Nonlinear Profile Data Analysis for System Performance Improvement

by

Kamran Paynabar

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2012

Doctoral Committee:
Professor Jionghua Jin, Chair
Professor Jack Hu
Professor Vijayan Nair
Associate Professor Ji Zhu

I dedicate this dissertation to my parents Shahnaz and Kazem for their endless love, and unlimited supports, endurance and encouragement.

Acknowledgements

I would like to sincerely thank my advisor, Professor Judy Jin for her great guidance, innumerous support and encouragement, and for everything she has taught me throughout the period of my Ph.D. study, without which this dissertation would not have been completed. Because of Professor Jin, my dream of accomplishing my Ph.D. and pursuing my career in academia came true. Special thanks go to other members of the committee Professor Vijayan Nair, Professor Ji Zhu and Professor Jack Hu for their valuable comments, suggestions and guidance during the course of this dissertation.

I would also like to thank Professor Matthew Reed, Professor Richard Hughes, Professor Arthur Yeh, Professor Jianjun Shi, and Professor Massimo Pacella for providing me with the collaboration opportunities and for their great encouragement and support.

My special gratitude also goes to my former advisors in my undergraduate and masters studies, Professor Rassoul Noorossana, Professor Abbas Saghaei, and Professor Mirbahador Arianezhad for teaching me the basics of conducting research and for encouraging me to continue my graduate study.

Finally, I would like to express my extreme gratitude to my parents for their endless love and supports during whole my life, and to my siblings and friends for their continuous encouragement.

Table of Contents

List of Figures

List of Tables

Abstract

The rapid development of distributed sensing and computer technologies has facilitated a wide collection of various nonlinear profiles during system operations, thus resulting in a data-rich environment that provides unprecedented opportunities for improving complex system operations. At the same time, however, it raises new research challenges on data analysis and decision making due to the complex data structures of nonlinear profiles, such as high-dimensional and non-stationary characteristics. In this dissertation, for the purpose of system performance improvement, new methodologies are proposed to effectively model and analyze nonlinear profile data. Specifically, three major research tasks are accomplished. First, the problem of informative sensor and feature selection among massive multi-stream sensing signals is discussed. In this research, a new hierarchical regularization approach called hierarchical non-negative garrote (HNNG) is proposed. At the first level, a group non-negative garrote is developed to select important signals, and at the second level, the individual features within each signal are selected using a modified version of non-negative garrote that can guarantee nice properties for the estimated coefficients. Second, a new methodology has been developed to analyze cyclic nonlinear profile signals for fully characterizing process variations and enhancing the fault diagnosis capability. In the proposed method, both within- and between-profile variations are taken into consideration. In order to accomplish this, a new mixed-effect model integrated with multiscale wavelets analysis

has been developed. Third, the problem of modeling and monitoring of binary survival profiles has been studied. A general Phase I risk-adjusted control chart is being proposed based on a likelihood ratio test derived from a change-point model. Furthermore it is shown that binary survival outcomes depend on not only patients' health conditions prior to surgery, but also other categorical operational covariates, such as different surgeons. The efficacy of the proposed methods in each chapter is validated and demonstrated by Monte-Carlo simulations and real-world case studies.

CHAPTER I
Introduction

1.1 Motivation

The rapid development of distributed sensing and computer technology has facilitated a wide collection of data during system operations, resulting in a temporally and spatially data-rich environment that provides unprecedented opportunities for improving complex system operations in various fields of applications, including manufacturing, healthcare, and emerging energy systems. In practice, sensor measurements are often represented as a function of one or more variables such as time or location, and are known in the relevant literature as "waveform signals", "profiles", "functional data", and "trajectories". Examples of these types of data include the ram force signals used to press valve seats into engine heads in engine head assembly processes (Paynabar and Jin, 2011), post-operation survival profiles of patients in a surgical system (Paynabar et al., 2012), and the body motion trajectories of drivers' movements in vehicle ingress and egress testing for evaluating the comfort of various vehicle designs (Paynabar et al., 2012).

Profile data consist of rich information that can be used for system analysis, monitoring and diagnosis, as well as effective decision-making to improve system performance. However, analysis of these data also raises new research challenges due to the complex structures of profile data, such as high-dimensional and non-stationary

characteristics of signals. These critical problems are to be addressed in this dissertation as follows:

1. Data dimension reduction through the selection of informative waveform signals among massive multistream sensing signals and the extraction of important features from each selected informative signal.
2. Variation modeling and characterization of profiles, which can effectively analyze both within-profile and between-profile variations to enhance the root cause diagnosis of process variations.
3. Parametric risk-adjusted modeling and monitoring of binary survival profiles with categorical operational covariates.

The case studies shown in this dissertation are conducted in different field applications, indicating the generality and broad application opportunities of the proposed methodologies.

1.2 Overview of Dissertation

In this section, the research topics highlighted in the previous section is briefly discussed in the following subsections. For each topic, an overview of research objectives, challenges, and the proposed methodology are provided.

1.2.1 Selection of Informative Sensing Signals and Features via Hierarchical Non-Negative Garrote Method

In many real-world applications, multiple waveform signals are recorded using a distributed sensing system, providing useful information about system performance. However, it is impossible or unaffordable to place sensors at every station of a process or

every element of a system. Moreover, not all of the collected sensing data are equally important with respect to a specific performance requirement. Therefore, it is necessary to develop a systematic method to optimally select important sensing variables and extract useful information for optimal decision-making. One example of multiple sensing signals includes the Electroencephalography (EEG) signals that record the electrical activity of the brain by distributing sensors over different locations of the scalp. EEG signals are exploited to diagnose and/or predict brain disorders such as epilepsy. However, depending on the purpose of study, only a small subset of EEG signals provides useful information for decision making. Alternatively, in another example, the problem of interest is to evaluate vehicle comfort based on the motion trajectories of drivers. For this purpose, the selected representative subjects are tested under various design configurations. During each test trial, the subjects are asked to get into and out of the vehicle, while their corresponding motion trajectories are recorded by the motion sensors mounted at critical positions on the subjects. Figure 1-1(a) shows a sample of the 3D spatial trajectory, and Figure 1-1(b) shows the corresponding profile signals at x, y and z coordinates for two measurement positions on the right and left ankles. After each trial, each tested subject is asked to give a score of his/her comfort feeling on a scale of 1 to 10, with 1 as the lowest and 10 as the highest comfort index. One critical issue in this research is how to develop a model to represent the relationship between the motion trajectories and the comfort score of each design configuration. As not all of motion trajectories are equally important in affecting the comfort score, it is essential to develop a systematic method for identifying the informative trajectory signals.

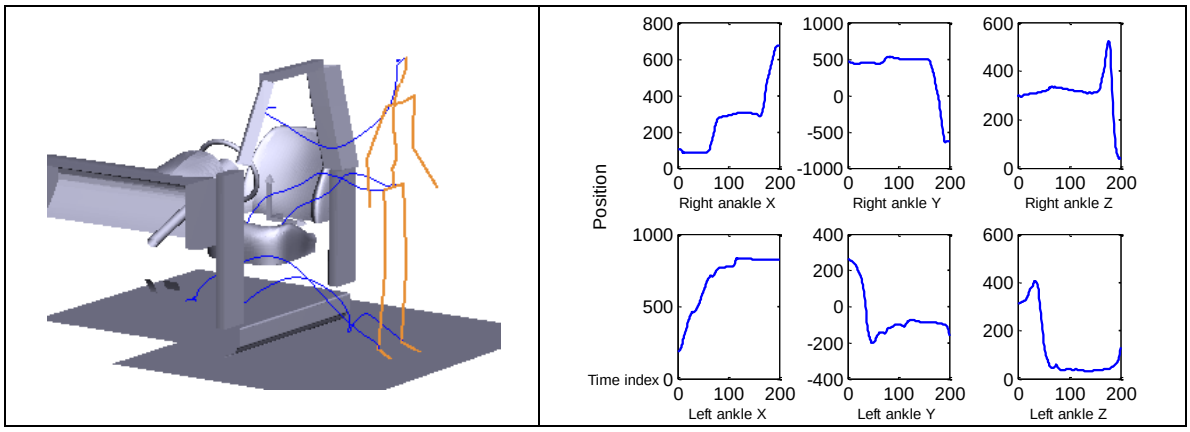

Figure 1-1. Body movement trajectories

Figure 1-1(a) (left panel). A sample of 3D body movement trajectories. Figure 1-1(b) (right panel). A sample of trajectories for right and left ankles in x-, y-, and z-coordinate

Although the problem of selecting important variables has been extensively studied in the literature, our problem is more challenging because we need to select not only important sensors/signals among multi-stream signals, but also a low dimension of interpretable features from the high-dimensional vector of a selected signal. Therefore, the objective of Chapter 2 of the dissertation is to develop a systematic method to effectively extract information from a distributed sensing system by identifying the informative sensors (i.e., selecting the important waveform signals) and selecting important features within the selected waveform signals.

To achieve this objective, we encountered three research challenges. The first one is how to perform variable selection at both between-signal and within-signal levels. The second is a result of the large number of waveform signals that are shown as high-dimensional data vectors. The third is due to the fact that the number of samples may often be less than the dimension of a waveform signal vector.

In order to overcome these challenges, we are proposing a new hierarchical method called “Hierarchical Non-Negative Garrote” (HNNG). HNNG consists of two levels of hierarchy: at the first level, informative sensors/signals are selected using group

non-negative garrote (GNNG) proposed by Yuan and Lin (2007); at the second level, a modified version of non-negative garrote (MNNG) is proposed to select important features within waveform signals selected at the first step. We show that the proposed HNNG benefits from the following useful properties:

- HNNG can effectively perform variable selection at both between- and within-signal levels;
- The solution path for both HNNG criteria is piece-wise linear, and thus can easily be obtained by the least angle regression (Efron et al. 2004);
- HNNG is applicable in cases when the sample size is less than the number of sensors and/or the dimension of waveform signals;
- The estimated coefficients by HNNG have the grouping effect (similar effect) property (Zou and Hastie, 2005), meaning that the estimated coefficients for highly-correlated variables tend to be the same.

We use simulations to evaluate the performance of the proposed HNNG, which are further compared with the existing methods in the literature. Additionally, the analysis of the driver-motion-trajectory data is used as a real case study to illustrate the potential applications and the effectiveness of the HNNG method.

1.2.2 Characterization of Nonlinear Profile Variations Using Nonparametric Mixed-Effect Models and Wavelets

Cyclic waveform profiles with complicated nonlinear structures are widely used for online process monitoring and quality control in automatic manufacturing systems. Chapter 3 of this dissertation is devoted to developing a new methodology for analyzing

cyclic waveform profiles in order to fully characterize process variations and achieve quick fault diagnosis. Extensive research on modeling and analysis of waveform signals for monitoring and fault diagnosis purposes has been done (see for example, Gardner et al., 1997, Williams et al., 2007, Ding et al., 2006, Zou et al., 2008, Zou et al., 2009). However, most of these studies assume that the total variability of profiles can be modeled by random noises, which are mainly used to reflect the within-profile variation that has a constant variance over all measurement points of individual profiles. In many practical situations, however, the variation among in-control profiles is too large to be handled when only using random noises. For example, consider the insertion force of a pressing machine used to press valve seat rings into an engine head. The overlapped multiple samples of signals collected at different cycles of in-control operations are shown in Figure 1-2. As can be seen here, the significant amount of total variation is due to between-profile variations, which cannot be modeled only by random noises. In practice, between-profile variations are usually caused by part-to-part variations, fixture or tooling tolerance, and/or process operation condition variations, while within-profile variations are mainly a result of measurement errors and environmental disturbances. Therefore, in order to effectively monitor the process and identify the sources of variations, it is essential to model and characterize both within-profile and between-profile variations.

In many applications, waveform profile signals are often shown as complicated shapes with local sharp changes and non-differentiable points. Such local information for profiles is very important in both fault detection and diagnosis, and thus must be accurately modeled. Therefore, the objective of Chapter 3 is to develop a modeling

approach for nonlinear profiles for the following two purposes: (1) to characterize both global and local segmental variation patterns; and (2) to characterize nonlinear profile variations by considering both between-profile and within-profile variations.

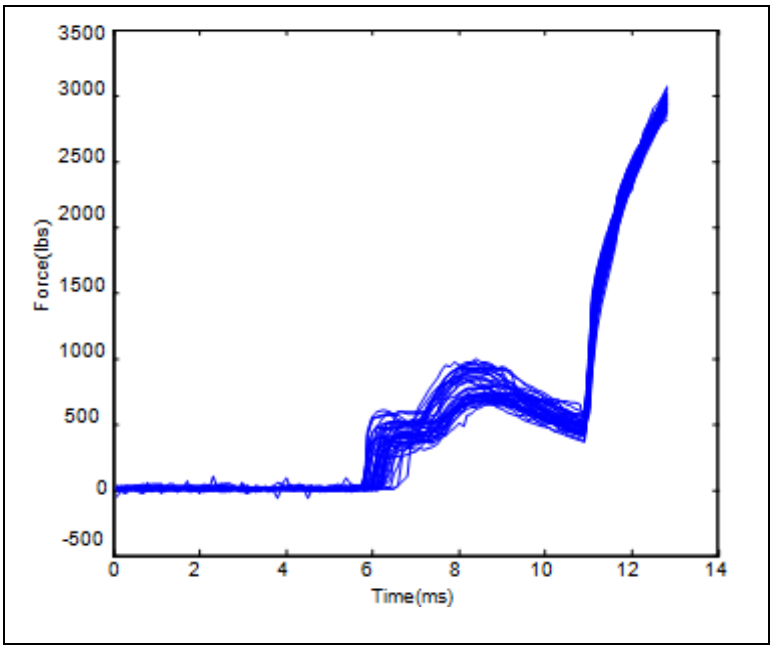

Figure 1-2. Overlapped multiple samples of force in a valve seat assembly operation

Wavelets analysis is a multi-resolution transform that can be used to effectively characterize both global and local segmental variations of the profiles with non-differentiable points. Additionally, mixed-effect models can be used to represent both between- and within-profile variations. In order to take advantage of both the wavelet technique and the mixed-effect models, and to achieve the objectives described previously, we have developed a new wavelet-based mixed effect model.

There are two research challenges in developing and implementing wavelet-based mixed-effect models. The first is to ensure that the collected profile samples used for model estimation follow an identical mixed-effect model distribution because otherwise combining samples from different distributions would lead to a large estimation error for the model parameters, resulting in a misleading model. This challenge is addressed by

applying a change-point model derived from a series of likelihood ratio tests that would be able to detect profile clusters with different model parameters. The second challenge is the computational complexity posed by the large number of wavelet coefficients that are modeled as random effects in the mixed-effect models. We address this issue by a two-step estimation approach that decouples within- and between-profile variations using the wavelet multi-resolution property and estimates each of these separately. The performance of the proposed wavelet-based mixed-effect model is evaluated and compared with other existing methods via a Monte-Carlo simulation. The force profile dataset in the head-engine assembly process is used to illustrate this general methodology and show the performance of the proposed method in a real-world application.

1.2.3 Parametric Risk-adjusted Modeling and Monitoring of Binary Survival Profiles with Categorical Operational Covariates

Survival profiles are also typical profile data that are commonly found in healthcare systems. In Chapter 4 of this dissertation, we develop a new parametric method for a risk-adjusted model to monitor the binary survival profiles with categorical operational covariates. For example, assessing the system performance of surgical operations is vital for improving hospital operations and performance in order to ensure patient safety. However, only tracking the number of successful operations as a metric of surgical performance may be misleading since a surgical outcome depends on not only surgical performance, but also the patients' risk factors before surgery. Therefore, the surgery output, modeled as binary profile data, should be adjusted based on the patients' risk factors. There is an increasing research interest in Phase II risk-adjusted monitoring of binary survival profiles in the literature. For example, Steiner et al. (2000) introduced a

risk-adjusted cumulative sum (RA-CUSUM) chart to monitor the binary survival profiles. Cook et al. (2003) developed a risk-adjusted Shewhart p-chart with variable control limits. Spiegelhalter et al. (2003) proposed resetting a sequential probability ratio test (RSPRT) chart. Grigg and Farewell (2004) proposed a risk-adjustment method to monitor the number of operations between two unsuccessful operations. Grigg and Spiegelhalter (2007) developed a risk-adjusted exponentially weighted moving average (RA-EWMA) chart. In contrast to the existing methods, we show that the binary surgical outcomes depend on not only the patient conditions described by the Parsonnet scores, but also other categorical operational covariates, including different surgeons. The proposed risk-adjustment model should incorporate dummy variables to reflect different surgeon groups' performances. Moreover, all existing research focuses on Phase II monitoring, where it is assumed that the parameters of the risk-adjustment model are known or can be accurately estimated from historical data collected from a stable process. However, this assumption is not valid in real applications, and historical data may follow different distributions due to instability of the surgery system, affecting the accuracy of the estimated parameters in the risk-adjusted model. Therefore, Phase I control is crucial in practice for checking the quality of historical data, and for obtaining accurate estimates of the model's parameters. Constructing risk-adjusted control charts in Phase I is very challenging since each sample represents an individual operation for each patient, making it impossible to fit a risk-adjustment model for each patient based on individual observations. To address this issue, we develop a likelihood ratio test derived from a change-point model (LRT-CP) based on the risk-adjustment logistic regression.

Performance of the proposed method is evaluated and compared with other existing methods through Monte-Carlo simulations. Furthermore, the efficacy of our method in real applications is demonstrated via the analysis of a cardiac surgery dataset.

1.3 Outline of Dissertation

In this dissertation, a generic methodology is developed for the modeling and analysis of nonlinear profiles for the purpose of system monitoring and diagnosis, as well as for effective decision making to improve system performance. Throughout the dissertation the implementation of this methodology is demonstrated using different real-world case studies. The organization of the dissertation associated with the methodology development is shown in Figure 1-3.

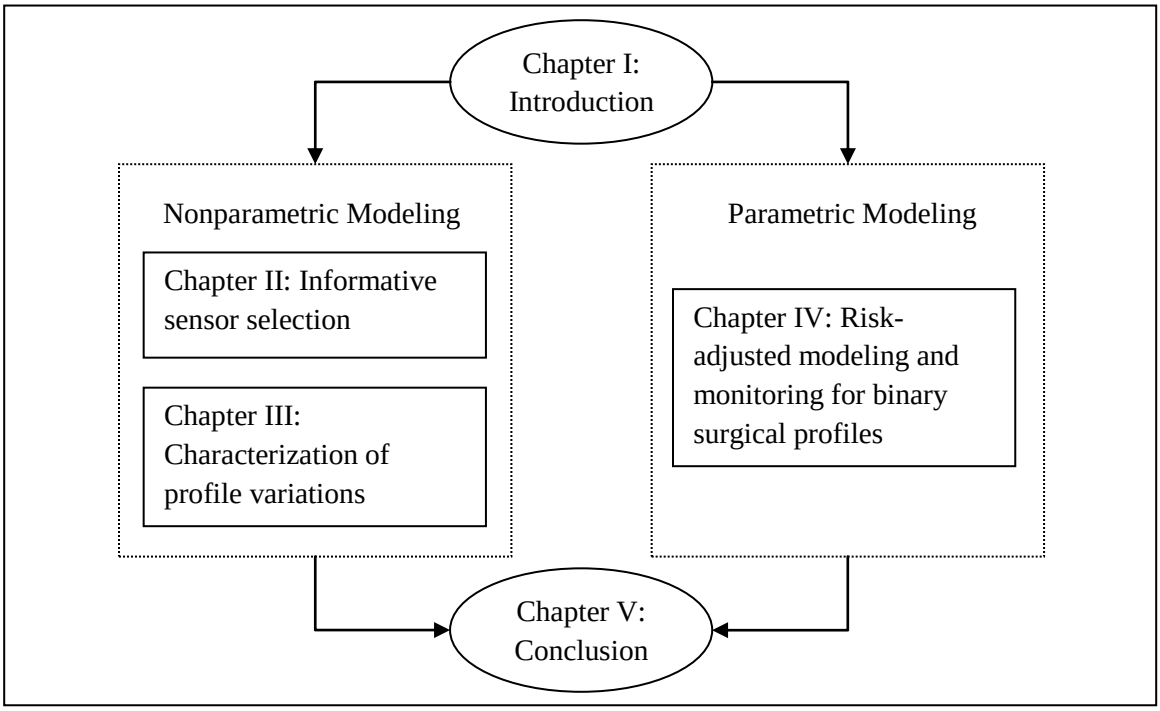

Figure 1-3. Outline of dissertation.

Chapter I presents the key research topics to be discussed in the dissertation. Research motivations and challenges associated with each research topic is also presented in this chapter.

Chapter II proposes a new method for selecting informative sensors and their corresponding waveform signals in a distributed sensing system. The method is also capable of selecting important features within each selected high-dimensional waveform signal. A hierarchical regularization approach called Hierarchical Non-Negative Garrote is used for this purpose.

Chapter III studies variation modeling and characterization of nonlinear profiles with a complex shape for the purpose of process monitoring and fault diagnosis. In this chapter, a novel nonparametric mixed-effect model is developed based on wavelets, which not only characterizes both within- and between-profile variations, but also can extract local information of nonlinear profiles important for root-cause identification.

Chapter IV deals with parametric risk-adjusted modeling and monitoring of binary survival profiles with heterogonous operational covariates. For this purpose, first, the binary survival profiles are modeled using a logistic regression model with dummy variables to model categorical operational covariates. Then, the logistic model is used to construct a likelihood ratio test derived from change-point models for Phase I monitoring.

Chapter V summarizes the major work and new contributions of the dissertation. Besides, some research topics are suggested for future research.

