 29–36.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, *43*(1), 129–159.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, *86*(2), 221–241.

Clubb, J. M., Flanagan, W. H., & Zingale, N. H. (1981). *Party Realignment*. Beverly Hills: Sage.

Cohen, I. B. (1985). *Revolution in Science*. Boston: Harvard University Press.

Elder, G. H, Jr. (1985). Perspectives on the life course. In Glen H. Elder, Jr. Ithaca (Eda.), *Life course dynamics: Trajectories and transitions* (pp. 1968–1980) Cornell University Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732.

Greene, S. (2001). The role of character assessments in presidential approval. *American Politics Research*, *29*(2), 196–210.

Gronke, P., & Brehm, J. (2002). History, heterogeneity, and presidential approval: A modified ARCH approach. *Electoral Studies*, *21*(3), 425–452.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, *57*, 1–17.

Hoefling, H. (2010). A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, *19*(4), 984–1006.

Isaac, L. W., & Griffin, L. J. (1989). Ahistoricism in time-series analyses of historical process: Critique, redirection, and illustrations from US labor history. *American Sociological Review*, *54*, 873–890.

Jacobson, G. C. (1999). Impeachment politics in the 1998 congressional elections. *Political Science Quarterly*, *114*(1), 31–51.

Kagay, M. R. (1999). Presidential address: Public opinion and polling during presidential scandal and impeachment. *The Public Opinion Quarterly*, *63*(3), 449–463.

Key, V. O, Jr. (1955). A theory of critical elections. *The Journal of Politics*, *17*(1), 3–18.

Kim, S.-J., Koh, K., Boyd, S., & Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM Review*, *51*(2), 339–360.

Knight, K., & Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*, 1356–1378.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (2nd ed.). Chicago: University of Chicago Press.

Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges (with discussion). *Statistica Sinica*, *11*, 303–408.

Lasser, W. (1985). The supreme court in periods of critical realignment. *The Journal of Politics*, *47*(4), 1174–1187.

Lawrence, R. G., & Bennett, W. L. (2001). Rethinking media politics and public opinion: Reactions to the Clinton's Lewinsky scandal. *Political Science Quarterly*, *116*(3), 425–446.

Lee, T., Won, J-H., Lim, J., & Yoon, S. (2017). Large-scale structured sparsity via parallel fused lasso on multiple GPUs. *Journal of Computational and Graphical Statistics*.

Leng, C., Lin, Y., & Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, *16*, 1273–1284.

Levy, J.S. (1989). The diversionary theory of war: A critique. In Manus I. Midlarsky (Ed.), *Handbook of war studies* (pp. 259-288). Unwin Hyman, Boston.

Li, K., Sharpnack, J., Rinaldo, A., & Tibshirani, R. (2017). Approximate recovery in changepoint problems, from ℓ_2 estimation error rates. arXiv:1606.06746v2.

Mammen, E., & van de Geer, S. (1997). Locally adaptive regression splines. *The Annals of Statistics*, *25*(1), 387–413.

Parker, S. L. (1995). Towards an understanding of "rally" effects: Public opinion in the Persian Gulf War. *Public Opinion Quarterly*, *59*(4), 526–546.

Rinaldo, A. (2009). Properties and refinements of the fused lasso. *The Annals of Statistics*, *37*(5B), 2922–2952.

Russett, B. (1990). Economic decline, electoral pressure, and the initiation of interstate conflict. In Charles S. Gochman & Alan Ned Sabrosky (Eds.), *Prisoners of War? Nation-States in the Modern Era* (pp. 123–140). Lexington: Lexington Books.

Sampson, R. J., & Laub, J. H. (2005). A life-course view of the development of crime. *The Annals of the American Academy of Political and Social Science*, *602*(1), 12–45.

Sabato, L. (1991). *Feeding frenzy: How attack journalism has transformed American politics*. Free Pr.

Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *The Annals of Statistics*, *14*, 361–404.

Tibshirani, R. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, *42*(1), 285–323.

Tibshirani, R., & Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, *39*(3), 1335–1371.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, *58*, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, *67*(1), 91–108.

Willard, K. L., Guinnane, T. W., & Rosen, H. S. (1996). Turning points in the civil war: Views from the greenback market. *American Economic Review*, *86*(4), 1001–1018.

Williams, R. (2000). *Political Scandals in the USA*. Chicago: Fitzroy Dearborn Publishers.

Western, B., & Kleykamp, M. (2004). A Bayesian change point model for historical time series analysis. *Political Analysis*, *12*(4), 354–374.

Yang, T., & Kuo, L. (2001). Bayesian binary segmentation procedure for a Poisson process with multiple changepoints. *Journal of Computational and Graphical Statistics*, *10*(4), 772–785.

Yu, D., Won, J.-H., Lee, T., Lim, J., & Yoon, S. (2015). High-dimensional fused lasso regression using majorizationminimization and parallel processing. *Journal of Computational and Graphical Statistics*, *24*(1), 121–153.

Zaller, J. R. (1998). Monica Lewinsky's contribution to political science. *PS. Political Science & Politics*, *31*(2), 182–189.

Zellner, A., Hong, C., & Min, C. K. (1991). Forecasting turning points in international output growth rates using Bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques. *Journal of Econometrics*, *49*(1–2), 275–304.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Chapter 10
Detecting the Change of Variance by Using Conditional Distribution with Diverse Copula Functions

Jong-Min Kim, Jaiwook Baik and Mitch Reller

Abstract We propose new method for detecting the change of variance by using conditional distribution with diverse copula functions. We generate the conditional asymmetric random transformed data by employing asymmetric copula function and apply the conditional transformed data to the cumulative sum control (CUSUM) statistics in order to detect the change point of conditional variance by measuring the average run length (ARL) of CUSUM control charts by using Monte Carlo simulation method. We show that the ARLs of change point of conditional variance by CUSUM are affected by the directional dependence by using the bivariate Gaussian copula beta regression (Kim and Hwang 2017).

10.1 Introduction

The statistical process control chart was designed by Walter Shewhart at Bell Laboratories in 1924. Since then, Shewhart control charts have been applied in many different fields: hospital infection control (Sellick 1993), prediction of business failures (Theodossiou 1993), quality management of higher education (Mergen et al. 2000), corroborating bribery (Charnes and Gitlow 1995), athletic performance improvement (Clark and Clark 1997), and fault detection in NMOS fabrication (Lahman-Schalem et al. 2002). Shewhart control charts can monitor the stability or capability of the process by plotting an appropriate statistic on the graph and are easy to implement

J.-M. Kim (✉) · M. Reller
Statistics Discipline, Division of Science and Mathematics,
University of Minnesota, Morris, USA
e-mail: jongmink@morris.umn.edu

M. Reller
e-mail: relle041@morris.umn.edu

J. Baik
Department of Information Statistics, Korea National Open University,
Seoul, Republic of Korea
e-mail: jbaik@mail.knou.ac.kr

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_10

in manufacturing and service industries and good for detecting a large shift in the parameter. But they do not detect small or moderate parameter shift since they do not make use of previous observations when calculating statistic of interest for control. So Page (1954) developed cumulative sum (CUSUM) which is a time-weighted control chart that displays the cumulative sums (CUSUMs) of the deviations of each sample value from the target value.

Definition 1 CUSUM The standardized cumulative sum (CUSUM) control chart is formed by plotting the quantity:

$$Z_t = \frac{\overline{X}_t - \hat{\mu}_0}{\hat{\sigma}_{\overline{X}}}, \tag{10.1}$$

for each subgroup t and

$$S_{H_t} = \max\{Z_t - k + S_{H_{t-1}}, 0\},$$
$$S_{L_t} = \min\{Z_t + k + S_{L_{t-1}}, 0\}.$$

where μ_0 is the in-control mean, $\overline{X}_t$ is the average of the t-th sample, $\hat{\sigma}_{\overline{X}}$ is the known (or estimated) standard deviation of the sample mean. The CUSUM chart is made by plotting the values S_{H_t} and S_{L_t} against time; as long as the process remains in control centered at μ_0, the CUSUM plot will show variation in a random pattern centered about zero. If the process mean shifts upward, the charted CUSUM points will eventually drift upwards, and vice versa if the process mean decreases, (see Montgomery (1996)).

In quality control, Verdier (2013) applied copulas to multivariate charts, Dokouhaki and Noorossana (2013) proposed copula Markov CUSUM chart for monitoring the bivariate autocorrelated binary observations, Long and Emura (2014) proposed a control chart using copula-based Markov chain models, Emura (2015) developed R routines for performing estimation and statistical process control under copula-based time series models, and Busababodhin and Amphanthong (2016) reviewed copula modeling for multivariate statistical process control.

Recently, Kim and Hwang (2017) proposed a copula directional dependence method to explore a relationship between two financial time series based on two useful methods, the Gaussian copula marginal regression (GCMR) method by Masarotto and Varin (2012) and the beta regression model by Guolo and Varin (2014).

The aim of the paper investigates how much the directional dependence effect can influence on detecting the regime switching of variance by using copula directional dependence method with employing the traditional Inclán and Tiao (1994) CUSUM. We generate the conditional asymmetric random transformed data by employing asymmetric copula functions and stochastic volatility model and apply the conditional transformed data to the CUSUM so that we can evaluate the change point analysis of conditional variance by measuring the ARL of CUSUM control charts by using Monte Carlo simulation method.

The remainder of this paper is organized as follows: Sect. 10.2 describes the copula concepts, the directional dependence by copula, and the detection of change of the variance by copula. Section 10.3 is the illustrated example with a bivariate simulated data. Finally, conclusions are presented in Sect. 10.4.

10.2 Copula Method

10.2.1 Copula

A copula is a multivariate uniform distribution representing a way of trying to extract the dependence structure of the random variables from the joint distribution function. It is a useful approach to understanding and modeling dependent random variables. A copula is a multivariate distribution function defined on the unit $[0, 1]^n$, with uniformly distributed marginals. In this paper, we focus on a bivariate (two-dimensional) copula, where $n = 2$. Sklar (1959) shows that any bivariate distribution function, $F_{XY}(x, y)$, can be represented as a function of its marginal distribution of X and Y, $F_X(x)$ and $F_Y(y)$, by using a two-dimensional copula $C(\cdot, \cdot)$. More specifically, the copula may be written as

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)) = C(u, v),$$

where u and v are the continuous empirical marginal distribution function $F_X(x)$ and $F_Y(y)$, respectively. Note that u and v have uniform distribution $U(0, 1)$.

Therefore, the copula function represents how the function, $F_{XY}(x, y)$, is coupled with its marginal distribution functions, $F_X(x)$ and $F_Y(y)$. It also describes the dependent mechanism between two random variables by eliminating the influence of the marginals or any monotone transformation of the marginals.

Definition 3 A r-dimensional copula is a function $C : [0, 1]^r \to [0, 1]$ with the following properties:

1. For all $(u_1, \ldots, u_r) \in [0, 1]^r$, $C(u_1, \ldots, u_r) = 0$ if at least one coordinate of $(u_1, \ldots, u_r)$ is 0,
2. $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$, for all $u_i \in [0, 1]$, $(i = 1, \ldots, r)$,
3. C is r-increasing.

The n-dimensional random vector $X = (X_1, \ldots, X_n)$ is said to have a (non-singular) multivariate Student-t distribution with ν degrees of freedom, mean vector μ, and positive-definite dispersion or scatter matrix Σ, denoted $X \sim t_n(\nu, \mu, \Sigma)$, if its density is given by

$$f(x) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)^n|\Sigma|}}\left(1 + \frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{\nu}\right)^{-\frac{\nu+n}{2}},$$

Note that in this standard parameterisation $cov(X) = \frac{\nu}{\nu-2}\Sigma$ so that the covariance matrix is not equal to Σ and is in fact only defined if $\nu > 2$. Useful reference for the multivariate t is Demarta and McNeil (2005).

The multivariate n-variate Clayton copula is:

$$C^C(u_1, u_2, \ldots, u_n) = \varphi^{-1}\left(\sum_{i=1}^{n} \varphi(u_i)\right),$$

where φ is a function from [0, 1] to $[0, \infty)$ such that

(i) φ is a continuous strictly decreasing function,
(ii) $\varphi(0) = \infty$ and $\varphi(1) = 0$,
(iii) φ^{-1} is completely monotonic on $[0, \infty)$.

If the generator is given by $\varphi(u) = u^{-\theta} - 1$, then we get the multivariate n-variate Clayton copula as follows:

$$C^C(u_1, u_2, \ldots, u_n) = \left[\sum_{i=1}^{n} u_i^{-\theta} - n + 1\right]^{-1/\theta} \quad with \;\; \theta > 0$$

The multivariate n-variate Gumbel copula is given by

$$C^G(u_1, u_2, \ldots, u_n) = \exp\left\{-\left[\sum_{i=1}^{n}(-\ln u_i)^{\theta}\right]^{1/\theta}\right\} \quad with \;\; \theta > 1.$$

The generator is given by $\varphi(u) = (-\ln(u))^{\theta}$, and hence, $\varphi^{-1}(t) = \exp(-t^{1/\theta})$; it is completely monotonic if $\theta > 1$ (Table 10.1).

10.2.2 Directional Dependence by Copula

Guolo and Varin (2014) developed a marginal extension of the beta regression model for time series analysis and the cumulative distribution function of a normal variable.

Table 10.1 Archimedean copula functions

Copula	Copula function
FGM	$C^{FGM}(u, v, \theta) = uv + \theta uv(1-u)(1-v), \quad \theta \in (-1, 1]$
Clayton	$C^C(u, v, \theta) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}, \quad \theta \in (0, \infty)$
Frank	$C^F(u, v, \theta) = -\frac{1}{\theta}\log\left[1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right], \theta \in \mathbb{R} \setminus \{0\}$
Gumbel	$C^G(u, v, \theta) = \exp\left[-\left((-\log u)^{\theta} + (-\log v)^{\theta}\right)^{1/\theta}\right], \theta \geq 1$

The Guolo and Varin (2014) marginal beta regression model exploited the probability integral transformation to relate response Y_t to covariates x_t and to a standard normal error ϵ_t. The probability integral transformation implies that Y_t is marginally beta distributed, $Y_t \sim Beta(\mu_t, \kappa_t)$. Kim and Hwang (2017) assume that U_t given $V_t = v_t$ follows a beta distribution $Beta(\mu_t, \kappa_t)$ where $\mu_{U_t} = E(U_t|v_t)$. Kim and Hwang (2017) obtain the dependence of the response U_t on the covariate v_t by assuming a logit model for the mean parameter, $logit(\mu_{U_t}) = \mathbf{x}_t^T \beta_{\mathbf{x}}$, where $\beta_{\mathbf{x}}$ is a 2-dimensional vector of coefficients.

$$\text{logit}(\mu_{U_t}) = \log\left[\frac{\mu_{U_t}}{1-\mu_{U_t}}\right] = \beta_0 + \beta_1 v_t, \text{ where } t = 1, \ldots, n,$$

so that $\mu_{U_t} = \text{E}(U_t|v_t) = \frac{\exp(\beta_0+\beta_1 v_t)}{1+\exp(\beta_0+\beta_1 v_t)}$ with the correlation matrix of the errors corresponding to the white noise process.

$$\rho^2_{V_t \to U_t} = \frac{\text{Var}(\text{E}(U_t|v_t))}{\text{Var}(U_t)} = 12\text{Var}(\mu_{U_t}) = 12\sigma_U^2.$$

Directional dependence of the response U_t on the covariates $v_{1t}, v_{2t}, \ldots, v_{kt}$ is obtained by assuming a logit model for the mean parameter, $\text{logit}(\mu_{U_t}) = \mathbf{x}_t^T \beta_{\mathbf{x}}$, where $\beta_{\mathbf{x}}$ is a $k+1$-dimensional vector of coefficients,

$$\text{logit}(\mu_{U_t}) = \log\left[\frac{\mu_{U_t}}{1-\mu_{U_t}}\right] = \beta_0 + \sum_{k=1}^{k} \beta_k v_{kt}, \text{ where } t = 1, \ldots, n,$$

so that $\mu_{U_t} = \text{E}(U_t|v_{1t}, v_{2t}, \ldots, v_{kt}) = \frac{\exp(\beta_0+\sum_{k=1}^{k}\beta_k v_{kt})}{1+\exp(\beta_0+\sum_{k=1}^{k}\beta_k v_{kt})}$.

$$\rho^2_{(V_{1t}, V_{2t}, \ldots, V_{kt}) \to U_t} = \frac{\text{Var}(E(U_t|v_{1t}, v_{2t}, \ldots, v_{kt}))}{\text{Var}(U_t)}.$$

10.2.3 Detection of Change of Variance by Copula

By using direction dependence by copula, Kim and Hwang (2017), we can calculate the multivariate directional dependence measures based on the order of combinations of n-multivariate copula. In this research, we want to study how much the directional dependence by copula can affect to detect regime shift in the variance by employing the currently available methods of regime shift detection in the variance to exist in quantitative finance, where the concept of stock market volatility is very important. Since a change in the variance affects both the mean and variance charts, the CUSUM statistic for detecting these change points of variance has been much interested in the last two decades.