References

1. Cook, D. A., Steiner, S. H., Farewell, V. T. and Morton, A. P. (2003) Monitoring the evolutionary process of quality: Risk adjusted charting to track outcomes in intensive care. *Critical Care Medicine*, **31**, 1676–1682.

2. Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics*, 32, 407–451.

3. Gardner, M., Lu, J., Gyurcsik, R., Wortman, J., Hornung, B., Heinisch, H., Rying, E., Rao, S., Davis, J. and Mozumder, P. (1997) Equipment fault detection using spatial signatures. *IEEE Transaction on Components, Packaging, and Manufacturing Technology, Part C*, **20**, 294-303.

4. Grigg, O. A. and Farewell, V. T. (2004) A risk-adjusted sets method for monitoring adverse medical outcomes. *Statistics in Medicine*, **23**, 1593-1602.

5. Grigg, O., and Spiegelhalter, D. (2007) A simple risk-adjusted exponentially weighted moving average. *Journal of the American Statistical Association*, **102**, 140-152.

6. Paynabar, K., Jin, J. (2011) "Characterization of Nonlinear Profiles Variations using Mixed-effect Models and Wavelets," *IIE Transactions on Quality and Reliability Engineering*, **43**, 275–290.

7. Paynabar, K., Jin, J., and Yeh. B. A. (2011) "Phase I Risk-Adjusted Control Charts for Monitoring Surgical Performance by Considering Categorical Covariates," *Journal of Quality Technology*, **44**, 39-53.

8. Paynabar, K., Jin, J., and M. Reed, (2012) “Hierarchical Non-Negative Garrote for Group Variable Selection”, Technical Report, Department of Industrial and Operation Engineering, The University of Michigan.

9. Spiegelhalter, D. J., Grigg, O. A., Kinsman, R. and Treasure, T. (2003) Sequential probability ratio tests (SPRTS) for monitoring risk-adjusted outcomes. *International Journal for Quality in Health Care*, **15**, 1–7.

10. Steiner, S. H., Cook, R. J., Farewell, V. T., and Treasure T. (2000) Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics*, **1**, 441-452.

11. Williams, J. D., Woodall, W. H., and Birch, J. B. (2007) Phase I analysis of nonlinear product and process quality profiles. *Quality and Reliability Engineering International*, **23**, 925–941.

12. Yuan, M., and Lin, Y. (2006) Model selection and estimation in regression with grouped variable. *J. R. Statist. Soc.B*, **68**, 49-67.

13. Zou, C., Qiu, P., and Hawkins, D. (2009) Nonparametric Control Chart for Monitoring Profiles Using Change Point Formulation and Adaptive Smoothing. *Statistica Sinica*, **19**, 1337-1357.

14. Zou, H., and Hastie, T. (2005) Regularization and variable selection via the elastic net *J. R. Statist. Soc. B*, **67**, 301–320.

CHAPTER II

Selection of Informative Sensing Signals and Features via Hierarchical Non-Negative Garrote Method

2.1 Introduction

Recent advancements in sensing technology and data acquisition systems have facilitated large-scale data collection during system operations. In many real-world applications, multiple waveform signals are recorded using a distributed sensing system, providing useful information about system performance. However, it is impossible or unaffordable to place sensors at every station of a process or every element of a system. Moreover, not all of the collected sensing data are equally important with respect to a specific performance requirement. Therefore, it is necessary to develop a systematic method to optimally select important sensing variables and extract useful information for optimal decision-making. One example of multiple sensing signals includes the Electroencephalography (EEG) signals depicted in Figure 2-1(a). EEG signals represent the electrical activity of brain, which are recorded using several sensors located at different locations of scalp (Figure 2-1(b)). Those signals are exploited to diagnose and/or predict focal brain disorders such as epilepsy, stroke, etc. (See for example Chaovalitwongse et al., 2006). Depending on the purpose of analysis, only a small subset of EEG signals provides useful information for diagnosis and decision making. Alternatively, in another example, the problem of interest is to evaluate vehicle comfort

based on the motion trajectories of drivers. For this purpose, the selected representative subjects are tested under various design configurations. During each test trial, the subjects are asked to get into and out of the vehicle, while their corresponding motion trajectories are recorded by the motion sensors mounted at critical positions on the subjects. The spatial motion trajectories are measured over time in x-,y-, and z-coordinate. Figure 1-1(a) (in Chapter 1) shows a sample of the 3D spatial trajectories for right and left ankles, right and left hips, and head in an egress trial and Figure 1-1(b) shows the trajectories of one egress trial at x, y and z coordinates two measurement positions on the right and left ankles. After each trial, each tested subject is asked to give a score of his/her comfort feeling on a scale of 1 to 10, with 1 as the lowest and 10 as the highest comfort index. One critical issue in this research is how to develop a model to represent the relationship between the motion trajectories and the comfort score of each design configuration. As not all of motion trajectories are equally important in affecting the comfort score, it is essential to develop a systematic method for identifying the informative trajectory signals.

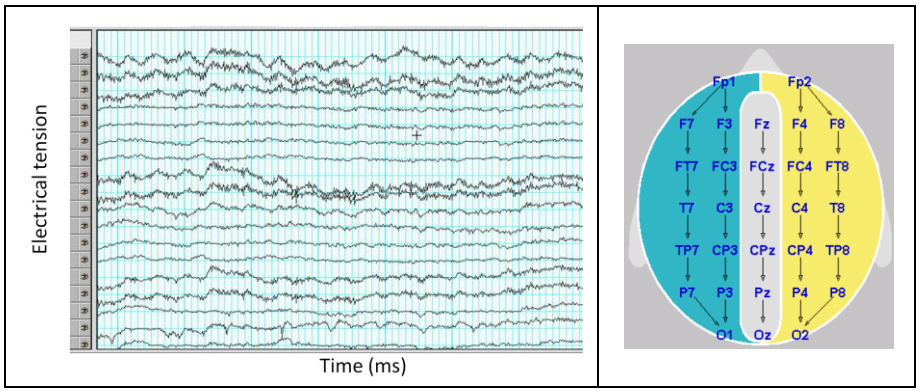

Figure 2-1. Electroencephalography signals
Figure 2-1(a) (left panel). A sample of multichannel EEG signals. Figure 2-1(b) (right panel). Location of sensors on scalp.

The problem of selecting informative sensors can be modeled as a group variable selection problem in which each group variable consists of recorded signals or signal features corresponding to one sensor. In this case, the significance of a group's contribution in the model can reflect the importance of the corresponding sensor. For example, in the ingress/egress experiment, the motion trajectory captured by each individual sensor can be considered as one group of variables. Therefore, the selection of an important sensor can be assessed by the significance of the group corresponding to its motion curve. It should be noted that the application of group variable selection is not limited to sensor selection. Generally, the group relationship can be identified based on either the physical functionality of individual variables or the model structure. For example, in multi-way analysis-of-variance (ANOVA), a factor can be considered as a group of dummy variables, which represent the levels of the factor. Group variables can also be found in additive models, in which each component of the model is represented by a set of polynomials, splines, wavelets and/or other basis functions. Thus, each set of basis functions forms a natural group.

In regression and classification problems, variable selection is a crucial issue since incorrect inclusion of unimportant variables in the model may seriously affect the prediction accuracy of the model. Moreover, from the practical standpoint, the parsimonious models that can be interpreted by the domain knowledge are often preferred. Owing to the importance of this topic, extensive research has been conducted and various methods can be found in the literature on variable selection. For example, best subset selection (Miller 2002) has been popularly used for variable selection. However, due to the discreteness of this method, it may yield unstable modeling results,

i.e., if data are slightly changed, the selected subset of variables may vary, thus leading to a very different model. Regularization techniques, on the other hand, are continuous processes that improve the prediction accuracy and model parsimony through the trade-off of between model's bias and variance. The examples of regularization techniques include Bridge regression (Frank and Freidman 1993), nonnegative garrote (Breiman, 1995), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive LASSO (Zuo, 2006), structured variable selection (Yuan et al., 2009), to name a few.

The aforementioned methods are often employed when there is no group relationship among predictor variables (called predictors henceforth, for simplicity), and thus all predictors are treated individually. However, in the case of sensor selection, the variable selection procedure should also consider the group relationship among individual variables. In the literature, there is extensive research on group variable selection via regularization methods. In order to review these methods, a general regression model with the consideration of the group variable structure is presented as follows:

$$y_i = \beta_0 + \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} x_{i,kj} + \varepsilon_i \,, \tag{2-1}$$

where y_i denotes the response variable, $x_{i,kj}$ is predictor j (j=1,..,p_k), belonging to group k (k=1,..,K), β_0 and β_{kj} are the corresponding model parameters, ε_i is the random error, and index i $(i = 1,...,n)$ indicates the sample number. The general regularization method for estimating the parameters of model (2-1) can be written as

$$\min_{\beta_{kj},\beta_0} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{K}\sum_{j=1}^{p_k}\beta_{kj}x_{i,kj}\right)^2 + J(\lambda_t,\beta_{kj}), \tag{2-2}$$

where $J(\lambda_t,\beta_{kj})$ is a penalty function with tuning parameters $\lambda_t; t=1,2,...,T$. In practice, usually not more than two tuning parameters are used. Before introducing our method, a brief review of various penalty functions $J(\lambda_t,\beta_{kj})$ for the purpose of variable selection is given as follows.

Atoniadis and Fan (2001) developed a group thresholding method for wavelet coefficient shrinkage. Yuan and Lin (2006) extended the regular lasso and developed the group lasso by using an L_2-norm penalty in the form of $J(\lambda_t,\beta_{kj}) = \lambda_1\sum_{k=1}^{K}\left\|\boldsymbol{\beta}_k\right\|$ with $\left\|\boldsymbol{\beta}_k\right\| = \sqrt{\sum_{j=1}^{p_k}\beta_{kj}^2}$. Meier et al. (2008) further modified the group lasso for logistic regression models. Moreover, Zhao et al. (2009) proposed an alternative method using an L_∞-norm penalty, i.e., $J(\lambda_t,\beta_{kj}) = \lambda_1\sum_{k=1}^{K}\left\|\boldsymbol{\beta}_k\right\|_\infty$ with $\left\|\boldsymbol{\beta}_k\right\|_\infty = \max_j\left\{\left|\beta_{kj}\right|\right\}$. Since both $\left\|\boldsymbol{\beta}_k\right\|$ and $\left\|\boldsymbol{\beta}_k\right\|_\infty$ are singular at $\boldsymbol{\beta}_k = \mathbf{0}$, unimportant groups can be removed from the model by choosing an appropriate tuning parameter.

In the case of grouped variables, a desirable variable selection method should consider two levels of variable selection. The first level is at the group level, which is used to judge if a group variable should be included in the model. The second level is at the individual level, by which those non-significant individual variables within the selected important groups are further removed from the model. In the case of sensor

selection, at the first level, the important sensors and their corresponding signals are identified and at the second level, the individual features within each important signal are selected. However, the above-reviewed methods perform variable selection only at the group level. As a result, all variables within the selected group will be included in the model, which may affect both prediction performance and model parsimony. To address this issue, Huang et al. (2009) and Zhou and Zhu (2010) proposed group bridge and hierarchical lasso (Hlasso) by using the following penalty: $J(\lambda_t, \beta_{kj}) = \lambda_1 \sum_{k=1}^{K} \left\| \boldsymbol{\beta}_k \right\|_1$ with $\left\| \boldsymbol{\beta}_k \right\|_1 = \sqrt{\sum_{j=1}^{p_k} \left| \beta_{kj} \right|}$. Both between- and within-group parts of this penalty function (i.e., $\sum_{k=1}^{K} \left\| \boldsymbol{\beta}_k \right\|_1$, and $\sqrt{\sum_{j=1}^{p_k} \left| \beta_{kj} \right|}$) are singular at $\boldsymbol{\beta}_k = \mathbf{0}$ and $\beta_{kj} = 0$, respectively. Thus, this penalty function can perform variable selection at both group and individual levels. However, the penalty function of hierarchical lasso is non-convex function, which makes it difficult to solve the optimization problem. Zhou and Zhu (2010) proposed to break down the optimization criterion into two convex optimization sub-problems and claimed that solving these two problems iteratively can converge to an optimum solution. In addition to expensive computation caused by their proposed iterative algorithm, there is no guarantee that the obtained solution is global optimum or a local optimum. This is true since the penalty function is non-convex and the iterative algorithm may converge to a stationary point which may or may not be a local optimum.

In the regression models with group variables, it often happens that individual variables are highly correlated both within and between groups. If two variables are highly correlated, their corresponding coefficients should tend to be equal (up to a sign

change if negatively correlated). This is called "grouping effects," in the literature. Note that the "grouping effects" is different from the "group variable" structure we described earlier. In the former, the grouping is defined only based on the correlation of predictors, while in the latter grouping is defined based on any common characteristics among one set of variables such as different levels of a factor, measured response from each sensor, etc. To avoid confusion, between "group variable" and "grouping effects", in the dissertation, we use the term *similar effects* to represent "grouping effects". Zou and Hastie (2005) proposed elastic net with a combined L_1- and L_2-norm penalty (i.e., $J(\lambda_t, \beta_{kj}) = \lambda_1 \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\beta_{kj}| + \lambda_1 \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj}^2$) and showed that the elastic net benefits from the *similar effects* property. Bondell and Reich (2007) developed OSCAR (octagonal shrinkage and clustering algorithm for regression) by combining regularization and clustering for the purpose of variable selection and clustering the highly-correlated variables into predictor clusters. Neither of these methods, however, considered group variable selection.

The main goal of this chapter of dissertation is to propose a new group variable selection method for sensor selection that not only is capable of identifying informative sensors and signals, but also can select important features within each selected signal. For this purpose, we define a two-step hierarchical regularization based on NNG. We show that the optimization criteria in both steps are convex and propose an iterative algorithm based on least angle regression (LARS) (Efron et al., 2004) to simply obtain the whole solution path of each coefficient. The proposed method also benefits from the "similar effects" property for highly correlated individual variables. Moreover, another advantage

of the proposed method is that it can be used when the sample size is less than number of important variables, while other methods such as lasso and NNG cannot.

The rest of the chapter is organized as follows. In Section 2.2, we introduce two-step hierarchical non-negative garrote (HNNG) and present an iterative algorithm for solving the optimization criteria and obtain the solution paths of coefficients. We also discuss the useful properties of HNNG regarding *similar effects* and small sample size. The performance of the proposed method is evaluated and compared with other exiting methods using Monte-Carlo simulation in Section 2.3. Section 2.4 is devoted to a real case study, in which the proposed method is applied to vehicle ingress data to select important sensors for predicting the comfort score of a design configuration.

2.2 Proposed Method: Two-step Hierarchical Non-negative Garrote

In this section, we present our two-step hierarchical regularization method based on NNG. The original NNG was proposed by Breiman (1995) as an effective alternative to the subset method for variable selection. Yuan and Lin (2007) studied the properties of NNG in terms of solution path and consistency, and showed that despite lasso, NNG is consistent in identifying important variables. Yuan (2007) applied NNG to component selection in functional ANOVA. Cantoni et al. (2011) used NNG for variable selection in additive models. Yuan and Lin (2006) developed group NNG (GNNG) for group variable selection. However, similar to group lasso, their proposed group NNG is not capable of variable selection within a group. To address this issue, we propose HNNG. Our proposed method consists of two steps, each of which corresponds to one level of the hierarchy. In the first step, we perform variable selection at the group level of the hierarchy, select important groups, and remove unimportant groups from the set of

predictors. This is done by a group NNG (GNNG) model. The second step is devoted to variable selection at the individual level. In this step, we use a modified NNG (MNNG) for individual variable selection and coefficients estimation within important groups selected at the first step. In the following subsections, each step is elaborated.

2.1.1 Step 1: Variable Selection at Group Level

In this step, to select the important group variables, we use GNNG (Yuan and Lin, 2006). Without loss of generality, it is assumed that all predictors are transformed such that $\sum_{i=1}^{n} x_{i,kj} = 0$ and $\sum_{i=1}^{n} x_{i,kj}^2 = 1$, and the response variable is centered such that $\sum_{i=1}^{n} y_i = 0$. Consider the following reparameterization: $\beta_{kj} = d_k \hat{\beta}_{kj}^{\text{ols}}; d_k \geq 0$, where $\hat{\beta}_{kj}^{\text{ols}}$ is the ordinary least square estimate of β_{kj} and d_k is the shrinking factor for group *k*. d_k is considered non-negative to avoid nonidentifiability issue. The following penalized least square criterion is used for GNNG (Yuan and Lin, 2006):

$$\min_{d_k} \frac{1}{2} \left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_k \hat{\boldsymbol{\beta}}_k^{\text{ols}} d_k \right\|^2 + \lambda \sum_{k=1}^{K} d_k, \quad \text{subject to}: d_k \geq 0, \ k = 1,2,...,K, \tag{2-3}$$

where $\mathbf{y}$ is the vector of the observed response variable, $\mathbf{X}_k$ is the matrix of the predictor variables for group *k*, $\hat{\boldsymbol{\beta}}_k^{\text{ols}}$ is the vector of the estimated coefficients using the ordinary least square method corresponding to group *k*, and λ is the tuning parameter. Since $\sum_{k=1}^{K} d_k; (d_k \geq 0)$ is singular at $d_k = 0$, this penalty can effectively identify the unimportant groups.

After optimizing criterion (2-3), if $\hat{d}_k > 0$, we keep the variables of group *k* for the next step, otherwise the corresponding variables are removed from the model. Let *S* denote the set of indices for the selected groups. We define $\tilde{\mathbf{X}}$ as the matrix of all predictor variables whose groups are identified as important, i.e., $\tilde{\mathbf{X}} = [\mathbf{X}_k], k \in S$. The dimensions of matrix $\tilde{\mathbf{X}}$ is *n* by $\sum_{k \in S} p_k$.

2.2.1 Step 2: Variable Selection and Coefficient Estimation at Individual Level

After selecting important groups, this step is used to further identify important individual variables, and estimate their corresponding coefficients within the selected groups. As mentioned earlier, individual variables are often correlated. Therefore, it is preferred to develop a model that has *similar effects* property for highly correlated variables. We modify the NNG (Brieman 1995) as follows such that it gains the *similar effect* property.

Let $\boldsymbol{\beta}$ denote the vector of coefficients of all individual variables. $\boldsymbol{\beta}$ can be reparameterized as $\boldsymbol{\beta} = (\hat{\boldsymbol{\beta}}^{\mathrm{r}} \bullet \mathbf{d}), \mathbf{d} \geq \mathbf{0}$, where $\hat{\boldsymbol{\beta}}^{\mathrm{r}}$ is the vector of ridge estimates calculated by $(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} + \mathbf{I}\lambda_2)^{-1} \tilde{\mathbf{X}}^T \mathbf{y}$, $\mathbf{d}$ is the vector of shrinking factors of individual variables and operator "$\bullet$" represents the element-wise vector product. The proposed MNNG criterion is defined as follows:

$$\min_{\mathbf{d}} \frac{1}{2} \left\| \mathbf{y} - \tilde{\mathbf{X}}(\hat{\boldsymbol{\beta}}^{\mathrm{r}} \bullet \mathbf{d}) \right\|^2 + \lambda_1 \|\mathbf{d}\|_1 + \frac{\lambda_2}{2} \|\mathbf{d}\|^2, \quad \text{subject to}: \mathbf{d} \geq \mathbf{0}, \tag{2-4}$$

where $\lambda_t \geq 0, t = 1,2$ are tuning parameters. If $\lambda_1 = 0$, criterion (2-4) becomes a ridge-type criterion, and if $\lambda_2 = 0$, it becomes an NNG criterion, thus it has the characteristics of both ridge and NNG criteria. Note that the penalty function used in (2-4) can be

considered as the weighted penalty function in the elastic net criterion (Zou and Hastie, 2005) with $\hat{\boldsymbol{\beta}}^{r}$ as the vector of inverse weights. Therefore, similar to elastic net, the MNNG has the *similar effects* property and also can handle the situations where the sample size is less than number of important variables. We discuss these two properties in more details in Subsections 2.2.5 and 2.2.6. Despite Hlasso, which lacks a convex criterion, HNNG uses convex criteria at both group- and individual-levels, thus the global optimum can be easily obtained.

2.2.2 Hierarchical Non-negative Garrote with Orthogonal Predictors

To gain more insight about the proposed two-step HNNG, we study the cases with orthogonal predictors. Orthogonal predictors can be seen in many situations, for example, orthogonal wavelets, orthogonal spline, and/or orthogonal polynomials are often used as the predictors for nonparametric curve fitting. In the case of orthogonal bases, where $\sum_{i=1}^{n} x_{i,kj}^2 = 1$ and $\sum_{i=1}^{n} x_{i,kj} x_{i,k'j'} = 0$ if $k \neq k'$ or $j \neq j'$, both group- and individual-level criteria have closed-form solutions.