One of the most popular among the CUSUM methods is the Iterated Cumulative Sum of Squares (ICSS) algorithm developed by Inclán and Tiao (1994). Suppose $\{X_i\}$,

$i = 1, 2, \ldots, n$ is a series of independent, normally distributed random variables with zero mean and variance σ^2. The ICSS by Inclán and Tiao (1994) for retrospective detection of changes of variance is defined as

$$D_k = \frac{C_k}{C_n} - \frac{k}{n}, \quad \text{where } k = 1, \ldots, n,$$

so that $C_k = \sum_{i=1}^{k} X_i^2$. The algorithm consists of several steps, dividing time series $\{X_i\}$ into pieces and applying D_k to each of them iteratively.

In this research, we use diverse copula functions to generate the conditional transformed data by using conditional distribution by Archimedean copula functions. First of all, we want to explain how we perform the conditional transformed data by copulas. For illustration, Clayton Copula, Clayton (1978), is employed as follows:

Corollary 1 *Using one of Archimedean copulae, Clayton Copula, is*

$$C^C(u_1, u_2, \theta_{12}) = (u_1^{-\theta_{12}} + u_2^{-\theta_{12}} - 1)^{-1/\theta_{12}},$$

for $\theta_{12} > 0$, for two random variables X_1 and X_2, we can derive the conditional distribution of X_1 given X_2, $F_{1|2}(X_1|X_2; \theta_{12})$, as follows:

$$F_{1|2}(X_1|X_2; \theta_{12}) = \frac{\partial C^C(u_1, u_2, \theta_{12})}{\partial u_2}.$$

where $u_1 = F(X_1)$ and $u_2 = F(X_2)$.

Corollary 2 *Suppose we have three random variables X_1, X_2, X_3. Using one of Archimedean copulae, Clayton Copula, by Corollary 1, we can derive the following ones:*

$$C^C(u_2, u_3, \theta_{23}) = (u_2^{-\theta_{23}} + u_3^{-\theta_{23}} - 1)^{-1/\theta_{23}}$$

and

$$F_{3|2}(X_3|X_2; \theta_{23}) = \frac{\partial C^C(u_2, u_3, \theta_{23})}{\partial u_2},$$

where $u_2 = F(X_2)$ and $u_3 = F(X_3)$.

Since $F_{1|2}(X_1|X_2)$ and $F_{3|2}(X_3|X_2)$ are independent identically distributed as $U(0, 1)$, we can derive the conditional cumulative distribution function as follows:

$$F_{13|2}(X_1, X_3|X_2; \theta_{13|2}) = C^C(F_{1|2}(X_1|X_2; \theta_{12}), F_{3|2}(X_3|X_2; \theta_{23}); \theta_{13|2}), \quad (10.2)$$

where we denote $u_{1|2} = F_{1|2}(X_1|X_2; \theta_{12})$ and $u_{3|2} = F_{3|2}(X_3|X_2; \theta_{23})$.

We can derive the conditional distributions $F_{1|23}(X_1|X_2, X_3; \theta_{13|2})$, $F_{2|13}(X_2|X_1, X_3; \theta_{12|3})$ and $F_{3|12}(X_3|X_1, X_2; \theta_{13|2})$ by Corollary 1 as follows:

$$F_{1|23}(X_1|X_2, X_3; \theta_{13|2}) = \frac{\partial C^C(u_{1|2}, u_{3|2}; \theta_{13|2})}{\partial u_{3|2}},$$

where we denote $u_{1|23} = F_{1|23}(X_1|X_2, X_3; \theta_{13|2})$, likewise,

$$F_{2|13}(X_2|X_1, X_3; \theta_{12|3}) = \frac{\partial C^C(u_{1|3}, u_{2|3}; \theta_{12|3})}{\partial u_{1|3}},$$

where we denote $u_{2|13} = F_{2|13}(X_2|X_1, X_3; \theta_{12|3})$, and

$$F_{3|12}(X_3|X_1, X_2; \theta_{13|2}) = \frac{\partial C^C(u_{1|2}, u_{3|2}; \theta_{13|2})}{\partial u_{1|2}},$$

where we denote $u_{3|12} = F_{3|12}(X_3|X_1, X_2; \theta_{13|2})$.

Similarly, the conditional cumulative distribution functions for the multivariate n-variate t-copula, Frank copula, and Gumbel copula can be derived by Corollary 2. The procedure to estimate parameters of the copula for the conditional cumulative distribution function can be summarized by:

Step 1 Uses the empirical CDF to transform the observations to uniform distribution data in [0, 1],

Step 2 The parameters θ_{ij}, θ_{jk} of the joint CDF's $F(X_i, X_j)$ and $F(X_j, X_k)$ are estimated by the IFM method by Joe (1997),

Step 3 The conditional CDFs $F(X_i|X_j; \hat{\theta}_{ij})$ and $F(X_k|X_j; \hat{\theta}_{jk})$ are computed with the estimates $\hat{\theta}_{ij}$ and $\hat{\theta}_{jk}$,

Step 4 The parameter $\theta_{ik|j}$ of the CDF's $C(F(X_i|X_j), F(X_k|X_j))$ are estimated by the IFM method by Joe (1997),

Step 5 Compute $u_{i|jk}$ with u_i, u_j and u_k by using Corollary 2.

By using this procedure, we obtain the copula-transformed conditional values which we plug into CUSUM statistic for detecting the change of variance.

10.3 Simulation Study

The datasets in the area of finance usually do not follow the usual assumption of constant variance. It is important to see how much the directional dependence can affect the detection of the change of variance. So, we employ the Kim and Hwang (2017) copula directional dependence for detecting the change of variance by CUSUM statistic. By Corollary 2, we generate the copula-transformed conditional values with simulation dataset and then use Monte Carlo method to compare the efficiencies of the different copula-transformed conditional values with the unconditional values with CUSUM statistic according to the average run length (ARL).

Table 10.2 Directional dependence of Gaussian copula beta regression model

Directional dependence	$Y \to X$	$X \to Y$
$\mathrm{Var}(\mathrm{E}(X\|Y))$	0.0495	0.0535
$\mathrm{Var}(X)$	0.0833	0.0833
$\rho^2_{Y\to X}$	0.5937	**0.6421**

Table 10.3 ARLs with asymmetric copula-simulated data with $\rho = 0.8$ where S is the number of simulations and n is the size of data out of 8,888 simulated data

Y|X	S = 1000			X|Y	S = 1000			Estimate
	n = 300	n = 700	n = 1000		n = 300	n = 700	n = 1000	
CUSUM (Y)	150.83	356.77	497.93	CUSUM (X)	151.87	351.16	491.96	
Clayton-CUSUM	148.64	353.17	494.25	Clayton-CUSUM	149.41	349.02	501.65	2.174
Frank-CUSUM	149.90	347.01	495.39	Frank-CUSUM	149.90	347.01	495.39	8.740
Gumbel-CUSUM	150.01	361.00	501.25	Gumbel-CUSUM	146.15	344.35	502.23	2.582
t-CUSUM	145.56	352.18	491.42	t-CUSUM	151.60	348.61	499.55	(0.831, 3.176)

For the simulation study, we use an asymmetric copula (Frank(5) $\times$ Gumbel(30)) simulated data with $\rho = 0.8$ with the verified directional dependence (see Kim and Hwang (2017)). Table 10.2 shows that the directional dependence of Y given X is higher than the directional dependence of X given Y. For the calculation of the ARLs, we collect the sample datasets (n = 300, 700 and 1000) one thousand times (S = 1000) from total dataset size 8,888. Table 10.3 shows that the ARLs of copula conditional statistic of CUSUM in case of the Y given X and $n = 1000$ are more efficient than the ARLs of unconditional statistic of CUSUM but the ARLs of copula conditional statistics (CUSUM) in case of the X given Y and $n = 1000$ are not more efficient than the ARLs of the unconditional statistic of CUSUM. We can see that there exists a directional dependence effect to detect the change of variance. The values of the ARLs when we apply Clayton copula, Frank tail dependence (no tail dependence) t-copula (symmetric tail dependence)-based CUSUM to the simulated data in case of the Y given X, are more efficient than the ARLs of Gumbel-copula statistics of CUSUM.

From this observation, we can conclude that we need to consider the directional dependence by copula for detecting the change of variance of a bivariate financial data.

10.4 Conclusion

Simulation study with the bivariate data proves that our new method for detecting the change of variance by using conditional distribution with diverse copula functions is more efficient than the current available unconditional CUSUM statistic when we take the directional dependence effect into consideration. It will be also more applicable to the multivariate conditional directional dependence with several conditional variables in quantitative finance. In a future study, we will construct the directional dependence of high-order moments by using the new conditional control charts.

References

Busababodhin, P., & Amphanthong, P. (2016). Copula modelling for multivariate statistical process control: A review. *Communications for Statistical Applications and Methods*, *23*(6), 497–515.

Charnes, J. M., & Gitlow, H. (1995). Using control charts to corroborate bribery in Jai Alai. *The American Statistician*, *49*, 386–389.

Clark, T., & Clark, A. (1997). Continuous improvement on the free throw line. *Quality Progress*, *30*, 78–80.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*(1), 141–151.

Demarta, S., & McNeil, A. J. (2005). The t copula and related copulas. *International statistical review*, *73*(1), 111–129.

Dokouhaki, P., & Noorossana, R. (2013). A copula Markov CUSUM chart for monitoring the bivariate auto-correlated binary observations. *Quality and Reliability Engineering International*, *29*(6), 911–919.

Emura, T., Long, T.-H., & Sun, L.-H. (2015). R routines for performing estimation and statistical process control under copula-based time series models. *Communications in Statistics—Simulation and Computation*; In Press.

Guolo, A., & Varin, C. (2014). Beta regression for time series analysis of bounded data, with application to Canada Google flu trends. *The Annals of Applied Statistics*, *8*(1), 74–88.

Inclán, C., & Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, *89*(427), 913–923.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Kim, J.-M., & Hwang, S. (2017). Directional dependence via gaussian copula beta regression model with asymmetric GARCH Marginals. *Communications in Statistics: Simulation and Computation, 46*(10), 7639–7653.

Lahman-Schalem, S., Haimovitch, N., Shauly, E., & Daniel, R. (2002). MBPCA for fault detection in NMOS fabrication. *IEEE Transactions on Semiconductor Manufacturing*, *15*, 60–69.

Long, T.-H., & Emura, T. (2014). A control chart using copula-based markov chain models. *Journal of the Chinese Statistical Association*, *52*, 466–496.

Masarotto, G., & Varin, C. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics*, *6*, 1517–1549.

Mergen, E., Grant, D., & Widrick, M. (2000). Quality management applied to higher education. *Total Quality Management*, *11*, 345–352.

Montgomery, D. (1996). *Introduction to statistical quality control*. New York: Wiley.

Page, E. (1954). Continuous inspection scheme. *Biometrika*, *41*, 100–115.

Sellick, J. J. (1993). The use of statistical process control charts in hospital epidemiology. *Infection Control and Hospital Epidemiology*, *14*, 649–656.

Sklar, A. (1959). Fonctions de repartition á n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris*, *8*, 229–231.

Theodossiou, P. (1993). Predicting the shifts in the mean of a multivariate time series process: An application to predicting business failures. *Journal of the American Statistical Association*, *88*, 441–449.

Verdier, G. (2013). Application of copulas to multivariate charts. *Journal of Statistical Planning and Inference*, *143*, 2151–2159.

Chapter 11
Clustering Methods for Spherical Data: An Overview and a New Generalization

Sungsu Kim and Ashis SenGupta

Abstract Recent advances in data acquisition technologies have led to massive amount of data collected routinely in information sciences and technology, as well as engineering sciences. In this big data era, a clustering analysis is a fundamental and crucial step in an attempt to explore structures and patterns in massive data sets, where clustering objects (data) are represented as vectors. Often such high-dimensional vectors are L_2 normalized so that they lie on the surface of unit hypersphere, transforming them into spherical data. Thus, clustering such data is equivalent to grouping spherical data, where either cosine similarity or correlation is a desired metric to identify similar observations, rather than Euclidean similarity metrics. In this chapter, an overview of different clustering methods for spherical data in the literature is presented. A model-based generalization for asymmetric spherical data is also introduced.

11.1 Introduction

A cluster analysis refers to finding of natural groups (clusters) from a data set, when little or nothing is known about the category structure. A cluster analysis divides data into groups (clusters) that are meaningful, useful, or both. A data clustering belongs to the core methods of data mining, in which one focuses on large data sets with unknown underlying structure. One can broadly categorize clustering approaches to be either model-based (parametric) or distance-based (nonparametric or prototype-based).

In distance-based methods, a cluster is an aggregation of (data) objects in a multi-dimensional space such that objects in a cluster are more similar to each other than to

S. Kim (✉)
Department of Mathematics, University of Louisiana, Lafayette, USA
e-mail: dr.sungsu@gmail.com

A. SenGupta
Applied Statistics Unit, Indian Statistical Institute, Kolkata, India
e-mail: amsseng@gmail.com

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_11

objects in other clusters, and the choice of a distance measure between two objects, called similarity (or dissimilarity) measure, is one of the key issues. Most distance-based methods for linear data are based on the K-means method, fuzzy C-means clustering algorithm, which are called flat partitioning, or hierarchical method (Johnson and Wichern 2008). Flat partitioning clustering algorithms have been recognized to be more suitable as opposed to the hierarchical clustering schemes for processing large data sets. In model-based methods, clusters represent different populations existing in a data set. Hence, a mixture model is a way to perform a parametric clustering analysis. Extensive details on mixture models for linear data are given by Everitt and Hand (1981). However, compared to linear data, researches on both distance-based and model-based clustering methods for spherical data are emerging only recently. In the next section, we present an overview of clustering methods for spherical data and an alternative model-based methods for asymmetric spherical data.

11.2 What Is Spherical Clustering?

In this big data era, a clustering analysis is a fundamental and crucial step in an attempt to explore structures and patterns in massive data sets, where clustering objects (data) are represented as vectors. Often such high-dimensional vectors are L_2 normalized so that they lie on the surface of unit hypersphere, transforming them into spherical data. In spherical (directional) clustering (i.e., clustering of spherical data), a set of data vectors is partitioned into groups, where the distance used to group the vectors is the angle between them. That is, data vectors are grouped depending on the direction into which they point, but the overall vector length does not influence the clustering result. The goal of spherical clustering is thus to find a partition in which clusters are made up of vectors that roughly point in the same direction. For distance-based methods, cosine similarity, instead of Euclidean distance, is mostly used, which measures the cosine of an angle formed by two vectors. For model-based methods, popular mixture models such as a mixture of multivariate Gaussian distributions are inadequate, and the use of a spherical distribution in a mixture model is required.

11.2.1 Applications of Spherical Clustering

Two main applications of spherical clustering are found in text mining and gene expression analysis. In document clustering (or text mining), text documents are grouped based on their features, often described in frequencies (counts) of words, after removing stop words and word stemming operation. Using words as features, text documents are often represented as high-dimensional and sparse vectors, a few thousands dimensions, and a sparsity of 95–99% is typical. In order to remove the

biases induced by different lengths of documents, the data are normalized to have the unit length, ignoring overall lengths of documents. In other words, documents with a similar composition but different lengths will be grouped together (Dhillon and Modha 2001).

Gene expression profile data are usually represented by a matrix of expression levels, with rows corresponding to genes and columns to conditions, experiments or time points. Each row vector is the expression pattern of a particular gene across all the conditions. Since the goal of gene expression clustering is to detect groups of genes that exhibit similar expression patterns, the data are standardized so that genes have mean zero and variance 1, removing the effect of magnitude of expression level (Banerjee et al. 2005).

As a special case of spherical clustering, the spatial clustering is used for agricultural insurance claims and earthquake occurrences (SenGupta 2016). Other applications of spherical clustering found in the literature include:

- fMRI, white matter supervoxel segmentation and brain imaging in biomedicine;
- spatial fading and blind speech segregation in signal processing;
- exoplanet data clustering in astrophysics;
- environmental pollution data in environmental sciences.