Lemma 2-1. *If predictor variables are orthonormal to each other,*

i) the solution of (2-3) in Step 1 is obtained by

$$\hat{d}_k = \left(1 - \lambda \Big/ \sum_{j=1}^{p_k} (\hat{\beta}_{kj}^{\text{ols}})^2\right)_+, \qquad \textit{(2-5)}$$

where $\hat{\beta}_{kj}^{\text{ols}} = \sum_{i=1}^{n} x_{i,kj} y_i \ (k = 1,...,K; j = 1,..., p_k)$, *is the ordinary least square estimator using the orthogonal predictors, and* $(a)_+ = \max(0, a)$.

ii) the solution of (2-4) in Step 2 is obtained by

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^{\text{ols}})\left(\left|\hat{\beta}_j^{\text{ols}}\right| \gamma_1 - \gamma_2\right)_+, \qquad \textit{(2-6)}$$

where $\hat{\beta}_j^{\text{ols}} = \sum_{i=1}^{n} \tilde{x}_{i,j} y_i \,; (j = 1,2,..., \sum_{k \in S} p_k)$ *is the ordinary least square estimator in the orthogonal case, with* $\tilde{x}_{i,j}$ *as the elements of matrix* $\tilde{\mathbf{X}}$,

$$0 \le \gamma_1 = \left(\frac{\left(\hat{\beta}_j^{\text{ols}}\right)^2}{\lambda_2 (1+\lambda_2)^2 + \left(\hat{\beta}_j^{\text{ols}}\right)^2} \right) \le 1 \textit{ and } \gamma_2 = \left(\frac{\lambda_1 (1+\lambda_2)}{\lambda_2 (1+\lambda_2)^2 + \left(\hat{\beta}_j^{\text{ols}}\right)^2} \right) \ge 0 .$$

The proof is given in Appendix 2.A.

It should be noted that both solutions in (2-5) and (2-6) are soft-thresholding estimates. The intuitive interpretation is given as follows. The soft-thresholding estimate in (2-5) shrinks each group effect, and the amount of shrinkage is inversely proportional to $\sum_{j=1}^{p_k} \left(\hat{\beta}_{kj}^{\text{ols}}\right)^2$. Therefore, if the ordinary least square coefficient of a group is small implying the group is not important, it is more likely that corresponding d_k is shrunk to zero. Furthermore, the soft-thresholding estimate in (2-6) shrinks the $\hat{\beta}_j^{\text{ols}}$ coefficients using two different shrinkage factors γ_1 and γ_2. The amount of shrinkage using both of these factors is inversely proportional to $\left(\hat{\beta}_j^{\text{ols}}\right)^2$. If a predictor is significant, we expect that its coefficient is slightly shrunk. In other words, for significant predictors with large $\left(\hat{\beta}_j^{\text{ols}}\right)^2$, γ_1 and γ_2 are close to 1 and 0, respectively, implying that the amount of

shrinkage for this coefficient is small. So, both (2-5) and (2-6) have a rational interpretation.

2.2.3 Computation: Solutions of Hierarchical Non-negative Garrote

To solve the optimization criterion at the group level for GNNG in (2-3), one can apply standard convex optimization method such as interior-points algorithms (Sturm, 1999), the shooting algorithm (Fu, 1999), and/or an efficient solution path algorithm (e.g., Efron et al., 2004, and Yuan and Lin, 2006). We recommend the GNNG solution path algorithm based on LARS proposed by Yuan and Lin (2006) because this algorithm can provide the entire solution path for a wide range of λ with the same order of computations as a single OLS fit.

To solve the optimization criterion at individual level for MNNG in (2-4), we propose an efficient solution path algorithm called LARS-MNNG, which is based on the LARS algorithm (Efron et al., 2004). The original LARS algorithm (Efron et al., 2004) was proposed for individual variable selection and applied to lasso and stagewise regression. This algorithm can be summarized as follows. Starting from all coefficients equal to zero, LARS selects the variable, which has the highest correlation with the response variable and proceeds in this direction. Despite the stepwise regression methods, instead of taking a full step towards the projection of Y on the selected variable, the LARS algorithm takes the largest possible step in this direction until another input variable has as much correlation with the current residuals. In this case, the projection of current residuals on the space spanned by the selected variables has the equal angle with the selected two variables. Then, the LARS algorithm proceeds in this direction until it finds the third variable. This procedure continues until all variables enter to the model.

The fact that the LARS solution path is piecewise linear, significantly reduces the computation complexity of the LARS algorithm. Therefore, in order to calculate the entire solution path, it suffices to find the break points defining the piecewise linear path. Yuan and Lin (2007) showed that the solution path for NNG is piecewise linear and modified the LARS algorithm to obtain the solution path for NNG. We show that, similar to the NNG, the solution path for modified NGG, given λ_2 is piecewise linear and thus the LARS algorithm can be adapted for calculating the solution path.

Lemma 2-2. *Define* $\vec{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$, *and* $\vec{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{X}} \\ \sqrt{\lambda_2 / 2}\mathbf{B} \end{bmatrix}$,*where* $\mathbf{B}$ *is a* $P \times P$ *diagonal matrix whose diagonal elements are equal to* $\left(\hat{\beta}_j^{\mathrm{r}}\right)^{-1}$ *with* $P = \sum_{k \in S} p_k$ *. The MNNG criterion in (2-4), can be reparameterized as*

$$\min_{\mathbf{d}} \frac{1}{2}\left\|\vec{\mathbf{y}} - \vec{\mathbf{X}}(\hat{\boldsymbol{\beta}}^{\mathrm{r}} \bullet \mathbf{d})\right\|^2 + \lambda_1 \|\mathbf{d}\|_1, \quad \text{subject to}: \mathbf{d} \geq \mathbf{0}. \tag{2-7}$$

The proof is given in Appendix 2.B.

Lemma 2-2 states that we can transform the MNNG problem into an equivalent NNG problem on augmented data. Therefore, given λ_2, the LARS algorithm is adapted to obtain the piecewise linear solution path for MNNG (LARS-MNNG) as follows.

Step a: start from $\mathbf{d}^{[l-1]} = 0, \mathbf{r}^{[l-1]} = \vec{\mathbf{y}}$, and $l = 1$, where superscript [l] denotes the algorithm iteration.

Step b: select the variable that is the most correlated with the augmented response $\ddot{\mathbf{y}}$, and define the active set C_l, which includes the selected variables until iteration *l*.

$$C_l = \arg\max_i \left\{ \left(\ddot{\mathbf{x}}_i \beta_i^{\mathrm{r}} d_i \right)^T \ddot{\mathbf{y}} \right\},$$

where $\ddot{\mathbf{x}}_i$ is the i^{th} column of matrix , β_i^{r} and d_i are the i^{th} element of vectors $\hat{\boldsymbol{\beta}}^{\mathrm{r}}$ and $\mathbf{d}$, respectively.

Step c: compute the projection direction $\boldsymbol{\gamma} = [\gamma_i]$, which is a *P*-dimensional vector with

$$\gamma_i = \begin{cases} \left((\ddot{\mathbf{x}}_i \beta_i^{\mathrm{r}})^T (\ddot{\mathbf{x}}_i \beta_i^{\mathrm{r}}) \right)^{-1} (\ddot{\mathbf{x}}_i \beta_i^{\mathrm{r}})^T \mathbf{r}^{[l-1]} & \text{if variable } i \in C_l \\ 0 \quad \text{Otherwise} \end{cases}$$

Step d: for $\forall i \notin C_l$, compute the amount of progress for the LARS-MNNG in direction $\boldsymbol{\gamma}$ before $\ddot{\mathbf{x}}_i$ enters the active set. This can be measured by an α_i as follows.

$$(\ddot{\mathbf{x}}_i \beta_i^{\mathrm{r}})^T \left(\mathbf{r}^{[l-1]} - \alpha_i \ddot{\mathbf{X}} \boldsymbol{\gamma} \right) = (\ddot{\mathbf{x}}_{i'} \beta_{i'}^{\mathrm{r}})^T \left(\mathbf{r}^{[l-1]} - \alpha_i \ddot{\mathbf{X}} \boldsymbol{\gamma} \right),$$

where i' is arbitrarily chosen from C_l

Step e: for $\forall i \notin C_l$, compute $\alpha_i = \min(-d_i^{[k-1]} / \gamma_i, 1)$.

Step f: if $\forall i, \alpha_i \le 0$, or $\min_{i;\alpha_i>0} \{\alpha_i\} > 1$ set $\alpha = 1$, otherwise set $\alpha = \min_{i;\alpha_i>0} \{\alpha_i\}$, and $i^* = \arg\min_{i;\alpha_i>0} \{\alpha_i\}$. Set $\mathbf{d}^{[l]} = \mathbf{d}^{[l-1]} + \alpha\gamma$. If $i^* \notin C_l$, update $C_{l+1} = C_l \bigcup \{i^*\}$; otherwise update $C_{l+1} = C_l - \{i^*\}$.

Step g: set $\mathbf{r}^{[l]} = \ddot{\mathbf{y}} - \ddot{\mathbf{X}}(\hat{\boldsymbol{\beta}}^{r} \bullet \mathbf{d})$, and $k = k+1$. Go back to step c until $\alpha = 1$.

According to Lemma 2-2, the MNNG criterion can be transformed to an NNG criterion. Therefore, as proven by Yuan and Lin (2007, Theorem 5) the trajectory of the above LARS-MNNG algorithm coincides with the solution path of modified NNG.

Performance of the proposed approach depends on the choice of tuning parameters. There are a few methods in the literature that can be used for this purpose (See Hastie et al., 2009 for details). If enough data are available, the data are divided into 3 subsets; training, validation, and test. Then, the validation subset is used for tuning the parameters. Otherwise, the *k*-fold cross-validation (CV) method is used. In either method, several values of tuning parameters within an applicable range are used to train the model. Then the prediction error is calculated for each of these tuning parameters using validation data, and the tuning parameter with the least prediction error is chosen.

2.2.4 Case of small *n*, large *p* ($n << p$)

In spite of the lasso and NNG methods, the MNNG method can handle the case when the number of significant variables is larger than the sample size. This property is due to the L_2-norm penalty in the MNNG criterion. To show that MNNG can handle such situations, we refer to Lemma 2-2. In the MNNG criterion in (2-7), the dimensions of $\ddot{\mathbf{X}}$ is $(n+P)\times P$ with the maximum rank of $P<(n+P)$, which implies that the MNNG can select all P variables if they are important. In other words, even if $n<P$, using the L_2-norm penalty, the sample size is augmented such that it is always larger than the rank of matrix $\ddot{\mathbf{X}}$.

2.2.5 Similar Effects for Highly Correlated Predictors

In this subsection, we show that the MNNG criterion in (2-4) enjoys the *similar effects* property. In the regression model with "similar effects" property, it is expected that predictors with high correlation have the similar coefficients (up to a sign change if negatively correlated). In the extreme case, particularly, when two predictors are identical, the estimation method should assign the same coefficients to the identical variables. The following theorem is proved to show the proposed MNNG can effectively reflect the *similar effects* property.

Theorem. *Given* λ_1 *and* λ_2, *let* $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ *denote the estimates obtained from the MNNG criterion in (2-4) and* $\rho_{ij} = \tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_j$ *be the sample correlation between predictors* $\tilde{\mathbf{x}}_i$ *and* $\tilde{\mathbf{x}}_j$. *Also, suppose* $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$. *Define* $D_{\lambda_1,\lambda_2}(i, j) = \left|\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)\right| \Big/ \|\mathbf{y}\|_1$, *then* $D_{\lambda_1,\lambda_2}(i, j) \to 0$ *as* $\rho_{ij} \to 1$ *and* $D_{\lambda_1,\lambda_2}(i, j) = 0$ *if* $\rho_{ij} = 1$

The proof is given in Appendix 2.C.

The unitless quantity $D_{\lambda_1,\lambda_2}(i, j)$ represents the difference between two coefficients of predictors *i* and *j* given the tuning parameters. The above theorem indicates that if predictors $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$ are highly correlated, i.e., $\rho_{ij} \to 1$ (if $\rho_{ij} \to -1$, replace $\tilde{\mathbf{x}}_j$ with $-\tilde{\mathbf{x}}_j$), then difference between two coefficients, $D_{\lambda_1,\lambda_2}(i, j)$ tends to zero. In the extreme case with $\rho = 1$, this difference is equal to zero implying that two coefficients are identical. Note that despite the MNNG method, the original NNG

(Breiman 1995) does not benefit from the *similar effects* property. Indeed, the L_2-norm penalty added to NNG criterion in (2-4) leads to this property in the modified NNG.

2.3. Performance Comparison using Simulation

In this section, we conduct a simulation study to evaluate the performance of the proposed HNNG and compare it with some existing methods in the literature. Specifically, we compare HNNG with lasso as an individual variable selection method; and with L_2-norm group lasso, L_∞-norm group lasso, GNNG, and Hlasso as the group variable section methods.

Two different scenarios are considered in this simulation study. In the first scenario, in order to evaluate and compare the prediction performance of different methods, we extend the simulation study conducted by Zhou and Zhu (2010). We first generate 17 independent standard normal variables, $Z_1, Z_2, ..., Z_{16}$, and W. We exploit these variables to generate 16 correlated variables using $X_k = (Z_k + W)/\sqrt{2}, k = 1,...,16$. Then, each of the first 8 variables is expanded through a fourth-order polynomial and each of remaining variables is discretized to 0, 1, 2, 3, and 4 by $\Phi^{-1}(0.2)$, $\Phi^{-1}(0.4)$, $\Phi^{-1}(0.6)$, and $\Phi^{-1}(0.8)$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution. This gives us a total of 8 continuous group variables and 8 discrete group variables. Each group variable consists of 4 individual variables. The following model is used to simulate independent responses.

$$Y = [2.5X_1 + 2.5X_1^3] + [-2X_2^2 - 2X_2^4] + [-3I(X_9 = 0)] + \varepsilon,$$

where I is the indicator function, and ε is the random error generated from a normal distribution $N(0,\sigma^2)$. σ is set such that $std(\sum_{k=1}^{64} X_k \beta_k)/std(\varepsilon)$ is equal to 3.

For this study, we generate 400 observations as training dataset, 400 observations as validation dataset, and 10,000 observations as test dataset. For each method, we try different values of tuning parameters to train the model using the training dataset and choose tuning parameters such that their validation error is minimum. There are 3 criteria used for performance evaluation and comparison, mean square error (MSE) for test dataset, the percentage of true important variables selected by each method (denoted by IV), and the percentage of unimportant variables removed from the model by each method (denoted by UV). We repeat the simulations 100 times and record these criteria. Figure 2-2 shows the boxplot of MSE values obtained from each method. Furthermore, the mean and standard error of each of the criteria over the 100 repetitions are summarized in Table 2-1. For each criterion, the number inside the parenthesis represents the standard error of simulation.

Table 2-1. Summary of simulation results for performance comparison of different methods.

	MSE	IV	UV
HNNG	2.63 (0.20)	91% (1.1%)	90% (1.0%)
Hlasso	4.08 (0.25)	79% (0.8%)	90% (0.8%)
GNNG	4.50 (0.40)	91% (1.1%)	76% (1.6%)
L_2	4.88 (0.24)	86% (0.9%)	51% (1.7%)
L_∞	4.66 (0.29)	85% (0.9%)	50% (1.7%)
Lasso	5.37 (0.33)	89% (1.4%)	72% (1.1%)
OLS	18.94 (1.50)	---	---

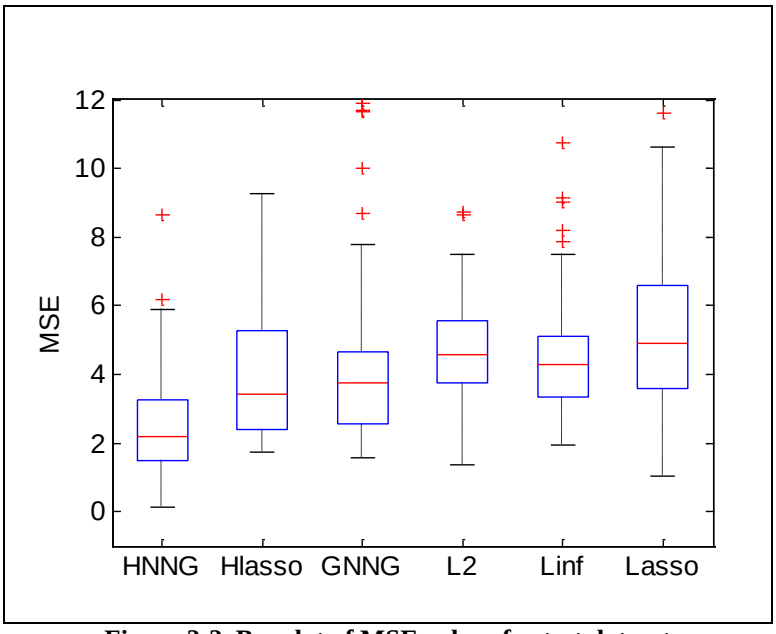

Figure 2-2. Boxplot of MSE values for test dataset

As can be seen from the results, in terms of MSE, the proposed HNNG significantly outperforms other methods with the average MSE of 2.63 and the standard error 0.20. After HNNG, Hlasso has the least MSE among others. The reason of HNNG's and Hlasso's superiority is that both HNNG and Hlasso can perform the variable selection at both the group- and individual-level, while others only perform the group variable selection. Lasso on average has the larger MSE (5.37) compared to L_2-norm group lasso and L_∞-norm group lasso since it cannot select group variables. L_2-norm group lasso and L_∞-norm group lasso perform almost equally. The very large MSE of OLS (18.94) also indicates the essential need of the regularization. Furthermore, from Table 2-1, GNNG and HNNG are equivalently effective in terms of selecting important variables, with IV of 91%. The small IV of Hlasso (79%) indicates that the model obtained by Hlasso is too sparse. This is the reason that the MSE of Hlasso is worse than the MSE of HNNG. In terms of removing unimportant variables, HNNG and Hlasso with

UV of 90% are much better than all other methods. Also, the small UV values for L_2-norm group lasso, L_∞-norm group lasso (51% and 50%) is because of the fact that these methods are not able to remove unimportant variable within the selected groups.

To study the *similar effect* property, in the second scenario, two groups of variables, each of which comprises four variables are considered. All predictors follow a standard normal distribution, i.e., $X_{kp} \sim N(0,1); k = 1,2; j = 1,2,3,4$. The predictors are generated so that $corr(X_{11}, X_{21}) > 0.90$ and $corr(X_{12}, X_{22}) > 0.90$, and other pair-wise correlations are equal to zero. The following data generating model is used to simulate independent responses obtained from highly correlated predictors.

$$Y = [2.5X_{11} - 2X_{12}] + [2.5X_{21} - 2X_{22}] + \varepsilon,$$

where ε is the random error generated from a normal distribution $N(0, \sigma^2)$. σ is set such that $std(\sum_{k=1}^{64} X_k \beta_k) / std(\varepsilon)$ is equal to 3. Similar to the first scenario, we generate 400 observations as training dataset, 400 observations as validation dataset used for choosing tuning parameters. Then, we calculate the $D(i, j)$ values for the highly correlated predictors in each simulation repetition, i.e., $D(X_{11}, X_{21})$ and $D(X_{12}, X_{22})$. The boxplots of $D(i, j)$ values obtained from each method are plotted in Figures 2-3(a) and 2-3(b), respectively.

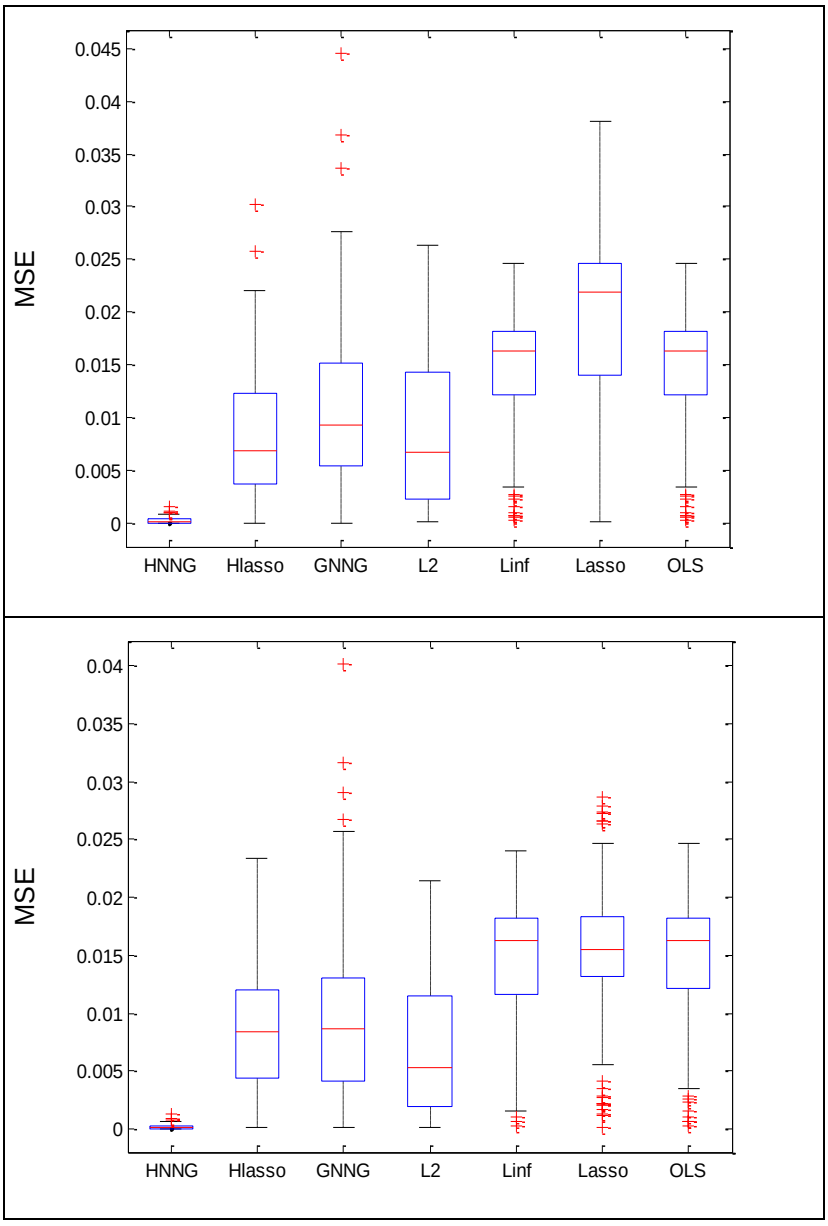

Figure 2-3. Boxplot of unitless difference between two coefficients of predictors
Figure 2-3(a) (top panel). Boxplot of $D(X_{11}, X_{21})$, Figure 2-3(b) (bottom panel). Boxplot of $D(X_{12}, X_{22})$

Both Figures 2-3(a) and 2-3(b) show that $D(i, j)$ values of the proposed HNNG for highly correlated variables on average are smaller than $D(i, j)$ values of all other methods and close to zero. For HNNG, the median of $D(X_{11}, X_{21})$ and $D(X_{12}, X_{22})$ are 0.0002 and 0.0001, respectively. This implies that the coefficients estimated via HNNG

tend to be the same for highly correlated variables. That is, despite other methods, the proposed HNNG benefits from the similar effect property. In short, the simulation results imply that considering MSE, IV, UV, and $D(i, j)$ criteria, HNNG has the best overall performance among all variable selection methods discussed in this chapter.