11.2.2 Distance-Based Methods

11.2.2.1 Similarity Measures in Spherical Clustering

Cosine similarity measure quantifies similarity between two spherical objects as the cosine of the angle between vectors. Cosine similarity measure is one of the most popular similarity measures performed in spherical clustering applications. Cosine similarity measure is nonnegative and bounded between [0, 1], and Pearson correlation is exactly cosine similarity measure when data are standardized to have mean zero and variance 1.

The Jaccard coefficient, which is sometimes referred to as the Tanimoto coefficient, measures similarity as the intersection divided by the union of two objects and ranges between 0 and 1. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms.

In information theory-based clustering, a vector is considered as a probability distribution of elements, and similarity of two vectors is measured as the distance between two corresponding probability distributions. The Kullback Leibler divergence (KL divergence), also called the relative entropy, is a widely applied measure for evaluating the difference between two probability distributions. However, unlike the previous similarity measures, the KL divergence is not symmetric; as a result, the averaged KL divergence is used in the literature.

11.2.2.2 Spherical K-means and Fuzzy C-direction Algorithms

When data lie on the unit circle, the circular distance between two objects is given by $\cos(\alpha_1 - \alpha_2)$, where α_1 and α_2 are corresponding angles (Jammalamadaka and SenGupta 2001). Generalizing the circular distance to unit hypersphere, cosine similarity between two unit vectors, say y_1 and y_2, is defined to be the inner product of y_1 and y_2, denoted $< y_1, y_2 >$. Suppose n spherical data points are subject to a classification into K groups. Spherical K-means algorithm minimizes $\sum_{k=1}^{K} \sum_{i=1}^{n} \mu_{ki} < y_i, p_k >$, where $\mu_{ki} = 1$ if y_i belongs to cluster k (and otherwise $\mu_{ki} = 0$), and p_k denotes a prototype (cluster center) vector for cluster k. The optimization process consists of alternating updates of the memberships and the cluster centers. Given a set of data objects and a pre-specified number of clusters K, K clusters are initialized. each one being the centroid of a cluster. The remaining objects are then assigned to the cluster represented by the nearest or most similar centroid, which is an updating membership step. Next, new centroids are re-computed for each cluster and in turn all data objects are re-assigned based on the new centroids, which is an updating cluster center step. These two steps iterate until a converged and fixed solution is reached, where all data objects remain in the same cluster after an update of centroids (Hornik et al. 2012).

It is known that for complex data sets containing overlapping clusters, fuzzy partitions model the data better than their crisp counterparts. In fuzzy C-means clustering algorithm for spherical data, each data point belongs to more than one cluster with a different membership value (Kesemen et al. 2016). Fuzzy C-means algorithm for spherical data uses the following criterion

$$\min_{B,M} \sum_{k=1}^{K} \sum_{i=1}^{n} v_{ki}^{m} d_{ki}^{2}, \tag{11.1}$$

where M is a matrix of fuzzy memberships denoted by v_{ki} and $m > 1$, B is a matrix with centroid column vectors, and d_{ki} denotes a similarity measure between object i and centroid k.

11.2.2.3 Issues Related to Distance-Based Methods

- The number of clusters needs to be provided.
- Different initialization of K clusters can produce difference clustering results.
- Convergence is local, and the globally optimal solution cannot be guaranteed. Though, fuzzy C-means algorithm is less prone to local or sub-optimal solutions.
- Convergence is relatively slow in high dimension.

11.2.3 Model-Based Methods: Mixture Models

Suppose a data set consists of n spherical objects (i.e., unit vectors) $\{y_1, y_2, \ldots, y_n\} \in S^{p-1}$ that one wants to divide into K homogeneous groups. Denoting by $g(\cdot)$ a probability density function of Y, the mixture model is

$$g(y) = \sum_{k=1}^{K} \pi_k f(y; \theta_k), \tag{11.2}$$

where π_k (with $\sum_{k=1}^{K} \pi_k = 1$), and f and θ_k represent mixing proportion, spherical density function, and parameter vector of kth mixture component, respectively. Inferences of a mixture model cannot be directly done through the maximization of the likelihood since group labels $\{z_1, z_2, \ldots, z_n\}$ of n objects are unknown. The set of pairs $\{(y_i, z_i)\}_{i=1}^{n}$ is usually referred to as the complete data set. The E-M algorithm iteratively maximizes the conditional expectation of the complete log-likelihood, beginning with initial values of $\theta^{(0)}$. Each expectation (E) step computes the expectation of the complete log-likelihood conditionally to the current value of $\theta^{(q)}$. Then, the maximization (M) step maximizes the expectation of the complete log-likelihood over $\theta^{(q)}$ to provide an update for θ, i.e., $\theta^{(q+1)}$. Computations with high-dimensional or large number of components can be quite demanding. In such cases, Bayesian approaches can lead to significant computational savings and have been quite popular.

11.2.3.1 Mixture of von Mises-Fisher Distributions

The most widely used mixture model is a mixture of von Mises-Fisher (vMF) distributions. The probability density function of vMF distribution is defined by

$$f(y|\mu, \kappa) = c_d(\kappa) e^{\kappa \mu' y}, \tag{11.3}$$

where μ is a mean vector, κ is a concentration parameter around μ, and c_d denotes the normalizing constant. It is not possible to directly estimate κ value in high-dimensional data, and an asymptotic approximation is used. vMF distribution is unimodal and symmetric with circular contours. Various contour shapes of vMF distribution are shown in Fig. 11.1.

11.2.3.2 Score Matching Algorithm

While the E-M algorithm is most widely being used in a mixture modeling of spherical clustering, it requires an approximation of the normalizing constant of a spherical probability distribution, for example, κ in case of vMF distribution. Alternatively,

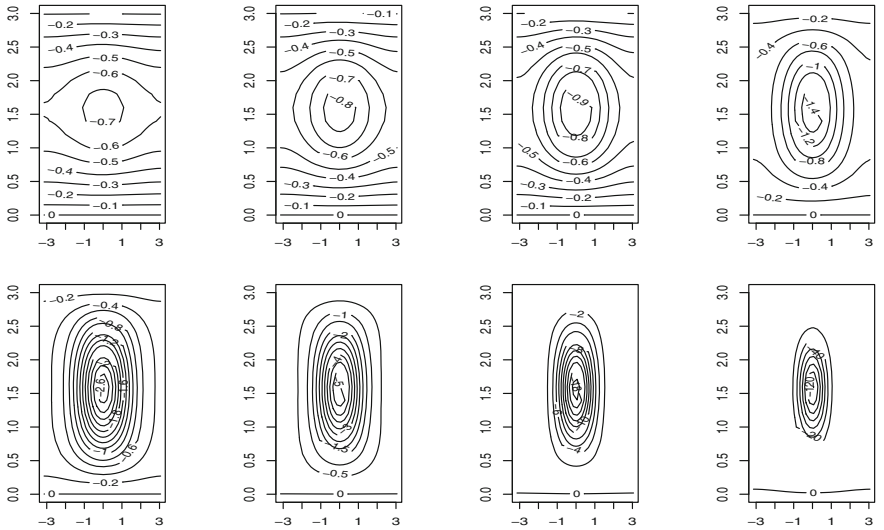


Fig. 11.1 Contour plots of von Mises-Fisher distribution

score matching algorithm (Rosenbaum and Rubin 1983) can be employed, which does not require any knowledge about a normalizing constant. Let $f(y;\pi)$ form a canonical exponential family on a compact-oriented Riemannian manifold with its density proportional to $\exp\{\pi' t(y)\}$, where π and $t(y)$ denote vectors of natural parameters and sufficient statistics, respectively. Then

$$\hat{\pi} = W_n^{-1} d_n, \tag{11.4}$$

where W_n and d_n are sample averages over n data points of $w_{ab} = E\{< t_a, t_b > (y)\}$ and $d_c = -E\{\Delta_M t_c(y)\}$, where $<,>$ is the gradient inner product and Δ_M is the Laplace–Beltrami operator.

11.2.3.3 Connection Between Spherical K-means and Mixture of von Mises-Fisher Distributions

Suppose the concentration parameters of all components in a mixture of von Mises-Fisher distributions are equal and infinite, and mixing proportions ($\pi_k' s$) are all equal as well. Under these assumptions, the E-step reduces to assigning a data point to its nearest cluster, where nearness is computed as cosine similarity between the point and cluster representatives. Hence, spherical K-means is a special case of the vMF mixture model.