2.4 Case Study

A case study is shown in this section to further demonstrate the effectiveness of the proposed HNNG method. We apply HNNG to the ingress/egress dataset described in the introduction. In a set of experiments, different drivers tried different design configurations of a vehicle by getting in and getting off the vehicle. During each trial, the motion profiles of 19 points of drivers' body including right hip, right ankle, left hip, left ankle, head, pelvis, T12L1 (on spine) ,T1T2 (on spine), neck, right knee, right toe, right shoulder, right elbow, right wrist, left knee, left toe, left shoulder, left elbow, and left wrist were captured. The spatial trajectories were measured over time in x-,y-, and z-coordinate. After each trial, each driver gave a score in the scale of 1 to 10 about comfort of the design configuration with 1 as the worse and 10 as the best comfort index. The objective of this case study is to identify informative sensors and their corresponding trajectories to construct a regression model for predicting the comfort score of a vehicle.

In this case study, we analyzed ingress profiles. Motion ingress profiles were registered (aligned) based on the starting and stopping time of movement determined by the 3D speed. The dataset contains 254 trials with 57 trajectories in each trial (19 profiles of 19 measurement points × 3 xyz-coordinates), in which each trajectory consists of 200 data points. If each data point in a trajectory were considered as a predictor, the number of predictors would be 11,400 (57×200). To reduce the dimension of predictors, we

modeled each curve by applying a cubic bspline with degrees of freedom of 13 (9 coefficients) and used the bspline coefficients as predictors. Before fitting a cubic bspline, each trajectory was scaled to the interval [0,1], thus the obtained bspline coefficients reflect variations in the shape of trajectories. Also, in order to consider variations in the magnitude of trajectories, two other features including startpoint, and (endpoint – startpoint) of each trajectory were added to the shape features. Therefore, each trajectory was reduced to a vector of 11 features.

In this chapter, a group variable is defined as each measurement position. Each group variable consists of 33 individual variables that correspond to the xyz three trajectories with 11 features representing each trajectory. Therefore, 19 measurement positions yield the total of 627 predictors (19 groups with 33 individual variables in each group). For model estimation and validation, we randomly selected 150 trials as the training dataset and remainders as the test dataset. Then, we used a 10-fold CV for choosing the tuning parameters and training the models. Finally, after training and tuning the parameters, each model was tested using the test dataset. Table 2-2 shows the comparison between HNNG and other existing methods, in which the mean square prediction error (MSPE) along with the number of selected group and individual variables are reported.

Table 2-2. MSPE and sparsity results of different methods for ingress/egress dataset.

	HNNG	GNNG	L_2	L_∞	Hlasso	Lasso	OLS
MSPE	3.28	4.18	4.14	4.11	3.91	4.30	132.03
No. of selected groups	13	13	12	13	9	16	19
No. of selected variables	159	429	396	429	99	17	627

As can be seen from Table 2-2, the proposed HNNG has the least MSPE among all methods with MSPE of 3.28. Also, all group variable selection methods predict the comfort index more accurately than lasso. In terms of sparsity, Hlasso results in the sparsest model by selecting 99 variables within 9 groups. However, it is so sparse that affects the prediction performance. HNNG, on the other hand, selects 271 variables in 13 groups, which results in a sparse model as well as accurate prediction. The number of selected individual variables in HNNG is much smaller than GNNG, L_2, and L_∞. This is because, despite other methods, HNNG is able to select variable in both group- and individual-levels.

Figures 2-4(a) and 2-4(b) also show the coefficients solution path for Step 1 and Step 2 of HNNG obtained from the Ingress dataset, respectively. In both figures, x-axis is the fraction of progress measuring how far the estimate has stepped on the solution path, and y-axis is the value of estimated coefficients. Each line represents a solution path corresponding to one coefficient and each dot represents one iteration of the LARS algorithm. As can be seen from the figures the solution paths for both GNNG and MNNG are linear piecewise. In both figures, the intersections of the vertical dashed line and the solution paths show the optimum coefficients obtained from the 10-fold CV.

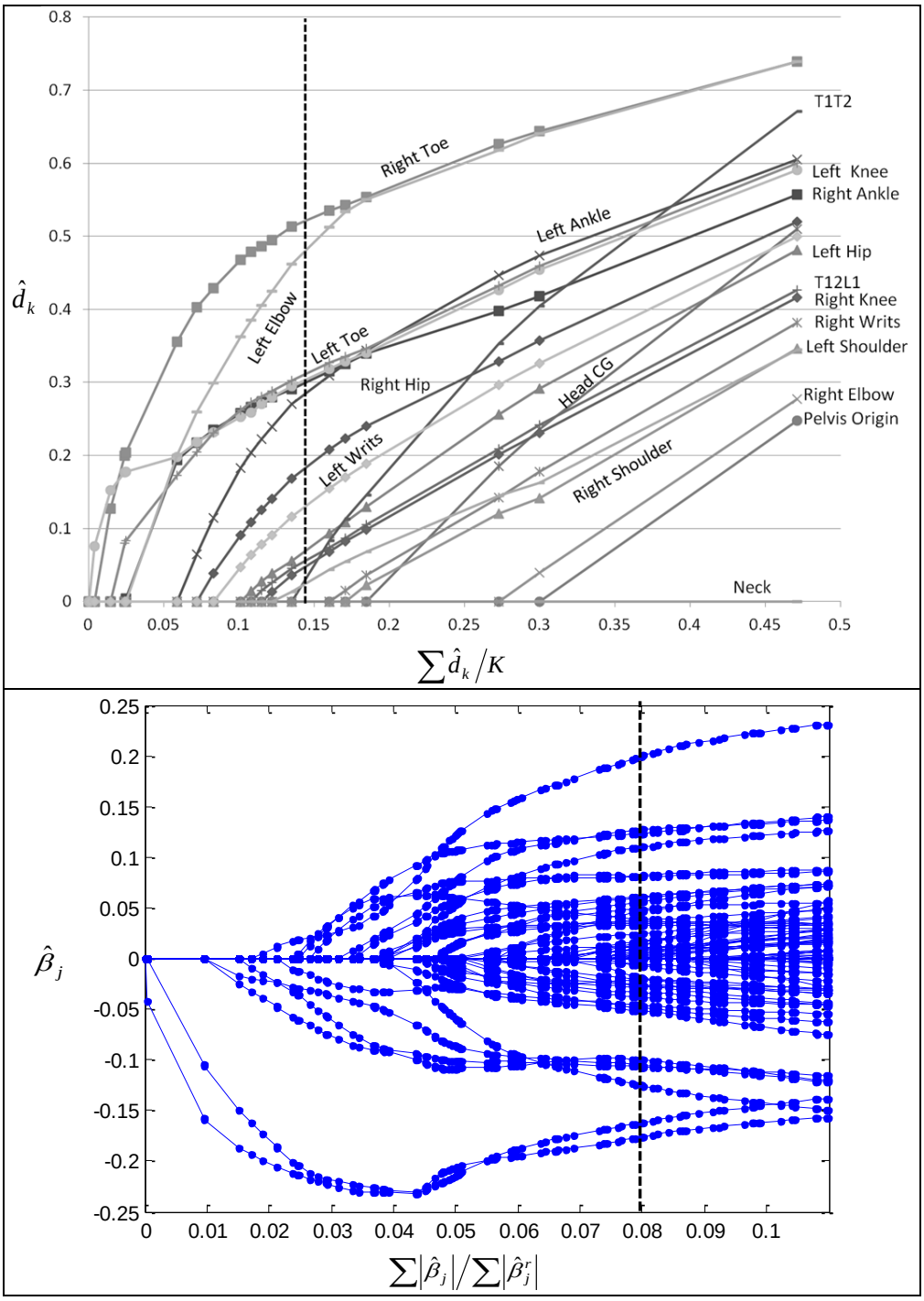

Figure 2-4. Solution paths for HNNG
Figure 2-4(a) (top panel). The solution path of Step 1 of HNNG obtained from GNNG-LARS
Figure 2-4(b) (bottom panel). Partial solution path of Step 2 of HNNG obtained from MNNG-LARS

As can be seen from Figure 2-4(a), 13 trajectories were selected that include left and right toe, left and right knee, left and right hip, left and right wrist, T1T2, T12L1, left

elbow. These trajectories can be grouped into a few clusters based on their location proximity of sensors on the driver's body. These clusters along with the average of $\hat{d}_k$ values for each cluster are reported in Table 2-3. From this table, it can be implied that both right and left foot clusters with average $\hat{d}_k$ of 0.29 are as equally important. However, the trajectories of left hand cluster are more influential on subjects' response than those of the right hand cluster. Right hip has also larger $\hat{d}_k$ than left hip. Another important cluster is the spine cluster including T1T2 and T12L1, which is associated with the drivers' bending when getting inside the car.

Table 2-3. Clusters of selected trajectories and their average importance.

	Cluster 1 (right foot)			Cluster 2 (left foot)			Cluster 3
Selected profiles	Right toe	Right ankle	Right knee	Left toe	Left ankle	Left knee	Right hip
Average $\hat{d}_k$	0.29			0.29			0.18
	Cluster 4	Cluster 5 (right hand)		Cluster 6 (left hand)		Cluster 7 (spine)	
Selected profiles	Left hip	Right wrist		Left wrist	Left elbow	T1T2	T12L1
Average $\hat{d}_k$	0.06	0.02		0.29		0.17	

Appendix 2.A: Proof of Lemma 2-1

Part *i.* In case of orthogonal predictors, Since $\sum_{i=1}^{n} y_i x_{ikj} = \hat{\beta}_{kj}^{\text{ols}}$, criterion (2-3) can be simplified as

$$\min_{d_k} \frac{1}{2}\left[\sum_{i=1}^{n} y_i^2 + \sum_{k=1}^{K}\sum_{j=1}^{p_k}\left(\hat{\beta}_{kj}^{\text{ols}}\right)^2 d_k^2 - 2\sum_{k=1}^{K}\sum_{j=1}^{p_k}\left(\hat{\beta}_{kj}^{\text{ols}}\right)^2 d_k\right] + \lambda\sum_{k=1}^{K} d_k, \qquad (2.A.1)$$
$$\text{subject to}: d_k \geq 0,\ k = 1,2,...,K.$$

Since (2.A.1) is convex, for $d_k > 0$, Karush–Kuhn–Tucker (KKT) sufficient condition can be given as $\sum_{j=1}^{p_k}\left(\hat{\beta}_{kj}^{\text{ols}}\right)^2 d_k - \sum_{j=1}^{p_k}\left(\hat{\beta}_{kj}^{\text{ols}}\right)^2 + \lambda = 0$. Solving this equation with respect to d_k results in $\hat{d}_k = \left(1 - \lambda \Big/ \sum_{j=1}^{p_k}\left(\hat{\beta}_{kj}^{\text{ols}}\right)^2\right)_+$.

<u>Part *ii*.</u> Since $\sum_{i=1}^{n} y_i x_{ij} = \hat{\beta}_j^{\text{ols}}$ and $\hat{\beta}_j^{\text{r}} = \hat{\beta}_j^{\text{ols}} \big/ (1+\lambda_2)$, criterion (2-4) can be simplified as

$$\min_{d_k} \frac{1}{2}\left[\sum_{i=1}^{n} y_i^2 + \sum_{j=1}^{P}\left(\frac{\hat{\beta}_j^{\text{ols}}}{1+\lambda_2}\right)^2 d_j^2 - 2\sum_{j=1}^{P}\frac{\left(\hat{\beta}_j^{\text{ols}}\right)^2}{1+\lambda_2} d_j\right] + \lambda_1\sum_{j=1}^{P} d_j + \frac{\lambda_2}{2}\sum_{j=1}^{P} d_j^2, \qquad (2.A.2)$$
$$\text{subject to}: d_k \geq 0, k = 1,...,K.$$

Because (2.A.2) is convex, for $d_j > 0$, Karush–Kuhn–Tucker (KKT) sufficient conditions can be given as $\left((\frac{\hat{\beta}_j^{\text{ols}}}{1+\lambda_2})^2 + \lambda_2\right) d_j + \frac{\left(\hat{\beta}_j^{\text{ols}}\right)^2}{1+\lambda_2} + \lambda_1 = 0$. Solving this equation with respect to $\beta_j = \hat{\beta}_j^{\text{ols}} d_j \big/ (1+\lambda_2)$ results in $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^{\text{ols}})\left(\left|\hat{\beta}_j^{\text{ols}}\right| \gamma_1 - \gamma_2\right)_+$, where $\gamma_1 = \left(\frac{\left(\hat{\beta}_j^{\text{ols}}\right)^2}{\lambda_2(1+\lambda_2)^2 + \left(\hat{\beta}_j^{\text{ols}}\right)^2}\right)$ and $\gamma_2 = \left(\frac{\lambda_1(1+\lambda_2)}{\lambda_2(1+\lambda_2)^2 + \left(\hat{\beta}_j^{\text{ols}}\right)^2}\right)$. ■

Appendix 2.B: Proof of Lemma 2-2

After plugging $\ddot{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$, and $\ddot{\mathbf{X}} = \begin{bmatrix} \widetilde{\mathbf{X}} \\ \sqrt{\lambda_2 / 2}\mathbf{B} \end{bmatrix}$ into $\frac{1}{2}\left\| \ddot{\mathbf{y}} - \ddot{\mathbf{X}}(\hat{\ddot{\boldsymbol{\beta}}}^{\mathrm{r}} \bullet \ddot{\mathbf{d}}) \right\|^2 + \lambda_1 \|\mathbf{d}\|_1$ and doing algebraic simplification, we will obtain the criterion in (2-4). ■

Appendix 2.C: Proof of Theorem

To prove this theorem, first we prove the following lemma.

Lemma 2-3 *let* $\hat{\boldsymbol{\beta}}^{\mathrm{r}}$ *denote the ridge estimate and suppose* $\hat{\beta}_i^{\mathrm{r}} \hat{\beta}_j^{\mathrm{r}} > 0$ *. Then* $\left| \hat{\beta}_i^{\mathrm{r}} - \hat{\beta}_j^{\mathrm{r}} \right| \to 0$ *as* $\rho_{ij} \to 1$.

Proof: According to Theorem 1 of Zou and Hastie (2005) with $\lambda_1 = 0$, it is true that

$$\left| \hat{\beta}_i^{\mathrm{r}} - \hat{\beta}_j^{\mathrm{r}} \right| \le \frac{\|\mathbf{y}\|}{\lambda_2} \sqrt{2(1-\rho)} \,. \tag{2.C.1}$$

Therefore $\left| \hat{\beta}_i^{\mathrm{r}} - \hat{\beta}_j^{\mathrm{r}} \right| \to 0$ *as* $\rho_{ij} \to 1$.■

Proof of Theorem: Suppose $\hat{d}_i$ and $\hat{d}_j$ are the solutions of (2-4), and $\hat{\beta}_i^{\mathrm{r}}$ and $\hat{\beta}_j^{\mathrm{r}}$ are the ridge estimates corresponding to $\hat{\beta}_i(\lambda_1, \lambda_2)$, and $\hat{\beta}_j(\lambda_1, \lambda_2)$, respectively. So, based on the KKT first order necessary condition for $\hat{d}_i$ and $\hat{d}_j$, it can be stated that $-\widetilde{\mathbf{x}}_i^T(\mathbf{y} - \widetilde{\mathbf{X}}\hat{\boldsymbol{\beta}})\hat{\beta}_i^{\mathrm{r}} + \lambda_1 + \lambda_2 \hat{d}_i = 0$ and $-\widetilde{\mathbf{x}}_j^T(\mathbf{y} - \widetilde{\mathbf{X}}\hat{\boldsymbol{\beta}})\hat{\beta}_j^{\mathrm{r}} + \lambda_1 + \lambda_2 \hat{d}_j = 0$. This results in

$$\hat{\beta}_i^{\mathrm{r}} \hat{d}_i - \hat{\beta}_j^{\mathrm{r}} \hat{d}_j = \frac{1}{\lambda_2}\left[\left(\widetilde{\mathbf{x}}_i^T \left(\hat{\beta}_i^{\mathrm{r}}\right)^2 - \widetilde{\mathbf{x}}_j^T \left(\hat{\beta}_j^{\mathrm{r}}\right)^2 \right) \hat{\mathbf{r}} - \lambda_1 (\hat{\beta}_i^{\mathrm{r}} - \hat{\beta}_j^{\mathrm{r}}) \right], \tag{2.C.2}$$

where $\hat{\mathbf{r}} = (\mathbf{y} - \widetilde{\mathbf{X}}\hat{\boldsymbol{\beta}})$ is the residuals vector.

Since $\tilde{\mathbf{X}}$ is normalized, $\left|\tilde{\mathbf{x}}_i^T\left(\hat{\beta}_i^{\mathrm{r}}\right)^2-\tilde{\mathbf{x}}_j^T\left(\hat{\beta}_j^{\mathrm{r}}\right)^2\right|=\sqrt{\left(\hat{\beta}_i^{\mathrm{r}}\right)^4+\left(\hat{\beta}_j^{\mathrm{r}}\right)^4-2\rho\left(\hat{\beta}_i^{\mathrm{r}}\hat{\beta}_j^{\mathrm{r}}\right)^2}$. From (2.C.2), and triangle inequality, it can be stated that

$$\left|\hat{\beta}_i-\hat{\beta}_j\right|\le\frac{1}{\lambda_2}\left[\sqrt{\left(\hat{\beta}_i^{\mathrm{r}}\right)^4+\left(\hat{\beta}_j^{\mathrm{r}}\right)^4-2\rho\left(\hat{\beta}_i^{\mathrm{r}}\hat{\beta}_j^{\mathrm{r}}\right)^2}\left\|\hat{\mathbf{r}}\right\|+\lambda_1\left|\hat{\beta}_i^{\mathrm{r}}-\hat{\beta}_j^{\mathrm{r}}\right|\right]. \qquad (2.C.3)$$

Because $\hat{\mathbf{d}}$ is the optimum solution to (2-4), we have $\frac{1}{2}\left\|\hat{\mathbf{r}}\right\|^2+\lambda_1\left\|\hat{\mathbf{d}}\right\|_1+\frac{\lambda_2}{2}\left\|\hat{\mathbf{d}}\right\|_2^2\le\frac{1}{2}\left\|\mathbf{y}\right\|^2$, which implies that $\left\|\hat{\mathbf{r}}\right\|\le\left\|\mathbf{y}\right\|$. From (2.C.1), (2.C.3), and $\left\|\hat{\mathbf{r}}\right\|\le\left\|\mathbf{y}\right\|$, it can be implied that

$$D_{\lambda_1,\lambda_2}(i,j)=\frac{\left|\hat{\beta}_i-\hat{\beta}_j\right|}{\left\|\mathbf{y}\right\|}\le\frac{1}{\lambda_2}\left[\sqrt{\left(\hat{\beta}_i^{\mathrm{r}}\right)^4+\left(\hat{\beta}_j^{\mathrm{r}}\right)^4-2\rho\left(\hat{\beta}_i^{\mathrm{r}}\hat{\beta}_j^{\mathrm{r}}\right)^2}+\frac{\lambda_1}{\lambda_2}\sqrt{2(1-\rho)}\right]. \qquad (2.C.4)$$

Lemma 2-3 and inequality (2.C.4) together imply that $D_{\lambda_1,\lambda_2}(i,j)\to 0$ as $\rho_{ij}\to 1$, and $D_{\lambda_1,\lambda_2}(i,j)=0$ if $\rho_{ij}=1$.■

References:

1. Antoniadis, A., and Fan, J. (2001), Regularization of wavelet approximations. *J. Am. Statist. Ass.*, **96**, 939–967.

2. Bondell, H. D., and Reich, B. J. (2008) Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics*, **64**, 115-123.

3. Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, 37, 373–384.

4. Cantoni, E., Flemming J. M., and Ronchetti, E. (2011) Variable selection in additive models by nonnegative garrote. *Statistical Modeling*, **11**, 237-252.

5. Chaovalitwongse, W., Iasemidis, L. D., Pardalos, P. M., Carney, P. R., Shiau, D., Sackellares, J. C., Performance of a seizure warning algorithm based on the dynamics of intracranial EEG. *Epilepsy Res* 2005, **64**, 93-113

6. Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics*, 32, 407–451.

7. Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

8. Frank, I.E., and Friedman, J.H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, **35**, 109–148.

9. Fu, W. J. (1999) Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist*., **7**, 397–416.

10. Hastie, T., Tibshirani,R., and Friedman. J. (2009) *The elements of statistical learning: prediction, inference and data Mining*. Springer, Verlag, NY.

11. Huang, J., Ma, S., Xie, H., and Zhang, C. (2009) A group bridge approach for variable selection. *Biometrika*, **96**, 339–355.