11.2.3.4 Issues Related to Model-Based Method

Some of the issues related to model-based methods include:

- curse of dimensionality;
- over-parameterizations in high dimension;
- observations are small compared to the number of variables;
- goodness-of-fit test is not available;
- sensitive to initial values for θ, which are usually given by a partitioning method such as spherical K-means;
- vMF distribution is not suitable if shapes of clusters are not circular symmetric.

11.3 Alternative Model-Based Method: Spherical Generalization of Asymmetric Circular Distributions

In this section, alternative spherical probability models are discussed, which are suitable to model non-symmetric cluster shapes.

The probability density function of Kent distribution (Kent 1982) is defined by

$$f(y|\zeta, \kappa) = C_\kappa \exp\left(\kappa(\zeta_1' y) + \beta\left[(\zeta_2' y)^2 - (\zeta_3' y)^2\right]\right), \tag{11.5}$$

where $\zeta_1, \zeta_2, \zeta_3$ are mean direction, major axis, and minor axis vectors, respectively, κ, β are shape parameters, and C_κ denotes the normalizing constant. The density has ellipse-like contours of constant probability density on the spherical surface. For a high dimension, maximum likelihood estimation is problematic and moment estimators are available (Peel et al. 2001).

By construction, the mixture of the Inverse Stereographic Projection of Multivariate Normal Distribution has the isodensity lines that are inverse stereographic mappings of ellipsoids, which allows asymmetric contour shapes. The necessary and sufficient condition for the density being unimodal is that the greatest eigenvalue of the variance–covariance matrix is smaller than $\frac{1}{2(p-1)}$, where p denotes the dimension of a multivariate normal distribution used in the projection. There is no closed form solution for μ_{MLE} (Dortet-Bernadet and Wicker 2008).

While mixture models using Kent distribution or inverse stereographic projection of normal distributions are suitable for elongated clusters in the data, using their elliptic contours, they will not perform well with clusters having shifted centers nor non-convex clusters. Spherical generalizations of two asymmetric circular distributions found in the following sections provide more flexible model-based spherical clustering.

11.3.1 Spherical Generalization of GvM

When data lie in the unit circle, the generalized von Mises (GvM) density is given by

$$f(\theta) = \frac{\exp(\kappa_1 \cos(\theta - \mu_1) + \kappa_2 \cos 2(\theta - \mu_2))}{\int_{-\pi}^{\pi} \exp(\kappa_1 \cos(\theta - \mu_1) + \kappa_2 \cos 2(\theta - \mu_2)) d\theta}, \tag{11.6}$$

where $\mu_1, \mu_2 \in (-\pi, \pi]$ are location parameters, and $\kappa_1, \kappa_2 > 0$ are shape parameters. GvM distribution is suitable for modeling asymmetric and bimodal circular data, and an extended model of the von Mises (vM) distribution.

A spherical generalization of GvM distribution has the density given by

$$f(y|\zeta, \kappa) = C_\kappa \exp\left(\kappa(\zeta_1' y) + \beta\left[(\zeta_2' y)^2 - (\zeta_3' y)^2\right]\right), \tag{11.7}$$

where ζ's are orientation vectors, κ, β are shape parameters, and C_κ denotes the normalizing constant.

Various contour shapes shown in Fig. 11.2 suggest that a mixture model based on spherical generalization of GvM distribution is appropriate for non-convex symmetric or asymmetric cluster shapes, as well as circular or elliptic symmetric cluster shapes. The Kent distribution is a special case of (11.7), where ζ_1, ζ_2, and ζ_3 are constrained to be orthogonal.

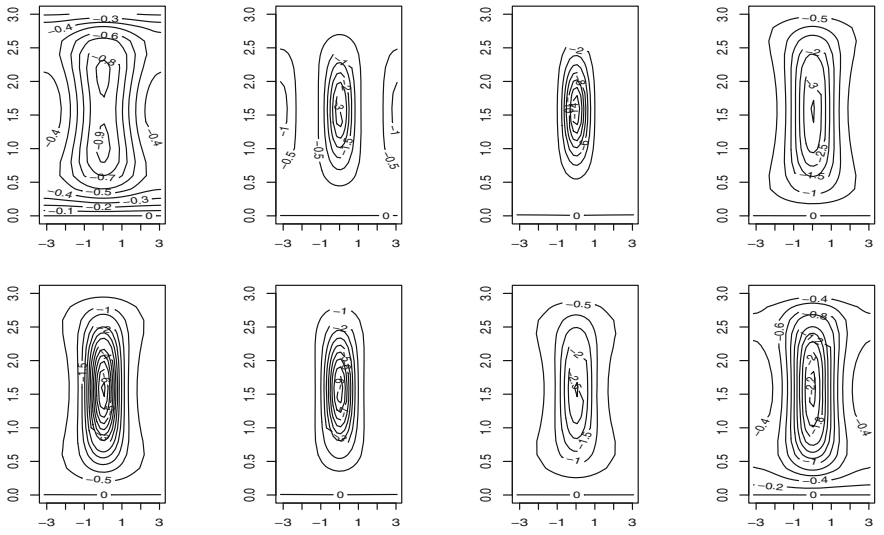


Fig. 11.2 Contour plots of spherical *GvM* distribution

11.3.2 Spherical Generalization of GvM3

The three-parameter generalized von Mises(GvM_3) density (Kim and SenGupta 2012) is given by

$$f(\theta) = \frac{\exp(\kappa_1 \cos(\theta - \mu) + \kappa_2 \sin 2(\theta - \mu))}{\int_{-\pi}^{\pi} \exp(\kappa_1 \cos(\theta - \mu) + \kappa_2 \sin 2(\theta - \mu)) d\theta}, \tag{11.8}$$

where $\mu \in (-\pi, \pi]$ is a location parameter, and $\kappa_1 > 0$ and $\kappa_2 \in [-1, 1]$ are concentration and skewness parameters, respectively. GvM_3 distribution has an advantage over GvM distribution with one less parameter and easier interpretation of the parameters.

A spherical generalization of GvM_3 has the density given by

$$f(y|\zeta, \kappa) = C_\kappa \exp\left(\kappa(\zeta_1' y) + \beta\left[(\zeta_2' y)(\zeta_3' y)\right]\right), \tag{11.9}$$

where ζ's are orientation vectors, κ, β are shape parameters, and C_κ denotes the normalizing constant.

Various contour shapes shown in Fig. 11.3 suggest that a mixture model based on spherical generalization of GvM_3 distribution is appropriate for clusters with shifted centers or clusters with a daughter cluster, as well as symmetric cluster shapes.

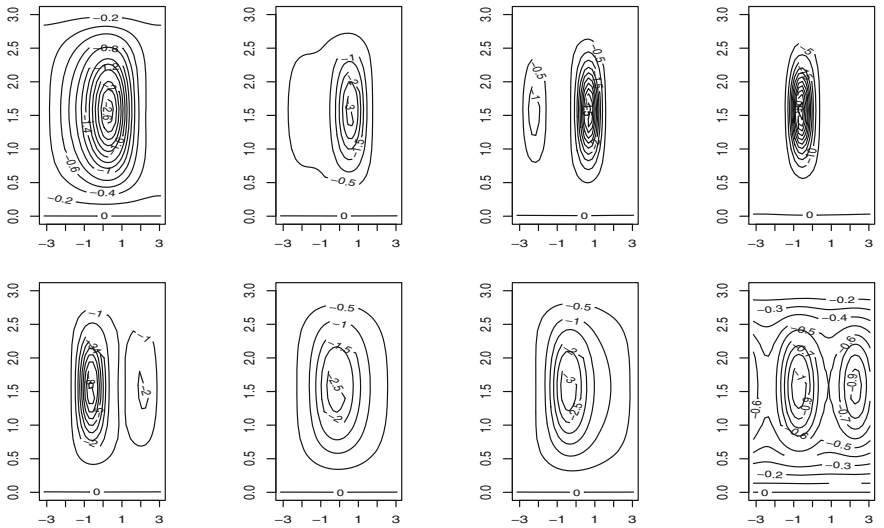


Fig. 11.3 Contour plots of spherical GvM_3 distribution

11.4 Concluding Remarks

In this chapter, an overview of spherical clustering was presented, and more flexible alternative model-based methods were discussed. The authors suggest our readers to consider the alternative model-based methods found in this chapter when cluster shapes in the data set seem to arise from populations which have neither circular nor elliptic contours. On the other hand, it is possible to consider more flexible alternative distance-based methods for asymmetric cluster shapes by developing suitable similarity measures.

References

Banerjee, A., Dhillon, I. S., Ghosh, J., & Sra, S. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, *6*, 1345–1382.

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text using clustering. *Machine Learning*, *42*, 143–175.

Dortet-Bernadet, J.-N., & Wicker, N. (2008). Model-based clustering on the unit sphere with an illustration using gene expression profiles. *Biostatistics*, *9*, 66–80.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman and Hall.

Hornik, K., Feinerer, I., Kober, M., & Buchta, C. (2012). Spherical K-means clustering. *Journal of Statistical Software*, *50*, 1–22.

Jammalamadaka, S., & SenGupta, A. (2001). *Topics in circular statistics*. Singapore: World Scientific.

Johnson, R. A., & Wichern, D. W. (2008). *Applied multivariate statistical analysis*. New York: Pearson.

Kent, J. T. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society Series B*, *44*, 71–80.

Kesemen, O., Tezel, Ö., & Özkul, E. (2016). Fuzzy c-means clustering algorithm for directional data (FCM4DD). *Expert Systems with Applications*, *58*, 76–82.

Kim, S., & SenGupta, A. (2012). A three-parameter generalized von Mises distribution. *Statistical Papers*, *54*, 685–693.

Peel, D., Whiten, W. J., & McLachlan, G. J. (2001). Fitting mixtures of Kent distributions to aid in joint set identification. *Journal of the American Statistical Association*, *96*, 56–63.

Rosenbaum, P. R., Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55

SenGupta, A. (2016). *High volatility, multimodal distributions and directional statistics*. Special Invited Paper, Platinum Jubilee International Conference on Applicaitons of Statistics, Calcutta University, 21–23 Dec

Chapter 12
A Semiparametric Inverse Gaussian Model and Inference for Survival Data

Sangbum Choi

Abstract This work focuses on a semiparametric analysis of survival and cure rate modeling approach based on a latent failure process. In clinical and epidemiological studies, a Wiener process with drift may represent a patient's health status and a clinical end point occurs when the process first reaches an adverse threshold state. The first hitting time then follows an inverse Gaussian distribution. On the basis of the improper inverse Gaussian distribution, we consider a process-based lifetime model that allows for a positive probability of no event taking place in finite time. Model flexibility is achieved by leaving a transformed time measure for disease progression completely unspecified, and regression structures are incorporated into the model by taking the acceleration factor and the threshold parameter as functions of the covariates. When applied to experiments with a cure fraction, this model is compatible with classical two-mixture or promotion time cure rate models. A case study of stage III soft tissue sarcoma data is used as an illustration.

12.1 Introduction

Survival models with cure rates have been received much attention over the last decade. A commonly used approach to facilitate a cure rate is by assuming that the underlying population is a mixture of subjects with different levels of risk. The population would be divided into two subpopulations such that a patient is either cured with a probability of $1-\phi$ or has a proper latency survival function of $S_0(t)$ with a probability of ϕ, which leads to a two-mixture model with an overall survival function $S_{\text{pop}}(t) = (1-\phi) + \phi S_0(t)$. Alternatively, one may use a promotion time model that specifies $S_{\text{pop}}(t) = \exp\{-\theta F_0(t)\},\ \theta > 0$, where $F_0(t) = 1 - S_0(t)$.

In this work, we propose an alternative cure rate modeling approach based on a latent failure process. Patients under study may experience deteriorating health prior to failure or death. One may think of a stochastic process to characterize the

S. Choi (✉)
Department of Statistics, Korea University, Seoul, South Korea
e-mail: choisang@korea.ac.kr

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_12

deteriorating health of a patient by assuming that the process triggers the event of interest when it first passes a critical threshold. Such process-based models have been well exploited in reliability analysis under the name of "degradation modeling" or an "accelerated life model" to anticipate latent failure times (Doksum and Hoyland 1992; Lee and Whitmore 2006; Aalen et al. 2008). For example, a Wiener process with drift may characterize a patient's health status, resulting in an inverse Gaussian (IG) distribution for a lifetime model.

12.2 Models and Methods

Let $A(t)$ be a nondecreasing function. Consider a Wiener process $W(t)$ with drift coefficient $-\mu A(t)$ and variance parameter $A(t)$, for which $W(0) = 0$ and independent increment $\Delta W(t) = W(t + \Delta t) - W(t)$ has a normal distribution with mean $-\mu \Delta A(t) = -\mu\{A(t + \Delta t) - A(t)\}$ and variance $\Delta A(t)$. When the amount of health depreciation reaches a critical level α, failure occurs. Let T denotes the failure time, hence $T = \inf\{t : W(t) \geq \alpha\}$. The monotonicity of $W(t)$ may be achieved by regarding $W(t)$ as $\max_{s \leq t} W(s)$, as their first times are the same. If $\mu \leq 0$, $W(t)$ eventually will reach the threshold with probability one. On the other hand, if $\mu > 0$, $W(t)$ tends to drift away from the absorbing boundary and offers a positive probability of avoiding failure

$$P(T > t) = P(W(t) < \alpha) = \mathrm{IG}(A(t); \alpha, \mu), \tag{12.1}$$

where

$$\mathrm{IG}(t; \alpha, \mu) = \Phi\left(\frac{\alpha + \mu t}{\sqrt{t}}\right) - e^{-2\alpha\mu}\Phi\left(-\frac{\alpha - \mu t}{\sqrt{t}}\right), \tag{12.2}$$

and $\Phi(\cdot)$ is the standard normal distribution function. Event time T in (12.1) has a natural decomposition of the failure time $T = \varepsilon T^* + (1 - \varepsilon)\infty$, where $T^* < \infty$ may denote the failure time for a susceptible patient and ε indicates, by a value of 1 or 0, whether the sample patient is susceptible or not. Likewise, distribution function (12.2) can be written as a two-mixture model, $P(T \geq t) = (1 - \phi) + \phi \cdot P(T^* \geq t)$, where $\phi = P(\varepsilon = 1) = e^{-2\alpha\mu}$. That is, the Wiener process may not reach the boundary with probability $1 - \phi$, representing nonsusceptible or "cured" patients. The underlying distribution for T^* can be obtained by conditioning on the ultimate failure and is given by $P(T^* > t) = \mathrm{IG}^*(A(t); \alpha, \mu)$, where

$$\mathrm{IG}^*(t; \alpha, \mu) = \Phi\left(\frac{\alpha - \mu t}{\sqrt{t}}\right) - e^{2\alpha\mu}\Phi\left(-\frac{\alpha + \mu t}{\sqrt{t}}\right).$$

This is known as the inverse Gaussian (IG) distribution with respective mean and variance, α/μ and α/μ^3, which is a proper distribution function for the susceptible or "non-cured" population in the sense that $P(T^* = 0) = 1$ and $P(T^* = \infty) = 0$.

Let $\tilde{T} = \min(T, C)$ and $\delta = I(T \leq C)$, and Z and X be vectors of time-independent covariates related to disease progression. Also define the counting process $N(t) = \delta I(\tilde{T} \leq t)$ and the at-risk process $Y(t) = I(\tilde{T} \geq t)$. The observations then consist of $(\tilde{T}_i, \delta_i, Z_i, X_i)$, for $i = 1, \ldots, n$, which are copies of $(\tilde{T}, \delta, Z, X)$. Following the idea of (12.1), we seek to conduct a semiparametric analysis to assess the effects of Z on T in the presence of nonsusceptible subjects. Specifically, we posit the survival model for the associated failure times

$$S_i(t|Z_i) = P(T_i \geq t|Z_i) = G_\alpha \left[\int_0^t Y_i(s) \exp\{\beta^T Z_i(s)\} dA(s) \right], \tag{12.3}$$

where $G_\alpha(t) = \text{IG}(t; \alpha, 1)$. We need $\mu = 1$ for model identification with A being nonparametric. Let $\Lambda_Z(t) = \int_0^t Y(s) e^{\beta^T Z(s)} dA(s)$. The IG-based model (12.3) is obtained from the level crossing of the degradation threshold at α by the nonhomogeneous process $W(t) = W_0(\Lambda_Z(t))$, where $W_0(t)$ denotes a homogeneous Wiener process with $T_0 = \inf\{t : W_0(t) > \alpha\}$. Under (12.3), the survival function levels off at the tail, leading to a cured fraction

$$P(\varepsilon = 0|Z_i) = \lim_{t \to \infty} S_i(t|Z_i) = 1 - e^{-2\alpha}.$$

Intuitively, for a large value of α, it is more difficult for the failure process to reach the boundary, resulting in a high cure rate, and the corresponding hazard function shows a high hazard rate early in time that decreases toward zero.