12. Meier, L., van der Geer, S., and Buhlmann, P., (2008) The group lasso for logistic regression. *J. R. Statist. Soc. B*, **70**, 53–71.

13. Miller, A. (2002) *Subset selection in regression*. Chapman and Hall, Boca Raton, FL.

14. Sturm, J., (1999) Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, **11**. 625–653.

15. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267–288.

16. Yuan, M. (2007) Nonnegative garrote component selection in functional ANOVA models. *In Proceedings of AI and Statistics, AISTATS*, 656-662.

17. Yuan, M., and Lin, Y. (2006) Model selection and estimation in regression with grouped variable. *J. R. Statist. Soc.B*, **68**, 49-67.

18. Yuan, M. and Lin, Y. (2007) On the nonnegative garrote estimate. *J. R. Statist. Soc. B*, **69**, 143–161.

19. Yuan, M., Roshan, J., and Zou, H. (2009). Structured variable selection and estimation, *Ann. of App. Stat.*, **3**, 1738-1757.

20. Zhao, P., Rocha, G., and Yu, B. (2009) The composite absolute penalties family for grouped and hierarchical variable selection. *The Ann. of Stat.*, **37**, 3468–3497.

21. Zhou, N., and Zhu, J. (2008) Group variable selection via a hierarchical lasso and its oracle property. *Technical report*, University of Michigan, Dept. of Statistics, 357-358.

22. Zou, H. (2006) The adaptive Lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418-1429

23. Zou, H., and Hastie, T. (2005) Regularization and variable selection via the elastic net *J. R. Statist. Soc. B,* **67**, 301–320.

CHAPTER III

Characterization of Nonlinear Profile Variations Using Nonparametric Mixed-Effect Models and Wavelets

3.1 Introduction

With the rapid development of embedded sensing and computer technologies, online sensing and monitoring systems have been increasingly used for manufacturing process control. In many practical situations, the sensor measurements are shown as time-dependent *functional data*, which are also called "profile data" or "waveform signals." Some examples include the welding force responses recorded in resistance welding operations at the uniform sampling time intervals (Chu et al., 2004), the tonnage signature signals measured in stamping processes at the equal crank angle sampling intervals (Jin and Shi, 1999), the vertical density profile of a particleboard measured at each fixed vertical depth (Walker and Wright, 2002), and the ram force signals used to press valve seats into engine heads in engine head assembly processes.

Most of the previous research has focused on linear profile monitoring. Some of these studies include the work done by Kang and Albin (2000), Kim et al. (2003), Mahmoud and Woodall (2004), Mahmoud et al. (2007), Zou et al. (2006), and Jensen et al. (2008). Meanwhile, nonlinear profile modeling and monitoring has also generated increasing interest in the statistical process control (SPC) research field for complicated profile data. For example, Gardner et al. (1997) utilized smoothing spline to model

nonlinear profiles. Williams et al. (2007) used a four-parameter logistic regression and smoothing spline to model dose-response profiles of a drug. To monitor and distinguish out-of-control nonlinear profiles in Phase I, Ding et al. (2006) considered each profile as a high-dimensional data set and applied principal component analysis (PCA) and independent component analysis to reduce the dimension of the nonlinear profile data while preserving the cluster structure of the profiles. Zou et al. (2008) used local linear smoothers to monitor nonlinear profiles. Zou et al. (2009) applied the generalized likelihood ratio test to develop a monitoring procedure for nonlinear profiles modeled by local linear kernel smoothing. To determine the control limit, they used the bootstrap method based on a few samples of in-control profiles.

In all of the studies mentioned above, it is assumed that the total variability of profiles can be modeled by random noises, which are typically assumed to be normally independently distributed (NID). In those cases, such random noises are mainly used to reflect the within-profile variation with a constant variance over all measurement points. In many practical situations, however, the variation among in-control profiles is too large to be handled by NID noises only. Growth curves (Ramsay and Silverman, 1997), as shown in Figure 3-1, are typical examples of such situations. Since the growth of one individual generally differs from that of others, a large amount of profile variations is due to the person-to-person growth variability, which may not be fully explained by only the random noises within each person's growth curve.

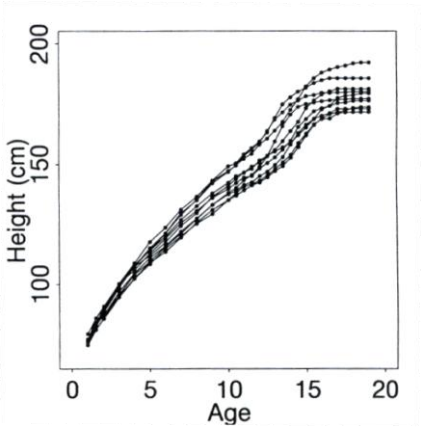

Figure 3-1: Growth curves of 10 Swiss boys

As another example (to be discussed in detail in Section 3.7), the inserting forces of a pressing machine (as shown in Figure 3-2) are used to press valve seat rings into an engine head. These force signals are continuously recorded during each cycle of repeated pressing operations. The overlapping multiple samples of signals collected at different cycles of in-control operations are shown in Figure 3-2(a). Furthermore, to show the magnitude of random noises within each profile, an individual signal is also depicted in Figure 3-2(b), in which the dotted points represent the actual measurements, and the solid line is the fitted profile through the wavelet-based denoising method. As can be seen in Figure 3-2(b), the within-profile variation obtained from the fitted model residuals e_t is much smaller than the part-to-part (i.e., curve-to-curve) variation shown in Figure 3-2(a). In other words, a significant portion of the total inherent variation is mainly reflected in the between-profile variation and is too large to be taken into account by only random noises corresponding to the within-profile variation.

In practice, there are many causes for such inevitable between-profile variations, such as part-to-part variation, fixture or tooling tolerance, and/or process operation condition variations. The between-profile variation may affect the local profile shape

differently at different segments of a profile. In contrast, the within-profile variation is mostly due to measurement errors and environmental disturbances, which independently and identically affect all observations of an entire profile. Therefore, characterization and estimation of between-profile and within-profile variations will not only help monitor the process more effectively, but also provide us with a better understanding of the root causes of process variations, which can expedite further decision making for variation reduction and process improvement.

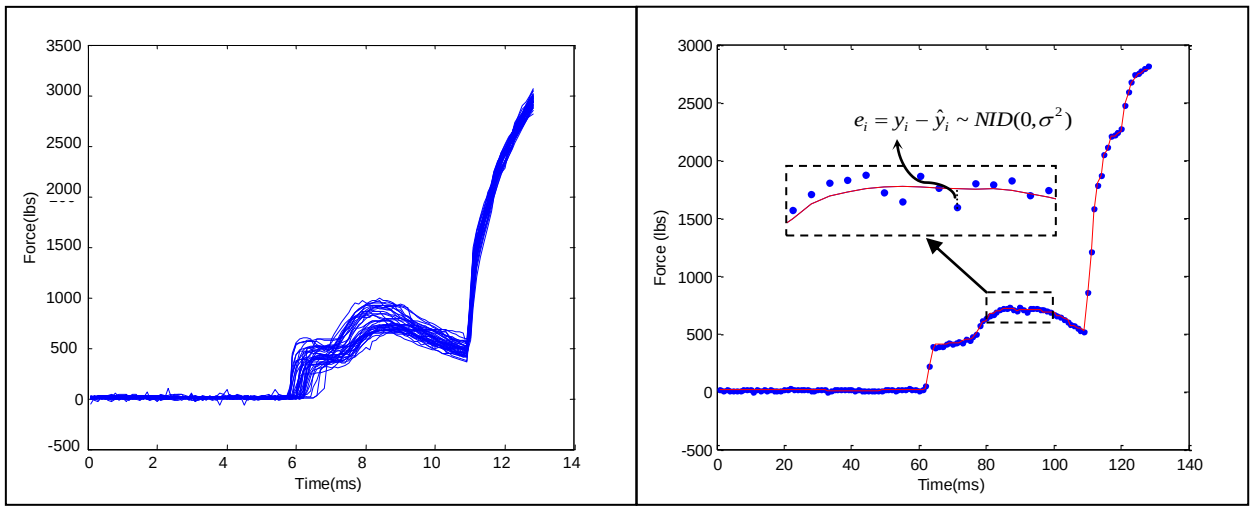

Figure 3-2. Pressing force profile signals in a valve seat assembly operation
Figure 3-2(a) (left panel) overlapped multiple samples of profile signals. Figure 3-2(b) (right panel) one original profile (dot) and fitted profile (line)

Recently, a few advanced modeling methods have been developed on variation modeling for nonlinear profiles by considering both within-profile and between-profile variations via mixed-effect models (hereafter called mixed models for simplicity). For example, Mosesova et al. (2006) and Jensen and Birch (2009) developed a parametric mixed model by including a few model parameters as the random effects to reflect the between-profile variation. Applying a parametric model, however, is not always achievable because it requires strong domain knowledge and tremendous modeling efforts to identify an appropriate parametric model structure. To overcome this challenge,

an alternative approach of using nonparametric mixed models has attracted recent research interests. Mosesova et al. (2006) developed a mixed model by using a B-spline basis. Shiau et al. (2009) used a random-effect B-spline model along with PCA for monitoring nonlinear profiles. Although splines and PCA have been considered effective nonparametric approaches for modeling and analyzing smooth nonlinear profiles, they are inherently ineffective for modeling complicated nonlinear profiles with local sharp jumps or nondifferentiable points. The wavelet transform is a nonparametric alternative that can be effectively used for modeling nonlinear profiles with sharp jumps.

One of the research challenges in using wavelets for process monitoring is determining how to select a low dimension of monitoring features from the large dimension of wavelet coefficients. Jin and Shi (1999) developed a feature-preserving wavelets-based thresholding method to extract monitoring features from complicated tonnage waveform signals, and then constructed a Hotelling T^2 control chart based on unthresholded wavelet coefficients for stamping process monitoring (Jin and Shi, 2001). However, their method is limited to detecting profile changes that are reflected by the selected unthresholded wavelet coefficients. To overcome this problem, Jeong et al. (2006) presented an adaptive thresholding procedure that threseholds the wavelet coefficients of each incoming profile and updates the selected coefficients based on those that are unthresholded.

Chicken et al. (2009) developed a change-point model based on the likelihood ratio test, in which all wavelet coefficients are taken into account. They showed that monitoring wavelet coefficients is equivalent to evaluating the hypothesis that the noncentrality parameter of a Chi-square distribution is equal to zero. They estimated the

noncentrality parameter based on the unthresholded coefficients, which can reduce the variance of the estimator and consequently improve the performance of monitoring methods. They also showed that the change-point model outperforms the methods proposed by Jin and Shi (2001) and Jeong (2006).

All of the aforementioned work on wavelet-based monitoring approaches only consider the within-profile variation in modeling of the total profile variability. Very little research has been done on wavelet-based profile modeling or on monitoring methods that can account for both within-profile and between-profile variations. Therefore, the objective of this chapter is to develop a mixed model based on wavelets for the following two purposes: (1) to characterize nonlinear profile variations by considering both between-profile and within-profile variations, thus going beyond the existing wavelets-based nonparametric modeling methods that account for only the within-profile variation; and (2) to characterize both global and local segmental variation patterns by mapping scale/detail wavelet coefficients into profile segments, which goes beyond the existing methods (such as PCA or splines) that mainly characterize global variations for smooth nonlinear profiles.

In this chapter, the wavelets transform is selected by considering its following three unique merits over other nonparametric approaches: (1) the wavelets-based modeling is capable of fitting complicated nonlinear profiles with sharp jumps and nondifferentiable points; (2) the multi-scale wavelet coefficients have the unique capability to separate the within-profile noises (at the high frequency range) from the true profile signal (mainly at the low frequency range), which can significantly simplify the computation for estimating the mixed model parameters; and (3) the mapping relationship between the

multiresolution wavelet coefficients and the local profile segments can facilitate the identification of the sources of the between-profile variation.

Implementing the proposed mixed model involves two critical research issues. The first is to ensure that the collected profile samples used for model estimation have an identical mixed model distribution. It is well known that combining samples from different distributions would lead to a large estimation error for the model parameters, thus resulting in a misleading model. For this purpose, a change-point model, which is used to group profiles based on their distributions, is developed based on a likelihood ratio test (LRT). The other critical issue is knowing how to reduce the computation for implementing a mixed model. It is well known that the computation required for estimating model parameters increases with the number of random parameters. Demidenko (2004) recommended a method for constructing a model starting with one random coefficient and then adding more random parameters one-by-one if needed. However, this step-by-step exploration approach may not be very effective considering the large number of wavelet coefficients transformed from profile signals. Therefore, although wavelet transform is an effective approach for modeling complex nonlinear profiles, implementing a wavelets-based mixed model is still challenging. This chapter also discusses how to effectively select a low dimension of wavelet random effects in the construction of the proposed mixed model, which can be well suited for characterizing the between-profile variation.

The remainder of the chapter is organized as follows. Section 3.2 provides an overview of the proposed methodology, and Section 3.3 gives a brief review of the wavelets transform used for profile signals. The development of the proposed wavelet-

based mixed model is elaborated in detail in Section 3.4. In Section 3.5, an LRT-based change point model is developed to check the distribution identicalness of the collected profile samples. The performance of the proposed approach is examined through both Monte-Carlo simulations and a case study in Sections 3.6 and 3.7, respectively.

3.2 Overview of the proposed methodology

A general framework of the proposed methodology is shown in Figure 3-3. Firstly, the measured nonlinear profile data are transformed into the wavelet domain by using a selected wavelet basis. Then, a mixed model is developed on the wavelet coefficients, in which a two-step modeling approach is developed to reduce the computational complexity in the mixed model estimation. At the first step, the wavelet denoising thresholding is conducted on each profile in order to separate within-profile noises from between-profile variations. At the second step, in order to reduce the dimension of the parameters in the mixed model, a few wavelet coefficients with random-effect are selected. Afterward, the LRT-based change-point model is applied to check the distribution identicalness of the collected profile samples. This result is used to group profile samples based on their distributions for further estimation of mixed model parameters. Finally, a mapping between the selected wavelet coefficients with the random-effect and the profile segments is conducted to facilitate the identification of variation sources. The detailed analysis of each step will be elaborated in subsequent sections.

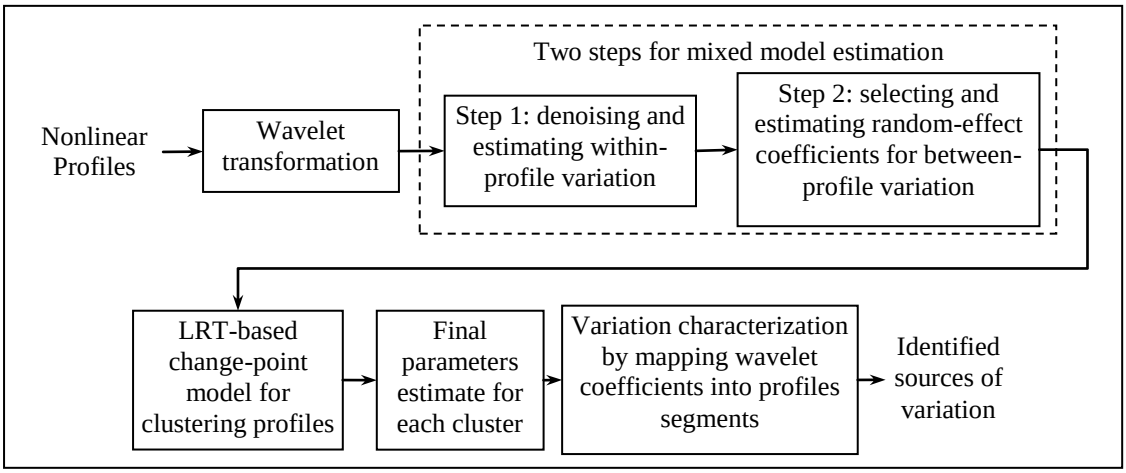

Figure 3-3. Flow diagram of the proposed methodology.

3.3 Wavelet transformation for profile signals

Suppose there are *m* available profiles, each of which consists of *n* pairs of (*t*, *y*) discrete observations that can be generally described by

$$\mathbf{y}_i = f_i(\mathbf{t}) + \boldsymbol{\varepsilon}_i \text{ for } i = 1,2,...,m, \tag{3-1}$$

where $\mathbf{y}_i$ is a vector of the discrete response measurements of profile *i*; $f_i(.)$ is an unknown nonlinear function of profile *i*; $\mathbf{t}$ is a vector comprising of equally spaced sampling time or distance; and $\boldsymbol{\varepsilon}_i$ is a vector of NID noises with $\boldsymbol{\varepsilon}_i \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I})$ to represent the within-profile variation, where $\mathbf{I}$ is an $n \times n$ identity matrix.

The first step of the proposed procedure, as shown in Figure 3-3, is to transform each profile into the wavelet domain. It is well known that any function g in $L^2(\Re)$, the square-integrable functions space can be expressed by a wavelet series of $g(t) = \sum_{k \in Z} c_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{\infty}\sum_{k \in Z} d_{jk}\psi_{jk}(t)$ (Daubechies, 1992). Functions $\phi(.)$ and $\psi(.)$ are known as father and mother wavelets basis, respectively. They are used to decompose

function g into two parts corresponding to low-frequency (coarse) and high-frequency (detail). The multiresolution decomposition property of wavelets is performed by a set of orthonormal wavelets basis of $\phi_{j_0k}(t) = 2^{j_0/2}\phi(2^{j_0}x - k)$ and $\psi_{jk}(t) = 2^{j/2}\psi(2^j x - k)$; for any nonnegative integer $j \geq j_0$. The decomposed coefficients c_{j_0k} and d_{jk} are called approximate and detail wavelet coefficients, which are determined by the inner product of g and the corresponding wavelet functions, i.e., $c_{j_0k} = \langle g, \phi_{j_0k} \rangle$, and $d_{jk} = \langle g, \psi_{jk} \rangle$, where $\langle\ \rangle$ represents the inner product operator.

When the number of discrete measurements (n) in each profile is dyadic, i.e., $n = 2^J$; with J as a positive integer number, a fast numerical algorithm called discrete wavelet transform (DWT), can be used to determine wavelet coefficients (Mallat, 1999). The matrix form of DWT is represented as $\mathbf{z} = \mathbf{W}\mathbf{y}$, where $\mathbf{W}_{n\times n}$ is an orthogonal real matrix depending on the selected orthogonal wavelet basis, and the vector $\mathbf{z} = [\mathbf{c}_{J-l_0}, \mathbf{d}_{J-l_0}, \mathbf{d}_{J-l_0+1}, ..., \mathbf{d}_{J-1}]^T$ represents all decomposed wavelet coefficients, where superscript T denotes the transpose operator. The elements of $\mathbf{c}_{J-l_0}$ denote the approximate coefficient vector at the decomposition level $l_0 (1 \leq l_0 \leq J)$, $\mathbf{y}$ can be represented by $\mathbf{c}_J$, and $\mathbf{d}_{J-l} (l = 1,2,...,l_0)$ denotes the detail coefficient vector at the decomposition level l. More details about the wavelets transform can be found in Mallat (1999) and Daubechies (1992).

In the chapter, an orthogonal Haar transform is used for the discretized profile data $\mathbf{y}_i = f_i(\mathbf{t}) + \boldsymbol{\varepsilon}_i$ and the resulting wavelet coefficients are represented as $\mathbf{z}_i = \boldsymbol{\theta}_i + \tilde{\boldsymbol{\varepsilon}}_i$;

where $\boldsymbol{\theta}_i = \mathbf{W}f_i(\mathbf{t})$ is a vector of the true wavelet coefficients transformed from the true profile function $f_i(\mathbf{t})$; $\mathbf{z}_i = \mathbf{W}\mathbf{y}_i$ is a vector of the empirical wavelet coefficients transformed from noisy profile $\mathbf{y}_i$; and $\tilde{\boldsymbol{\varepsilon}}_i = \mathbf{W}\boldsymbol{\varepsilon}_i$ is a random noise vector in the wavelet domain with $\tilde{\boldsymbol{\varepsilon}}_i \sim \mathrm{MVN}(0, \sigma^2\mathbf{I})$.

3.4 Mixed Model for Wavelet Coefficients

To consider the between-profile variation, a mixed model, in which a few wavelet coefficients are considered as random effects, is utilized. Davidian and Giltinan (1995), Pinheiro and Bates (2000), and Demidenko (2004) provided a comprehensive introduction of mixed-effect models. It is commonly known that constructing a parametric nonlinear mixed model and estimating the corresponding parameters is often based on numerical methods, which are computationally expensive and might not converge if the number of random effects is large. In contrast, the wavelets transform provides a multi-scale linear transformation capable of separating the within-profile variation from the true between-profile variation. This capability of the wavelets transform allows us to simplify the estimation of the mixed model parameters.

To implement the mixed model based on wavelet coefficients, let $\boldsymbol{\theta}_i = \boldsymbol{\mu} + \mathbf{b}_i$, where $\boldsymbol{\mu}$ is the vector of fixed effects common to all profiles, $\mathbf{b}_i$ is the vector of random effects of profile *i* with $\mathbf{b}_i \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Lambda})$, and $\boldsymbol{\Lambda}$ is a positive definite matrix representing the covariance structure of random effects. In this chapter, $\boldsymbol{\Lambda}$ is assumed to be a diagonal matrix, which implies that the random effects are uncorrelated. The reason for this assumption is that the maximum likelihood (ML) estimate of the covariance matrix could become negative definite if the sample size is less than the dimension of the covariance

matrix, especially in the wavelet-based mixed model where the number of wavelet coefficients is often large. However, if the sample size is large enough, the method presented in this is applicable to cases in which no restriction on the covariance matrix structure (such as a diagonality assumption) exists. Furthermore, such a diagonality assumption only restricts the covariance among the random effects of between-profile variations, which implies neither the independency nor a constant variance across different data points within a profile. We also assume that in the equation $\mathbf{z}_i = \boldsymbol{\mu} + \mathbf{b}_i + \tilde{\boldsymbol{\varepsilon}}_i$, $\mathbf{b}_i$ is independent from $\tilde{\boldsymbol{\varepsilon}}_i$. Based on this mixed model, the parameters of $\boldsymbol{\mu}$ and $\mathbf{b}_i$ can be effectively used to represent the profile mean and between-profile variation, respectively.

In order to construct a wavelet-based mixed model, a two-step modeling approach is proposed (Figure 3-3). At the first step, the within-profile variation is estimated and removed through wavelet denoising thresholding of each sample of profile signals. As a result, the sample-to-sample variability of the remaining wavelet coefficients mainly reflects the between-profile variation. Therefore, at the second step, the between-profile variation is estimated based on the remaining wavelet coefficients using all collected samples of profile signals if they follow an identical mixed model distribution. The investigation of distribution identicalness among samples will be discussed in Section 3.5. In order to effectively estimate the mixed model parameters and identify the major sources of variations, a data dimension reduction approach is further explored by selecting a few significant wavelet coefficients, which are sufficient to characterize the majority of the between-profile variation. The following two subsections will discuss the

details of the proposed two-step estimation of the mixed model parameters on the wavelet coefficients.

3.4.1 Characterizing Within-Profile Variation and Denoising

In this subsection, wavelet-based denoising thresholding is utilized to estimate and remove the within-profile variation of noises $\tilde{\boldsymbol{\varepsilon}}_i$. As Mallat (1989) indicates, only a few wavelet coefficients contribute to the original true function of profiles. Therefore, denoising thresholding can be effectively applied to wavelet coefficients to remove the within-profile variation.

Since the white noises equally contribute to the wavelet coefficients, the soft thresholding approach introduced by Donoho and Johnstone (1995) is applied with the following thresholding rule:

$$\eta(\mathbf{z}_i;\hat{\sigma}_i,n)=\operatorname{sign}(\mathbf{z}_i)\left(|\mathbf{z}_i|-\hat{\sigma}\sqrt{2\log n}\right)I\left(|\mathbf{z}_i|>\hat{\sigma}\sqrt{2\log n}\right), \tag{3-2}$$

where $\eta(\cdot)$ is the soft thresholding function, $I(\cdot)$ represents an indicator function, $\operatorname{sign}(\cdot)$ is the sign function, and $\hat{\sigma}$ denoting the estimated standard deviation of $\tilde{\boldsymbol{\varepsilon}}_i$, is calculated by $\hat{\sigma}^2=\sum_{i=1}^{m}\hat{\sigma}_i^2/m$; where $\hat{\sigma}_i=\operatorname{median}\left(|\mathbf{d}_{i,J-1}-\operatorname{median}(\mathbf{d}_{i,J-1})|\right)/0.6745$ with $\mathbf{d}_{i,J-1}$ denoting the detail coefficients at the lowest decomposition level (Donoho and Johnstone, 1995). Based on equation (3-2), if a coefficient is less than the threshold of $\hat{\sigma}\sqrt{2\log n}$, it will be shrunken to zero. The denoised profile data and wavelet coefficients are denoted as $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{z}}$, respectively. It is known that the maximum of *n* independent Gaussian white noises cannot exceed $\sigma\sqrt{2\log n}$ when *n* is large (Fan, 1996), i.e.,

$\Pr(\max|\tilde{\boldsymbol{\varepsilon}}| \le \sigma\sqrt{2\log n}) \to 1$ with $\tilde{\boldsymbol{\varepsilon}} \sim NID(0, \sigma^2\mathbf{I})$. This implies that with a high probability, all noises are shrunk towards zero when n is large. Thus, the remaining coefficients approximate the wavelet coefficients of the true profile signals.

As shown in Appendix 3.A, for each profile, the conditional variance of the denoised wavelet coefficients, given $\mathbf{b}_i$ denoted by $\sigma^2_{\tilde{z}_{ir}}$ ($r = 1,2,..,n$), can be calculated as follows:

$$\sigma^2_{\tilde{z}_{ir}} = \left\{\left(\mu^r_{z_{ir}}(\zeta) - \zeta\right)^2 + \left(\sigma^r_{z_{ir}}(\zeta)\right)^2\right\}\Phi\left(\frac{\mu_{z_{ir}} - \zeta}{\sigma}\right) + \left\{\left(\mu^l_{z_{ir}}(-\zeta) + \zeta\right)^2 + \left(\sigma^l_{z_{ir}}(-\zeta)\right)^2\right\}\Phi\left(\frac{-\mu_{z_{ir}} - \zeta}{\sigma}\right) - \mu^2_{\tilde{z}_{ir}} \quad (3\text{-}3)$$

where ζ is the threshold value $\sigma\sqrt{2\log n}$; $\mu^r_{z_{ir}}(\zeta)$ and $\mu^l_{z_{ir}}(\zeta)$ are the right and left truncated means of z_{ir} with truncation point ζ, respectively; $\sigma^r_{z_{ir}}(\zeta)$ and $\sigma^l_{z_{ir}}(\zeta)$ are the right and left truncated standard deviation of z_{ir} with truncation point ζ, respectively; $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution; $\mu_{\tilde{z}_{ir}}$ is the conditional mean of the denoised wavelet coefficients; and $\mu_{z_{ir}} = \mu_r + b_{ir}$ with μ_r and b_{ir} as the r^{th} element of vectors $\boldsymbol{\mu}$ and $\mathbf{b}_i$, respectively. The detailed derivations for obtaining $\mu_{\tilde{z}_{ir}}$, $\mu^r_{z_{ir}}(\zeta)$, $\mu^l_{z_{ir}}(\zeta)$, $\sigma^r_{z_{ir}}(\zeta)$ and $\sigma^l_{z_{ir}}(\zeta)$ are given in Appendix 3.A.

3.4.2 Characterizing between-profile variation

After denoising thresholding (Section 3.4.1), the true wavelet coefficients $\boldsymbol{\theta}_i$ can be estimated by the remaining denoised wavelet coefficients $\tilde{\mathbf{z}}$, i.e.,

$$\hat{\boldsymbol{\theta}}_i = \tilde{\mathbf{z}}_i \text{ and } \boldsymbol{\theta}_i = \boldsymbol{\mu} + \mathbf{b}_i . \tag{3-4}$$

As a result, the remaining unthresholded empirical wavelet coefficients can be used for modeling the between-profile variation and estimating the remaining unknown parameters of the mixed model. The ML estimate of $\boldsymbol{\mu}$ is obtained by

$$\hat{\boldsymbol{\mu}} = \sum_{i=1}^{m} \tilde{\mathbf{z}}_i / m . \tag{3-5}$$

Let $\boldsymbol{\Sigma}_{\tilde{\mathbf{z}}}$ represent the covariance matrix of the denoised wavelet coefficients. The ML estimate of $\boldsymbol{\Sigma}_{\tilde{\mathbf{z}}}$ is $\hat{\boldsymbol{\Sigma}}_{\tilde{\mathbf{z}}} = \underset{r=1,2,\ldots,n}{diag}\left[\hat{S}_r\right]; \hat{S}_r = \sum_{i=1}^{m} (\tilde{z}_{ir} - \bar{\tilde{z}}_r)^2 / m$, where $\tilde{z}_{ir}$ is the r^{th} denoised wavelet coefficient of the i^{th} profile and $\bar{\tilde{z}}_r$ is the sample mean of the r^{th} denoised wavelet coefficient among all profiles. Furthermore, It is well known that $\boldsymbol{\Sigma}_{\tilde{\mathbf{z}}} = \text{Var}\left\{\text{E}\left[\tilde{\mathbf{z}}|\mathbf{b}_i\right]\right\} + \text{E}\left\{\text{Var}\left[\tilde{\mathbf{z}}|\mathbf{b}_i\right]\right\}$. In other words, in the wavelet domain, the variance obtained from the denoised coefficients of all profiles, $\boldsymbol{\Sigma}_{\tilde{\mathbf{z}}}$, can be decomposed into the between-profile variation, $\text{Var}\left\{\text{E}\left[\tilde{\mathbf{z}}|\mathbf{b}_i\right]\right\}$, and the estimate's variation caused by denoising each profile $\text{E}\left\{\text{Var}\left[\tilde{\mathbf{z}}|\mathbf{b}_i\right]\right\}$. Based on equation (3-4), the term $\text{Var}\left\{\text{E}\left[\tilde{\mathbf{z}}|\mathbf{b}_i\right]\right\}$ can be estimated by $\hat{\boldsymbol{\Lambda}}$, where $\hat{\boldsymbol{\Lambda}}$ is the estimate of the covariance matrix of $\mathbf{b}_i$. Also, the term $\text{E}\left\{\text{Var}\left[\tilde{\mathbf{z}}|\mathbf{b}_i\right]\right\}$ can be estimated by (3-3) provided that $\mu_{z_{ir}}$ and σ are replaced with $\hat{\mu}_r$ and $\hat{\sigma}$, respectively, where $\hat{\mu}_r$ represents the r^{th} element of $\hat{\boldsymbol{\mu}}$. If the estimated $\text{E}\left\{\text{Var}\left[\tilde{\mathbf{z}}|\mathbf{b}_i\right]\right\}$ is denoted by $\hat{\boldsymbol{\Sigma}} = \underset{r=1,2,\ldots,n}{diag}\left[\hat{v}_r\right]$ with $\hat{v}_r$ as the estimate of $\sigma^2_{\tilde{z}_r}$, we can obtain

$\hat{S}_r = \hat{\lambda}_r + \hat{v}_r$. Therefore, $\hat{\lambda}_r$ can be estimated by $\hat{\lambda}_r = (\hat{S}_r - \hat{v}_r) I(\hat{S}_r - \hat{v}_r > 0)$, with $I(\cdot)$ as an indicator function.

3.4.2.1 Selecting wavelet coefficients with significant random effects

To reduce the dimensionality of random effects in the mixed model, only a small number of wavelet coefficients that have larger and more significant random effects, should be selected. For this purpose, two rules are suggested in this for selecting appropriate wavelet coefficients in the mixed model. Rule 1 is used to select wavelet coefficients with larger contributions to the between-profile variance. Rule 2 is used to further check whether the selected coefficients have a significant random effect. The detailed description of each rule is provided below.

Rule 1: Wavelet coefficients with a larger variance are chosen such that the cumulative variance contribution of the selected random effects exceeds a threshold of Q $(0 \le Q \le 1)$. The contribution of each wavelet coefficient as a random effect in the total between-profile variation is sorted by

$$q_r = \frac{\hat{\lambda}_r}{\text{trace}(\hat{\mathbf{\Lambda}})}; \hat{\lambda}_1 > \hat{\lambda}_2 > ... > \hat{\lambda}_n . \tag{3-6}$$

Thus, the set of the selected wavelet coefficients can be represented by $\mathrm{A} = \left\{ \tilde{z}_r \middle| r \le \arg\min_k \left\{ \sum_{d=1}^{k} q_d \ge Q \right\} \right\}$.

Justification of Rule 1: The reason we use this criterion is because we are often interested in identifying the root sources for only the top $100Q$ percent of total variations.

To show how q_r is related to the between-profile variation, let $\mathbf{\Sigma}_{f(\mathbf{t})}$ denote the between-profile covariance matrix. As proved in Appendix 3.B, the total between-profile variance, calculated by $\text{trace}(\mathbf{\Sigma}_{f(\mathbf{t})})$, can be explained by the total variance of the random effects. That is, $\text{trace}(\mathbf{\Sigma}_{f(\mathbf{t})}) = \text{trace}(\mathbf{\Lambda})$. Therefore, the selected wavelet coefficients would be sufficiently described as more than $100Q$ percent of the total between-profile variation. There is no fixed value for threshold Q. The choice of Q is subject to the specific application. Generally, a very small value of Q may result in information loss about between-profile variation. On the other hand, as will be explained in the next section, a very large value of Q may lead to too many selected wavelet coefficients that may affect the performance of the model used for grouping profile data.

Rule 2: Each wavelet coefficient in A is tested to check whether it is a significant random effect.

For this purpose, the following hypothesis is formulated:

$$\begin{cases} \mathrm{H}_0 : \lambda_r = 0 \\ \mathrm{H}_\mathrm{a} : \lambda_r \neq 0 \end{cases}, \quad r \in \mathrm{A}.$$

The statistic $F_r = \dfrac{\hat{\nu}_r}{\hat{S}_r}, r \in \mathrm{A}$ is used for testing the above hypothesis. If the calculated F_r is larger than the critical value F_C, then it can be implied that the random effect corresponding to wavelet coefficient r is significant. The critical value F_C is obtained by the $100(1-\alpha)^{\text{th}}, 0 < \alpha < 1$ percentile of the empirical distribution of F_r when there is no random effect in the model that is obtained by the Monte Carlo simulation.

Justification of Rule 2: As stated earlier, in the wavelet domain, the estimated variance obtained from the denoised coefficients of all profiles can be expressed by the estimated random effect plus the estimate's variation caused by denoising each profile, i.e., $\hat{S}_r = \hat{\lambda}_r + \hat{\nu}_r$. Therefore, when the value of $\hat{\nu}_r$ is close to the value of $\hat{S}_r$, it indicates that the corresponding random effect is not significant.

After using these two rules, the set of selected wavelet coefficients with a significant random effect is denoted as Ω_S, and Ω_U represents the remaining unselected wavelet coefficients. Furthermore, if the distribution of coefficients in Ω_S is identical, these coefficients can be used to estimate the between-profile variation in the wavelet domain. By applying the inverse discrete wavelet transformation (IDTW), we can map the coefficients in Ω_S to the corresponding local segments of profiles, which have significant between-profile variability. The mapped segments of profiles contributing to the between-profile variation about the process can further facilitate the identification of sources of variations along with engineering knowledge. A case study employing this mapping will be presented in Section 3.7.

3.5 LRT based change-point model for clustering profiles

In estimating a mixed model, it is essential to ensure that all the selected profile samples follow an identical distribution. Thus, a change-point model involving LRT (LRT-CP) is applied for checking this condition. To develop the LRT-CP model, it is assumed that if there is a change in the process, it would only affect the mean of the profile distribution and the covariance structure would remain constant. If LRT-CP finds more than one group of profiles, the parameters of each group should be reestimated.

3.5.1 Construction of monitoring features

If all coefficients are directly considered in the LRT-CP, the test power would decrease due to the "curse of dimensionality". To reduce the number of variables involved in LRT-CP, we use the taxonomy of coefficients presented in subsection 3.4.2.1. It is commonly known that the sample covariance matrix is sensitive to the change of sample mean, i.e., if the mean of the process changes globally or locally, the estimated sample covariance matrix is inflated. Therefore, the larger the sample variance of a wavelet coefficient, the more likely the coefficient mean has changed. Thus, coefficients with the random effect in Ω_S are chosen as the monitoring features to be directly used in LRT-CP. In order to further include the information of unselected coefficients in Ω_U, these coefficients are added together as a combined monitoring feature, which is defined as $\gamma_i = \sum_{r \in R} \tilde{z}_{ir}; i = 1,2,...,m$ with $R = \{r; \tilde{z}_{ir} \in \Omega_U\}$. Since the elements of vector $\tilde{\mathbf{z}}$ are normally independently distributed, it yields $\gamma_i \sim N(\sum_{r \in R} \mu_r, \sum_{r \in R} S_r)$. Therefore, if there is a change on the coefficients in Ω_U, it could also be detected by γ_i. If we use $\tilde{\boldsymbol{\gamma}}_i$ to denote all the selected coefficients in Ω_S, the monitoring feature vector can be formed as $\boldsymbol{\gamma}_i = [\tilde{\boldsymbol{\gamma}}_i^T \quad \gamma_i]^T$ with $\boldsymbol{\gamma}_i \sim \mathrm{MVN}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Lambda}_\gamma)$, where $\boldsymbol{\mu}_\gamma$ and $\boldsymbol{\Lambda}_\gamma$ represent the mean vector and covariance matrix of the monitoring feature, respectively.

3.5.2 LRT-based change-point model

Various change-point models have been developed with great successes in analyzing and grouping collected observations. For example, an LRT-based change-point model is developed for univariate normally distributed data (Sullivan and Woodall, 1996) and

multivariate normally distributed data (Worsely, 1979; Sullivan and Woodall, 2000; and Zamba and Hawkins, 2006). Sullivan (2002) developed a different change-point model based on the clustering approach. Recent research on process monitoring using a change-point model has also been explored, such as the change-point method for monitoring linear profiles (Zou et al., 2006 and Mahmoud et al., 2007) and for monitoring nonlinear profiles using wavelet coefficients in Phase-II control charts (Chicken et al., 2009). In addition to the change-point models, there is another set of methods for grouping observations based on the clustering approach (see the work by Fraley and Raftery, 1998; Kothari and Pitts, 1999; Ertoz et al., 2003; and Zhang and Albin, 2007). The LRT-based change-point model, however, is preferred in this for two reasons. First, for the purpose of grouping sequential observations, the change-point models directly utilize the information related to the data sequence order. Second, since the observations are assumed to follow a normal distribution in this , a parametric model is preferred over a nonparametric model. It is well known that the LRT-based change-point model is a commonly-used parametric model-based approach. We extend the existing change-point models by including the wavelet coefficients with random effect in order to account for the between-profile variability.

Suppose the mean of the profiles changes at an unknown time τ. Assuming the covariance matrix of nonlinear profiles is constant, the distribution of $\boldsymbol{\gamma}_i$ is written as

$$\boldsymbol{\gamma}_i \sim \begin{cases} \mathrm{MVN}(\boldsymbol{\mu}_\gamma^0, \boldsymbol{\Lambda}_\gamma) & i \le \tau \\ \mathrm{MVN}(\boldsymbol{\mu}_\gamma^1, \boldsymbol{\Lambda}_\gamma) & i > \tau \end{cases}, \tag{3-7}$$

where $\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{0}$ and $\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{1}$ represent the mean vectors before and after the change, respectively. If all profiles are from an identical distribution, then τ does not exist and $\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{0} = \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{1}$. Therefore, in order to check whether the collected profile data have an identical distribution, the following hypothesis test can be evaluated:

$$\begin{cases} \mathrm{H}_0 : \mathrm{E}\left[\boldsymbol{\gamma}_i\right]_{i\le\tau} = \mathrm{E}\left[\boldsymbol{\gamma}_i\right]_{i>\tau} = \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{0} \\ \mathrm{H}_{\mathrm{a}} : \mathrm{E}\left[\boldsymbol{\gamma}_i\right]_{i\le\tau} \neq \mathrm{E}\left[\boldsymbol{\gamma}_i\right]_{i>\tau} = \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{1} \end{cases}, \quad \tau = 1,2,...,m, \tag{3-8}$$

where $\mathrm{E}[\cdot]$ is the mathematical expectation operator. Here, a likelihood ratio test is utilized for evaluating this hypothesis test. As derived in Appendix 3.C, the log likelihood ratio can be expressed as

$$\Gamma(\tau) = \frac{\tau(m-\tau)}{m} \sum_{i=\tau+1}^{m} (\hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^{1} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^{0})^{T} \, \hat{\boldsymbol{\Lambda}}_{\boldsymbol{\gamma}}^{-1} \, (\hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^{1} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^{0}), \tag{3-9}$$

where $\hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^{0} = \sum_{i=1}^{\tau} \boldsymbol{\gamma}_i / \tau$, $\hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^{1} = \sum_{i=\tau+1}^{m} \boldsymbol{\gamma}_i / (m-\tau)$, and

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\gamma}} = \left\{ \sum_{i=1}^{\tau} (\boldsymbol{\gamma}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^{0})(\boldsymbol{\gamma}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^{0})^{T} + \sum_{i=\tau+1}^{m} (\boldsymbol{\gamma}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^{1})(\boldsymbol{\gamma}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^{1})^{T} \right\} \Big/ (m-2).$$

The values of $\Gamma(\tau);\ (\tau = 1,2,...,m)$ are compared with a limit of *L*. If $\underset{\tau=1,2,..m}{\Gamma(\tau)} < L$, all profiles are from an identical distribution. Otherwise, they can be categorized into different clusters and the value of τ maximizing $\Gamma(\tau)$ is an estimate of the change-point i.e., $\hat{\tau} = \underset{\tau=1,2,...,m}{\arg\max}\, \Gamma(\tau)$. If the LRT-CP detects a change, the profile samples can be separated into two groups based on the estimated change-point. Then, LRT-CP should be

applied for each group to check whether more clusters exist in each group. Using this approach, one can categorize multiple change-points and clusters. Additionally, knowing the estimated time at which the process changed could help process engineers effectively detect the root cause(s) of the change and identify the corresponding source(s) of variations. It should be noted that the LRT-CP model is often used to detect single and/or multiple sustained shifts in the historical profile data. If one is interested in identifying outlier profiles, other methods such as multivariate T^2 control charts with a robust estimator of the covariance matrix (Vargas, 2003), can be utilized to examine the selected wavelet coefficients with random effect and detect outliers.

The limit L could be determined based on the desired Type-I error (α). In this , L is determined via simulations since the values of $\Gamma(\tau)$ are not independent.