We propose a nonparametric maximum likelihood (ML) method for fitting model (12.3). Let $\Omega = (\alpha, \beta, A)$. Also, let us define $\Psi_\alpha = -\log G_\alpha$, $\eta_\alpha = (\partial/\partial t)\Psi_\alpha$, $\psi_\alpha = (\partial/\partial t) \log \eta_\alpha$, and $\varphi_\alpha = (\partial/\partial \alpha) \log \eta_\alpha$. For ease of presentation, we use $\Psi_i(t; \Omega)$ to denote $\Psi_\alpha\{\Lambda_i(t; \beta, A)\}$, where $\Lambda_i(t; \beta, A) = \int_0^t Y_i(s) \exp\{\beta^T Z_i(s)\} dA(s)$. We similarly define $\eta_i(t; \Omega)$, $\psi_i(t; \Omega)$ and $\varphi_i(t; \Omega)$. By the usual counting process and its associated martingale theory, $M_i(t; \Omega) = N_i(t) - \int_0^t d\Psi_i(s; \Omega)$, $(i = 1, ..., n)$, are martingale processes, where $d\Psi_i(t; \Omega) = \exp\{\beta^T Z_i(t)\}\eta_i(t-; \Omega)dA(t)$. Given this specification, we can write the observed log-likelihood function for (12.3) as

$$l(\Omega) = \sum_{i=1}^n \left[\int_0^\tau \{\beta^T Z_i(t)\} dN_i(t) + \int_0^\tau \log\{\eta_i(t-; \Omega)\} dN_i(t) \right.$$
$$\left. + \int_0^\tau \log\{dA(t)\} dN_i(t) - \int_0^\tau Y_i(t) d\Psi_i(t; \Omega) \right]. \tag{12.4}$$

Because the maximum of (12.4) does not exist if $A(\cdot)$ is restricted to be absolutely continuous, we regard A as a nonincreasing step function and maximize (12.4) with respect to $(\alpha, \beta, \{dA\})$ by taking the jump size of A, denoted by dA, to be zero except for the time at which an event occurs, as a discrete function for A leads to the largest contribution to the likelihood.

Taking derivatives of (12.4) with respect to each component of Ω leads to score process:

$$\mathscr{U}(\Omega) \equiv (\mathscr{U}_\alpha, \mathscr{U}_\beta^T, \{\mathscr{U}_{dA}\}^T)^T = \sum_{i=1}^{n} \int_0^\tau \left[\frac{\partial}{\partial \Omega} \log\{d\Psi_i(t;\Omega)\} \right] dM_i(t;\Omega).$$

The nonparametric ML estimator $\hat{\Omega} = (\hat{\alpha}, \hat{\beta}, \hat{A})$ is then defined as the solution to equation $\mathscr{U}(\Omega) = 0$. Further, it follows from martingale theory that the observed information matrix is approximated by

$$\mathscr{I}(\Omega) = \sum_{i=1}^{n} \int_0^\tau \left[\frac{\partial}{\partial \Omega} \log d\Psi_i(t;\Omega) \right]^{\otimes 2} Y_i(t) d\Psi_i(t;\Omega),$$

which can be used for inferences about Ω.

12.3 Data Example: Soft Tissue Sarcoma Data

Patients with a large (>5 cm), deep, high-grade, soft tissue sarcoma (STS) of the extremity are at significant risk for distant tumor recurrence and subsequent sarcoma-related death. Cormier et al. (2004) retrospectively examined a cohort of 674 patients with primary stage III STS who were treated at two cancer centers in USA from 1984 to 1999. The primary treatment for these patients is surgical resection of the tumor. The use of chemotherapy as adjuvant treatment, however, remains controversial: Explanations have been lacking for the many inconsistencies encountered in the literature that describe the effectiveness of chemotherapy on STS.

The data set is characterized by a considerable fraction of survivors. Of the patients who received chemotherapy (and those who did not), 45.5% (39.3%) died of STS, 9.3% (9.2%) died of competing risks, and 45.2% (51.5%) were still alive at the last study time. Here, death as a result of causes other than STS was treated as censored, which was assumed to be independent of the event time. It turns out that the two Kaplan–Meier curves crossed at about 2 years after the initiation of treatment, indicating a non-monotonic effect of chemotherapy on long-term survival as well as a pronounced monotonic short-term effect. The observed effect of chemotherapy seems to benefit patients early on, but disappears or is reversed in the longer term.

To apply the proposed methods, we considered that the acceleration factor includes $Z =$ (chemotherapy, radiation, amputation, pathologic margin, tumor size), and the threshold regression includes $X =$ (1, chemotherapy). The results from these models are presented in Table 12.1. It appears that the two classical models agree well except for the intercepts, which is expected. Since all those semiparametric models contain the same number of parameters, the log-likelihoods may be translated into Akaike information criterion (AIC) scores. Table 12.1 shows that the IG model achieved slightly higher likelihoods than the two classical methods, which favors the proposed model. For the IG model, the coefficients associated with chemotherapy are both negative, implying that treatment with chemotherapy may decelerate disease progression early on, but eventually results in a lower cured proportion. Those effects

Table 12.1 Estimates and standard errors in parentheses from fitting the inverse Gaussian cure rate model, two-mixture cure rate model, and promotion time cure rate model to soft tissue sarcoma data

Model	Covariate	Inverse Gaussian	Classical models	
			Two-mixture	Promotion time
Disease	Chemotherapy (y/n)	−1.251 (1.447)	−0.206 (0.228)	−0.394 (0.385)
Progress	Radiation (y/n)	−0.358 (0.135)	−0.356 (0.161)	−0.398 (0.174)
	Amputation (y/n)	0.564 (0.208)	0.446 (0.245)	0.506 (0.267)
	Pathologic margin (+/−)	0.435 (0.159)	0.428 (0.189)	0.498 (0.205)
	Tumor size (10–15 cm)	0.279 (0.133)	0.529 (0.166)	0.568 (0.178)
	Tumor size ($\geq$15 cm)	0.504 (0.165)	1.001 (0.200)	1.049 (0.218)
Threshold	Intercept	−1.030 (0.612)	0.714 (0.334)	0.135 (0.220)
	Chemotherapy (y/n)	−0.773 (1.433)	0.353 (0.453)	0.255 (0.296)
	Log-likelihood	−1999.2	−2005.4	−2004.9

may be tested via a likelihood-ratio test: P-values from the proposal for testing the short-term and long-term effects of chemotherapy were 0.08 and 0.24, respectively; whereas they were 0.55 and 0.60 for the two-mixture model. In either case, the effect of chemotherapy appears to be insignificant at a 5% confidence level.

12.4 Discussion

In this work, we considered a new process-based lifetime model, based on the concept that the event of interest occurs when the cumulative depreciation of health first crosses a threshold. By considering a Wiener process that drifts away from the threshold with a positive probability of avoiding failure, the inverse Gaussian distribution naturally arises to account for mixed subpopulations that include cured and uncured individuals. This model has parametric parsimony, which does not require additional formats for the cure rate, and the semiparametric method further allows for a plausible and comprehensive modeling structure that is particularly adequate in large follow-up studies. It may be considered as an alternative to the more traditional hazard-based approaches for cure rate modeling. Besides distributional considerations, we employed a threshold regression scheme and developed a novel nonparametric maximum likelihood estimation.

References

Aalen, O., Gjessing, H., & Borgan, Ø. (2008). *Survival and event history analysis: A process point of view*. Springer: New York.

Cormier, J. N., et al. (2004). Cohort analysis of patients with localized high-risk extremity soft tissue sarcoma treated at two cancer centers: Chemotherapy-associated outcomes. *Journal of Clinical Oncology*, *22*, 4567–4574.

Doksum, K. A., & Hoyland, A. (1992). Models for variable-stress accelerated life testing experiments based on Wiener processes and the inverse Gaussian distribution. *Technometrics*, *34*, 74–82.

Lee, M.-L. T., & Whitmore, G. A. (2006). Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. *Statistical Science*, *21*, 501–513.