3.6 Performance evaluation using simulations

In this section, the performance of the proposed approach is evaluated through the Monte Carlo simulations. This is accomplished in two stages. First, the performance of the proposed mixed-effect LRT-CP model (denoted by "M") is assessed under different change scenarios and compared with another wavelet-based method recently proposed by Chicken et al. (2009) (denoted by "C"), in which the between-profile variation is not considered. As mentioned earlier, if the collected profiles do not follow the same mixed model, the estimation results are misleading. In this , two criteria are used for the performance evaluation: probability of detecting change in the mean $(1-\beta)$ and the average estimation of change time ($\bar{\hat{\tau}}$). In the other stage, the accuracy of the proposed approach for estimating the between-profile variation is checked. This is evaluated based

on the ratio of the estimated to actual random effects standard deviations, denoted by $\hat{\lambda}/\lambda$. Since the case of multiple changes can be boiled down to a single change case, only single change is studied here.

To simulate profiles, the popularly used piecewise smooth function of Mallat (1999) is utilized for generating simulated profiles, as shown in Figure 3-4. This simulated function is a complicated function with several nondifferentiable points that cannot be easily modeled by parametric models or other nonparametric models such as splines. Additionally, Chicken et al. (2009) used this function to evaluate the performance of their monitoring procedure. Also, it is assumed that the between-profile variations only occur at three segments: I_1=[32,55], I_2=[146,153], and I_3=[207,236] as shown in Figure 3-4, which cover 62 data points, or 24% of the entire profile. In those segments, each response $y_{ir}; i = 1,2,...m, r = 1,2,...,n$ is generated based on $y_{ir} = f(t_r) + b_{ir}^f$, where $f(t_r)$ is the value of the Mallat's function at t_r, and $b_{ir}^f \sim N\left(0, s^2 f^2(t_r)\right)$ is the random effects with the coefficient of variation s, that sets the standard deviation of each response y_{ir} to be proportional to the value of its mean $f(t_r)$. Finally, to include the within-profile variation, normally independently distributed noises with mean zero and variance $\sigma^2 = 1$ are added to y_{ir}. Figure 3-4 shows the 50 simulated profiles with n=256, $t \in [1,256]$, and s=0.2. An alternative procedure that can be used for simulating random profiles is generating both within- and between-profile variations on the wavelet coefficients, and then transforming the coefficients back to the original domain using IDWT. In this , we prefer to use the first procedure since we are interested in modeling and characterizing variations of original profiles. For the comparison purpose, similar to Chicken et al.

(2009), the Haar basis with the complete decomposition (i.e., 8 levels of decomposition) is used to develop the wavelet-based mixed model.

In order to assess the capability of the methods in detecting different types of profile changes, three change scenarios with different magnitudes are examined:

Scenario 1: overall mean change, where the whole profile is shifted vertically, that is, $\boldsymbol{\mu}_{\gamma}^{1} = \boldsymbol{\mu}_{\gamma}^{0} + \delta(s\boldsymbol{\mu}_{\gamma}^{0} + \sigma)$.

Scenario 2: local mean change in the segments of [41,46] and [208,215] with $\boldsymbol{\mu}_{\gamma}^{1} = \boldsymbol{\mu}_{\gamma}^{0} + \delta(s\boldsymbol{\mu}_{\gamma}^{0} + \sigma)$. These two segments contain the between-profile variation.

Scenario 3: local mean change in the segments of [6,22], [89,106], and [129,145] with $\boldsymbol{\mu}_{\gamma}^{1} = \boldsymbol{\mu}_{\gamma}^{0} + \delta\sigma$. These three segments are not comprised of between-profile variations.

Furthermore, to investigate the performance sensitivity to the parameters of *m* and τ, the simulations are carried out for *m*=75 and 150; $\tau = 0.5m$ and $0.8m$ with 1000 replications. Also, *Q*=0.80 is used for all simulation scenarios.

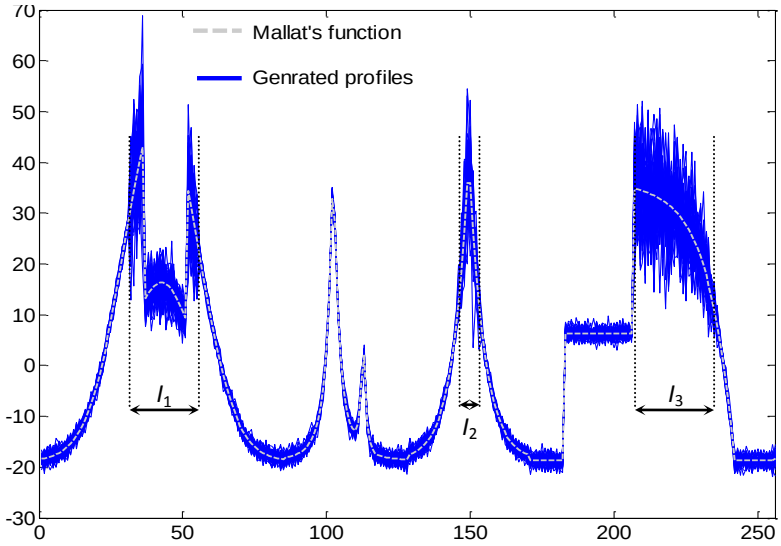

Figure 3-4. Mallat's function and randomly generated profiles

To compare the performance of the methods, the limit L is chosen so that the estimated probability of false signal (α) is approximately equal to 0.05. In the proposed method, the probability of the signal is estimated by the proportion of simulation runs where at least one of $\Gamma(\tau);\tau=1,2,...,m$ is plotted beyond L. So, the 95^{th} percentile of $\max\Gamma(\tau);\tau=1,2,...,m$ obtained from 1000 simulation runs is chosen as the limit L. It is clear that for a specific α, the value of L depends on the number of collected profiles m and the dimension of γ_i, denoted as p. The estimated values of L based on 1000 simulations for $\alpha=0.05$ under different m and p are provided in Table 3-1.

The estimated detection probabilities for different change scenarios are shown in Figure 3-5. As can be seen from these figures, under all scenarios and different parameters of m and τ, the proposed mixed LRT-CP method outperforms Chicken's method. This is because the LRT-CP model accounts for the between-profile variation, while the other does not. Moreover, the detection probability of LRT-CP is improved as the number of sampled profiles (m) increases. Clearly, by increasing m, the estimation of

parameters in the mixed model becomes more precise, thus resulting in this improvement. Also, the performance of LRT-CP is better when $\tau = 0.5m$ than when $\tau = 0.8m$. This is because in the case of $\tau = 0.5m$, there are equal samples available in each group to estimate the parameters, which leads to a better estimation for the two groups on average.

Table 3-1. *L* values for different *p* and *m* with α=0.05

	p				
m	5	10	15	20	25
75	25.81	41.38	58.70	80.12	108.22
100	24.73	38.77	52.08	67.62	86.72
125	24.40	36.27	49.31	62.39	78.12
150	23.84	35.55	47.99	58.90	73.84
175	23.16	34.94	45.50	56.54	70.51
200	23.30	34.64	45.24	55.81	66.10
250	23.70	33.56	42.55	53.42	63.68
300	22.16	33.58	43.73	52.53	62.32
350	22.96	32.86	43.64	52.12	61.99
400	23.67	33.30	42.65	51.74	60.76
450	23.21	33.94	41.87	50.67	59.76
500	23.03	32.99	42.43	51.49	59.77
	p				
m	30	35	40	45	50
75	140.97	187.64	254.12	351.40	529.93
100	106.90	130.81	164.97	206.79	256.82
125	93.92	111.80	133.16	159.12	186.31
150	86.49	102.60	118.03	139.07	157.97
175	82.54	96.46	110.71	125.19	140.76
200	78.13	89.66	104.15	118.21	132.76
250	73.70	86.45	98.34	108.15	120.35
300	72.05	82.87	94.42	105.10	115.19
350	70.89	79.92	90.93	101.09	112.71
400	70.47	78.77	89.99	97.83	106.43
450	69.56	77.73	88.48	97.94	105.58
500	67.41	76.83	87.27	96.02	104.79

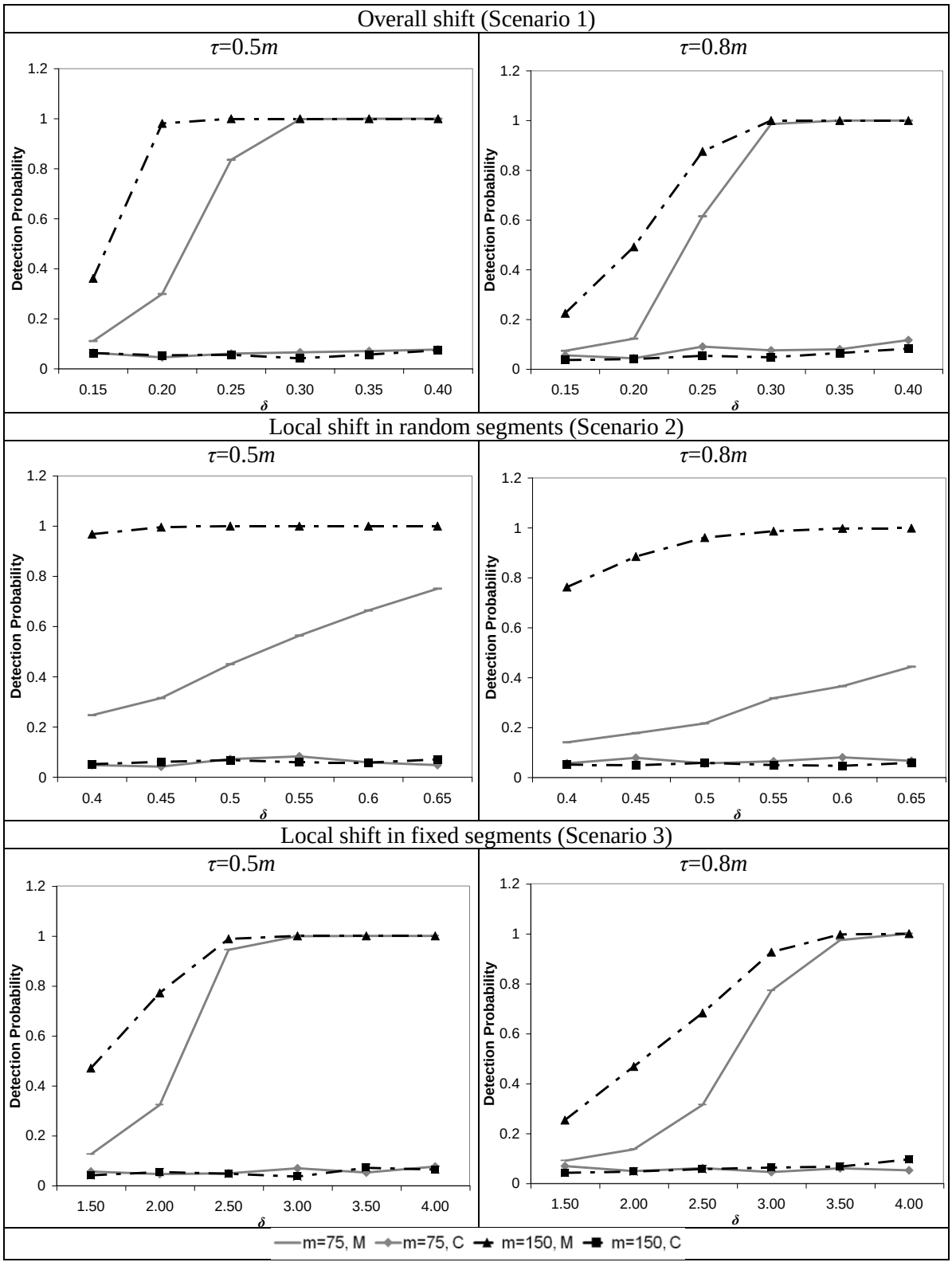

Figure 3-5. Detection probability of methods "M" and "C" under different shift scenarios.

Furthermore, the performance of change-point estimation is also studied. The mean value and the standard error (given in the parentheses) for each estimated τ by using the LRT-CP method are shown in Table 3-2. As an example, under Scenario 2 with *m*=150, $\tau = 75$, and $\delta = 0.60$, the mean of the estimated change-point is 74.97 with the standard error of 0.07. Similar to the effect of τ and *m* on the detection performance of LRT-CP, the performance of the change-point estimator also improves when *m* increases and/or when $\tau = 0.5m$. Generally, the higher the detection probability of LRT-CP, the more accurate and precise the change-point estimator will be. Based on Table 3-2, it can also be seen that the absolute biases of the estimated change-points are all less than 5, except for in a few cases (where values are underlined). Therefore, the estimation performance of change-points by using LRT-CP is quite reasonable.

Furthermore, in order to investigate the performance of the proposed approach in estimating the between-profile variation, the ratio of $\hat{\lambda}_r / \lambda_r ; r = 1,2,...,62$ is calculated for every point within all three random-effect segments [I_1, I_2, I_3]. In order to show the estimation performance, we use the median of all $\hat{\lambda}_r / \lambda_r$ ratios to assess the average estimation performance, and use the third and the first quartile values of $\hat{\lambda}_r / \lambda_r$, respectively denoted by Q_3 and Q_1, to show the estimation uncertainty. The results under different change scenarios with $\tau = 0.5m$ are presented in Figure 3-6. The values of $\hat{\lambda}_r / \lambda_r$ greater and less than 1 imply overestimation and underestimation, respectively, while values close to 1 indicate the unbiased estimates. From Figure 3-6, it is clear that median of $\hat{\lambda}_r / \lambda_r$ are close to 1, which shows that they have a very good average estimation performance. In the case of *m*=150, the estimates are more stable than those of *m*=75 across all δ values. Therefore, the standard deviation estimates become steadier

when m is large. Moreover, when m=75, the accuracy of the estimates becomes more stable as δ increases. The reason for this is that the detection probabilities of change-points increase as δ increases, thus resulting in the better change-point estimates.

Table 3-2. Mean and standard error of estimated change-point of LRT-CP under different scenarios

Overall Shift (Scenario 1)	***m***	***τ***	***δ***					
			0.15	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**
	75	**0.5*m*=38**	40.32	39.03	38.43	38.24	38.10	38.02
			(0.59)	(0.36)	(0.17)	(0.06)	(0.02)	(0.01)
		0.8*m*=60	43.86	52.50	58.98	60.18	60.07	60.02
			(0.72)	(0.61)	(0.25)	(0.03)	(0.01)	(0.01)
	150	**0.5*m*=75**	76.43	76.35	75.61	75.19	75.04	75.00
			(0.79)	(0.21)	(0.06)	(0.02)	(0.01)	(0.00)
		0.8*m*=120	104.79	111.83	118.87	120.12	120.05	120.00
			(1.18)	(0.85)	(0.35)	(0.01)	(0.01)	(0.01)
Local shift in random segments (Scenario 2)	***m***	***τ***	***δ***					
			0.40	**0.45**	**0.50**	**0.55**	**0.60**	**0.65**
	75	**0.5*m* =38**	37.73	37.41	38.34	37.46	37.75	37.86
			(0.44)	(0.35)	(0.31)	(0.27)	(0.24)	(0.21)
		0.8*m*=60	52.21	53.11	56.11	55.81	56.33	57.13
			(0.56)	(0.52)	(0.41)	(0.42)	(0.4)	(0.32)
	150	**0.5*m* =75**	74.77	75.00	75.25	74.97	75.04	75.06
			(0.25)	(0.17)	(0.12)	(0.08)	(0.07)	(0.06)
		0.8*m*=120	117.19	118.60	119.25	119.48	119.86	120.03
			(0.51)	(0.35)	(0.25)	(0.23)	(0.10)	(0.08)
Local shift in fixed segments (Scenario 3)	***m***	***τ***	***δ***					
			1.50	**2.00**	**2.50**	**3.00**	**3. 50**	**4.00**
	75	**0.5*m*=38**	37.58	39.11	38.17	38.03	38.00	38.00
			(0.57)	(0.39)	(0.1)	(0.02)	(0.01)	(0.002)
		0.8*m*=60	40.03	51.63	57.92	59.19	59.85	59.99
			(0.79)	(0.55)	(0.31)	(0.19)	(0.10)	(0.01)
	150	**0.5*m*=75**	74.39	75.27	75.20	75.03	75.00	75.00
			(0.70)	(0.46)	(0.11)	(0.01)	(0.00)	(0.00)
		0.8*m*=120	106.46	114.85	116.68	119.48	120.01	120.00
			(1.15)	(0.80)	(0.58)	(0.21)	(0.03)	(0.004)

From Figure 3-6, it can be observed that the variances of random effects could sometimes be overestimated under the scenario with a small shift, which is due to the poor performance in detecting and estimating the change point with a small shift. With the increase of the shift magnitude δ, the estimation performance is improved and the estimated variance approaches the true variance (i.e., the ratio of $\hat{\lambda}_i / \lambda_i$ is closer to 1, as shown in Figure 3-6). However, the stable value of $\hat{\lambda}_i / \lambda_i$ is generally less than 1 because in our mixed model, only a subset of random-effect coefficients is selected to explain the $100Q\%$ of the total between-profile variation. Therefore, without the effect of the estimation error of change-points, the estimated variance should be less than the true total variance.

In short, the simulation results show that the proposed methodology has reasonable performance in classifying different groups of profiles as well as in characterizing the variance of each group of profiles.

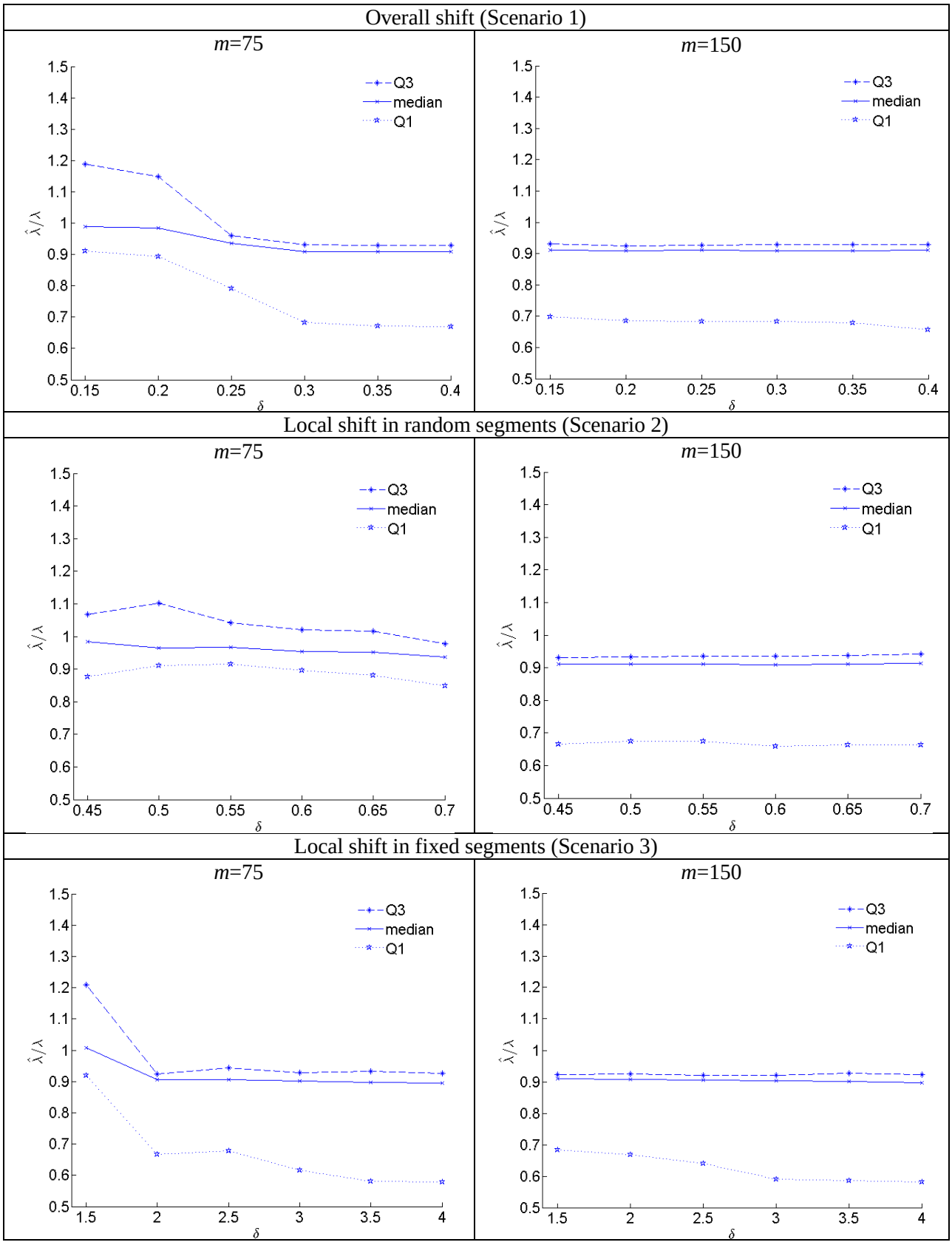

Figure 3-6. Q_1, Median, and Q_3 of $\hat{\lambda}_i/\lambda_i$ under different shift scenarios.

3.7 Case study

In this section, the proposed methodology is applied to real-world profile data, which are collected during valve seat pressing operations in an engine head assembly process. At every cycle of the pressing operations, each valve seat is pressed into a seat counterbore pocket of a cylinder head, which generates one cycle of press force signals as a sample of profile data. Pictures of the engine head (upper left), valve seat pocket (lower left), and pressing machine (right) are shown in Figure 3-7. In this process, one of the important quality characteristics is the gap between the seat bottom and the pocket. However, there is no automatic sensing technology for directly measuring that gap during production. Another aspect to take into account is that the product quality is very sensitive to the pressing force on the ram, which can be measured online by the load sensors installed on a pressing machine. Therefore, pressing force signals are often used for process control (i.e., reduction of the variation of pressing force signals will lead to the improvement of product quality). In this case study, 50 force profiles are collected for process variation evaluation by the following analysis (as shown in Figure 3-8).

A mixed model is developed to characterize the process variation according to the proposed methodology given in Figure 3-3. Based on Section 3.4.2.1, Q=0.80 is used as the threshold for selecting the wavelet coefficients with random-effect in Ω_S. The LRT-CP model presented in Section 3.5.2 is also used to check whether all 50 profiles follow an identical distribution. The LRT-CP result indicates a change-point of $\hat{\tau} = 42$, thus these 50 profiles are clustered into two groups. In Figure 3-8, the profiles corresponding to clusters 1 and 2 are plotted with a solid line and a dashed line, respectively. LRT-CP is also applied to the profiles within each group, but it does not find a new change-point

within each group of profiles. Since the number of profiles in cluster 2 is not large enough, only cluster 1 is used for further identification of the critical segments with a large variability. A similar method can be applied once more profiles are collected for cluster 2.

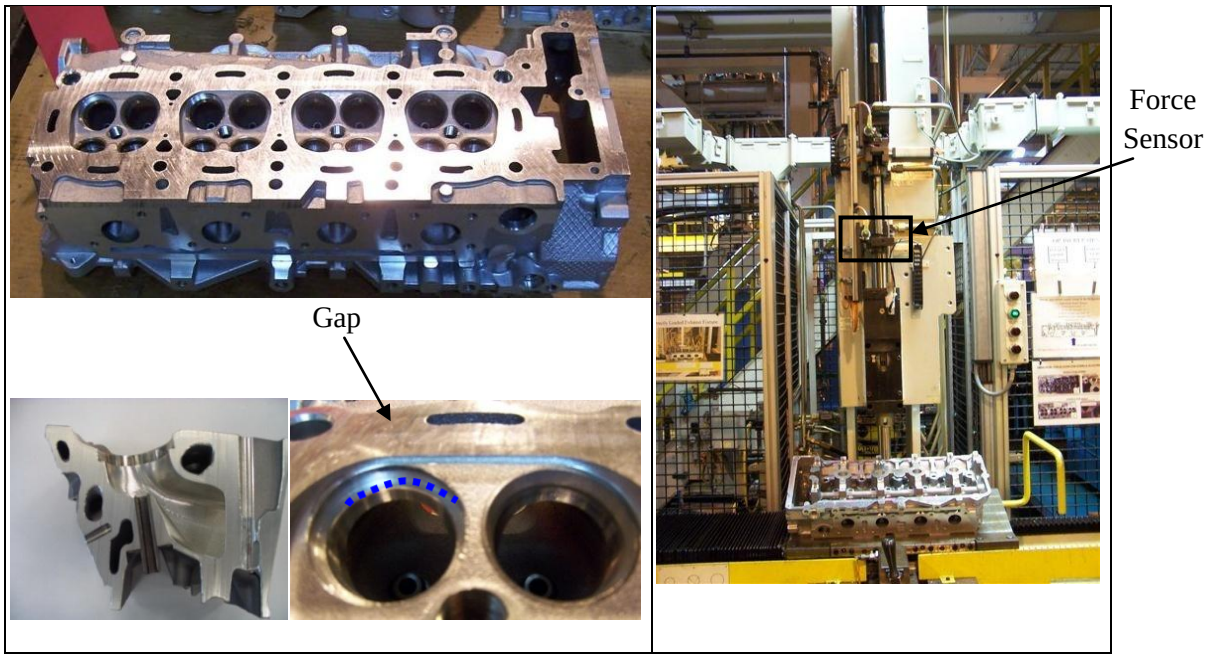

Figure 3-7. Engine head, cross-section view of valve seat pocket, and gap between valve seat and pocket (left panel), valve seat assembly process (right panel).

In order to identify sources of variations of cluster 1, a mixed model is constructed. Based on the fitted mixed model, the estimated within-profile variance ($\hat{\sigma}^2$) is equal to 82.28. The wavelet coefficients with random-effect, identified in Ω_S include $[c_{5,16}, c_{5,20}, c_{5,19}, c_{5,21}, c_{5,17}, c_{5,15}, c_{5,28}, c_{5,22}, c_{5,23}, c_{5,18}, c_{5,29}, d_{5,15}]^T$ with the descending order of their variances. Then, the mapping of these coefficients in Ω_S to the associated profile segments is conducted using IDWT. These mapped segments, along with their corresponding coefficients, are shown in Figure 3-8. There are three segments contributing to 80% of the total between-profile variation. Table 3-3 summarizes the

information about each segment and the corresponding between-profile variances. In Table 3-3, we see that the between-profile variance is much larger than the within-profile variance, which implies that the sources of variations causing between-profile variations are more important for process improvement. The wavelet coefficients with random-effect along with the estimated mean and variance (reported in Table 3-3) can be further served as the basis to implement control charts for process monitoring.

Table 3-3. Summary information of the fitted mixed model.

Segment	Corresponding profile observations	Corresponding wavelet coefficients	Wavelet coefficients mean (fixed effect)	Wavelet random effect variances	Average of segmental between-profile variance
A	y_{57}- y_{73}	$c_{5,16}$	5.736E+02	1.1163E+05	1.7026E+04
		$c_{5,19}$	1.076E+03	5.9810E+04	
		$c_{5,17}$	8.129E+02	4.8550E+04	
		$c_{5,15}$	1.238E+02	4.1900E+04	
		$c_{5,18}$	8.951E+02	2.4110E+04	
		$d_{5,19}$	-8.410E+01	2.0460E+04	
B	y_{74}- y_{92}	$c_{5,20}$	1.362E+03	6.8650E+04	1.1852E+04
		$c_{5,21}$	1.525E+03	5.0160E+04	
		$c_{5,22}$	1.538E+03	3.4070E+04	
		$c_{5,23}$	1.464E+03	2.4350E+04	
		$c_{5,19}$	1.076E+03	5.9810E+04	
C	y_{109}- y_{116}	$c_{5,28}$	1.998E+03	3.8790E+04	7.7238E+03
		$c_{5,29}$	3.651E+03	2.3000E+04	

Furthermore, based on the segments obtained from IDWT and engineering knowledge, the sources contributing to the between-profile variations can be identified. The variation in “segment a” is due to the position variations of engine head surfaces. This source is mainly related to the variation of initial contacting points induced by the

variation of an engine head's pocket depth due to previous manufacturing stages. The clearance tolerance between the valve seat and the seat packet is the major source of variations for "segment b." The pressure variation of the assembly machine could be causing the force signal variation in "segment c." The first two sources of the variations are considered part-to-part variations, but the source for "segment a" is related to the process variation at previous manufacturing stages while the source for "segment b" is related to the current assembly process variation. The average between-profile variance for each segment can also be obtained by adding up the variance of wavelet coefficients in each segment and dividing the sum by the length of the segment. These values, reported in Table 3-3, can be used to prioritize further actions for variation reduction and process improvement.

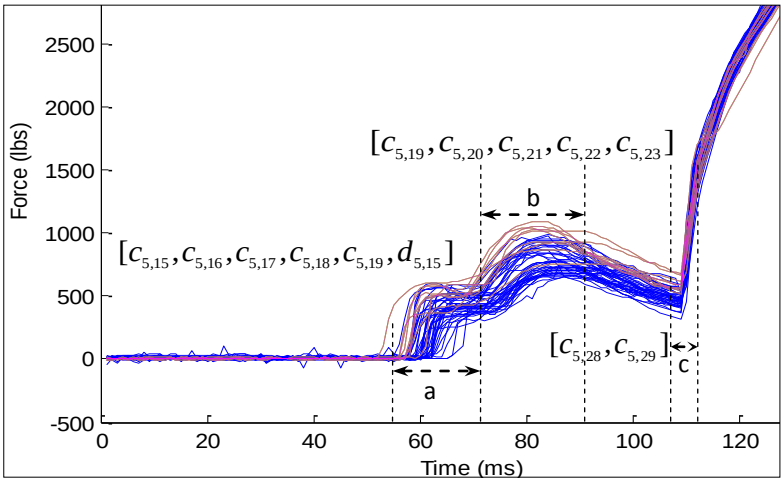

Figure 3-8. Force vs. time profiles.

Appendix 3.A: Derivation of $\mu_{\tilde{z}_{ir}}$ and $\sigma^2_{\tilde{z}_{ir}}$

First, we derive the conditional mean of denoised coefficients $\mu_{\tilde{z}_{ir}}$

$$
\begin{aligned}
&\mu_{\tilde{z}_{ir}} = \mathrm{E}\left[\tilde{z}_{ir}\middle|b_{ir}\right] = \mathrm{E}\left[\mathrm{sign}\left(z_{ir}\right)\left(\left|z_{ir}\right|-\zeta\right)I\left(\left|z_{ir}\right|>\zeta\right)\middle|b_{ir}\right] \\
&= \mathrm{E}\left[\mathrm{sign}\left(z_{ir}\right)\left(\left|z_{ir}\right|-\zeta\right)I\left(\left|z_{ir}\right|>\zeta\right)\middle|\left(\left|z_{ir}\right|<\zeta\right),b_{ir}\right]\Pr(\left|z_{ir}\right|<\zeta\middle|b_{ir}) + \\
&\quad \mathrm{E}\left[\mathrm{sign}\left(z_{ir}\right)\left(\left|z_{ir}\right|-\zeta\right)I\left(\left|z_{ir}\right|>\zeta\right)\middle|\left(\left|z_{ir}\right|>\zeta\right),b_{ir}\right]\Pr(\left|z_{ir}\right|>\zeta\middle|b_{ir}) \\
&= \mathrm{E}\left[\left(z_{ir}-\zeta\right)\middle|z_{ir}>\zeta,b_{ir}\right]\Pr(z_{ir}>\zeta\middle|b_{ir}) + \mathrm{E}\left[\left(z_{ir}+\zeta\right)\middle|z_{ir}<-\zeta,b_{ir}\right]\Pr(z_{ir}<-\zeta\middle|b_{ir}).
\end{aligned}
\tag{3.A.1}
$$

Since $z_{ir}|b_{ir} \sim N(\mu_{z_{ir}},\sigma^2)$, the random variable $\left(z_{ir}\middle|z_{ir}>\zeta,b_{ir}\right)$ follows a right truncated normal distribution with parameters $(\mu_{z_{ir}},\sigma^2,\zeta)$. Similarly, $\left(z_{ir}\middle|z_{ir}<-\zeta,b_{ir}\right)$ follows a left truncated normal distribution with parameters $(\mu_{z_{ir}},\sigma^2,-\zeta)$. Therefore, (3.A.1) can be written as

$$
\mu_{\tilde{z}_{ir}} = \left(\mu^r_{z_{ir}}(\zeta)-\zeta\right)\Phi\left(\frac{\mu_{z_{ir}}-\zeta}{\sigma}\right) + \left(\mu^l_{z_{ir}}(-\zeta)+\zeta\right)\Phi\left(\frac{-\mu_{z_{ir}}-\zeta}{\sigma}\right) \tag{3.A.2}
$$

where, $\mu^r_{z_{ir}}(\cdot)$ and $\mu^l_{z_{ir}}(\cdot)$, respectively are the right and left truncated means of z_{ir} with truncation point $(\cdot)$, and can be calculated by

$$
\mu^r_{z_{ir}}(\zeta) = \mu_{z_{ir}} + \frac{\phi\left((\mu_{z_{ir}}-\zeta)/\sigma\right)}{\Phi\left((\mu_{z_{ir}}-\zeta)/\sigma\right)}\sigma, \text{ and } \mu^l_{z_{ir}}(-\zeta) = \mu_{z_{ir}} + \frac{-\phi\left((-\mu_{z_{ir}}-\zeta)/\sigma\right)}{\Phi\left((-\mu_{z_{ir}}-\zeta)/\sigma\right)}\sigma, \tag{3.A.3}
$$

where $\phi(\cdot)$ is the probability distribution function of a normal standard random variable (Johnson and Kotz, 1970). The conditional variance $\sigma^2_{\tilde{z}_{ir}}$ is also obtained based on the derived $\mu_{\tilde{z}_{ir}}$.

$$\sigma^2_{\tilde{z}_{ir}} = \mathrm{E}\left[\tilde{z}^2_{ir}|b_{ir}\right] - \mu^2_{\tilde{z}_{ir}} = \mathrm{E}\left[(z_{ir}-\zeta)^2|z_{ir}>\zeta, b_{ir}\right]\Pr(z_{ir}>\zeta|b_{ir}) +$$
$$\mathrm{E}\left[(z_{ir}+\zeta)^2|z_{ir}<-\zeta, b_{ir}\right]\Pr(z_{ir}<-\zeta|b_{ir}) - \mu^2_{\tilde{z}_{ir}}$$
$$= \left\{\mathrm{E}^2\left[(z_{ir}-\zeta)|z_{ir}>\zeta, b_{ir}\right] + \mathrm{Var}\left[(z_{ir}-\zeta)|z_{ir}>\zeta, b_{ir}\right]\right\}\Pr(z_{ir}>\zeta|b_{ir}) +$$
$$\left\{\mathrm{E}^2\left[(z_{ir}+\zeta)|z_{ir}<-\zeta, b_{ir}\right] + \mathrm{Var}\left[(z_{ir}+\zeta)|z_{ir}<-\zeta, b_{ir}\right]\right\}\Pr(z_{ir}<-\zeta|b_{ir}) - \mu^2_{\tilde{z}_{ir}}$$
$$= \left\{\left(\mu^r_{z_{ir}}(\zeta)-\zeta\right)^2 + \left(\sigma^r_{z_{ir}}(\zeta)\right)^2\right\}\Phi\left(\frac{\mu_{z_{ir}}-\zeta}{\sigma}\right) + \left\{\left(\mu^l_{z_{ir}}(-\zeta)+\zeta\right)^2 + \left(\sigma^l_{z_{ir}}(-\zeta)\right)^2\right\}\Phi\left(\frac{-\mu_{z_{ir}}-\zeta}{\sigma}\right) - \mu^2_{\tilde{z}_{ir}},$$

where, $\left(\sigma^r_{z_{ir}}(\cdot)\right)^2$ and $\left(\sigma^l_{z_{ir}}(\cdot)\right)^2$, respectively are the right and left truncated varainces of z_{ir} with truncation point $(\cdot)$, and can be calculated by

$$\left(\sigma^r_{z_{ir}}(\zeta)\right)^2 = \sigma^2\left[1 - \frac{\frac{\mu_{z_{ir}}-\zeta}{\sigma}\phi\left(\frac{\mu_{z_{ir}}-\zeta}{\sigma}\right)}{\Phi\left(\frac{\mu_{z_{ir}}-\zeta}{\sigma}\right)} - \left(\frac{\phi\left(\frac{\mu_{z_{ir}}-\zeta}{\sigma}\right)}{\Phi\left(\frac{\mu_{z_{ir}}-\zeta}{\sigma}\right)}\right)^2\right], \text{ and}$$

$$\left(\sigma^l_{z_{ir}}(\zeta)\right)^2 = \sigma^2\left[1 - \frac{\frac{-\mu_{z_{ir}}-\zeta}{\sigma}\phi\left(\frac{-\mu_{z_{ir}}-\zeta}{\sigma}\right)}{\Phi\left(\frac{-\mu_{z_{ir}}-\zeta}{\sigma}\right)} - \left(\frac{\phi\left(\frac{-\mu_{z_{ir}}-\zeta}{\sigma}\right)}{\Phi\left(\frac{-\mu_{z_{ir}}-\zeta}{\sigma}\right)}\right)^2\right]. \qquad (3.A.4)\blacksquare$$

Appendix 3.B: Proof of $\mathrm{trace}(\boldsymbol{\Sigma}_{\tilde{\mathbf{z}}}) \approx \mathrm{trace}(\boldsymbol{\Sigma}_{f(\mathbf{t})})$.

It is known that $f(\mathbf{t})$ can be obtained by applying IDWT to true wavelet coefficients, i.e.,

$$f(\mathbf{t}) = \mathbf{W}^{-1}\boldsymbol{\theta} \qquad (3.B.1)$$

Therefore, the covariance matrix of $f(\mathbf{t})$ can be expressed by

$$\boldsymbol{\Sigma}_{f(\mathbf{t})} = \mathbf{W}^{-1}\mathrm{Var}(\boldsymbol{\theta})\left(\mathbf{W}^{-1}\right)^T = \mathbf{W}^{-1}\mathrm{Var}(\boldsymbol{\mu}+\mathbf{b})\left(\mathbf{W}^{-1}\right)^T = \mathbf{W}^{-1}\boldsymbol{\Lambda}\left(\mathbf{W}^{-1}\right)^T \qquad (3.B.2)$$

This is true because $\boldsymbol{\mu}$ is deterministic. Also, since $\mathbf{W}$ is an orthogonal wavelet basis, it can be implied $\left(\mathbf{W}^{-1}\right)^T = \mathbf{W}$.

After taking the trace(.), it yields

$$\text{trace}(\boldsymbol{\Sigma}_{f(\mathbf{t})}) = \text{trace}(\mathbf{W}^{-1}\boldsymbol{\Lambda}\mathbf{W}) = \text{trace}(\mathbf{W}^{-1}\mathbf{W}\boldsymbol{\Lambda}) = \text{trace}(\boldsymbol{\Lambda}) \qquad (3.B.3)\blacksquare$$

Appendix 3.C: Derivation of the likelihood ratio test statistic

The log likelihood function under the alternative hypothesis in (3-8) can be written as:

$$\begin{aligned} l_1 &= \log\left\{\prod_{i=1}^{\tau} h\left(\boldsymbol{\gamma}_i; \boldsymbol{\mu}_{\gamma}^0, \boldsymbol{\Lambda}_{\gamma}\right)\prod_{i=\tau+1}^{m} h\left(\boldsymbol{\gamma}_i; \boldsymbol{\mu}_{\gamma}^1, \boldsymbol{\Lambda}_{\gamma}\right)\right\} \\ &= -mc/2\log(2\pi) - m/2\log(\left|\boldsymbol{\Lambda}_{\gamma}\right|) - 1/2\sum_{i=1}^{\tau}(\boldsymbol{\gamma}_i - \boldsymbol{\mu}_{\gamma}^0)^T \boldsymbol{\Lambda}_{\gamma}^{-1}(\boldsymbol{\gamma}_i - \boldsymbol{\mu}_{\gamma}^0) \qquad (3.C.1) \\ &-1/2\sum_{i=\tau+1}^{m}(\boldsymbol{\gamma}_i - \boldsymbol{\mu}_{\gamma}^1)^T \boldsymbol{\Lambda}_{\gamma}^{-1}(\boldsymbol{\gamma}_i - \boldsymbol{\mu}_{\gamma}^1), \end{aligned}$$

where $h(.)$ is the multivariate normal probability distribution function, and c is equal to the cardinality of $\boldsymbol{\Omega}_s$.

Under H_0, the corresponding log likelihood function would be

$$\begin{aligned} l_2 &= \log\left\{\prod_{i=1}^{m} h\left(\boldsymbol{\gamma}_i; \boldsymbol{\mu}_{\gamma}^0, \boldsymbol{\Lambda}_{\gamma}\right)\right\} = -mc/2\log(2\pi) - m/2\log(\left|\boldsymbol{\Lambda}_{\gamma}\right|) \\ &-1/2\sum_{i=1}^{m}(\boldsymbol{\gamma}_i - \boldsymbol{\mu}_{\gamma}^0)^T \boldsymbol{\Lambda}_{\gamma}^{-1}(\boldsymbol{\gamma}_i - \boldsymbol{\mu}_{\gamma}^0). \end{aligned} \qquad (3.C.2)$$

The maximum likelihood estimators for mean parameters are $\hat{\boldsymbol{\mu}}_{\gamma}^0 = \sum_{i=1}^{\tau}\boldsymbol{\gamma}_i \big/ \tau$ and $\hat{\boldsymbol{\mu}}_{\gamma}^1 = \sum_{i=\tau+1}^{m}\boldsymbol{\gamma}_i \big/ (m-\tau)$, respectively. The estimate of $\boldsymbol{\Lambda}_{\gamma}$ is the pooled-sample covariance

matrix, i.e., $\hat{\boldsymbol{\Lambda}}_\gamma = \left\{ \sum_{i=1}^{\tau} (\boldsymbol{\gamma}_i - \hat{\boldsymbol{\mu}}_\gamma^0)(\boldsymbol{\varphi}_i - \hat{\boldsymbol{\mu}}_\gamma^0)^T + \sum_{i=\tau+1}^{m} (\boldsymbol{\gamma}_i - \hat{\boldsymbol{\mu}}_\gamma^1)(\boldsymbol{\gamma}_i - \hat{\boldsymbol{\mu}}_\gamma^1)^T \right\} \Big/ (m-2)$. Replacing $\boldsymbol{\mu}_\gamma^1$, $\boldsymbol{\mu}_\gamma^0, \boldsymbol{\Lambda}_\gamma$ in (3.C.1) by their maximum likelihood estimators, after simplification, the log likelihood ratio can be expressed as

$$\Gamma(\tau) = l_1 - l_2 = \frac{\tau(m-\tau)}{m} \sum_{i=\tau+1}^{m} (\hat{\boldsymbol{\mu}}_\gamma^1 - \hat{\boldsymbol{\mu}}_\gamma^0)^T \boldsymbol{\Lambda}_\gamma^{-1} (\hat{\boldsymbol{\mu}}_\gamma^1 - \hat{\boldsymbol{\mu}}_\gamma^0). \qquad (3.C.3)\blacksquare$$

References

1. Chicken, E., Pignatiello, J. J., and Simpson, J. R. (2009) Statistical process monitoring of nonlinear profiles using wavelets. *Journal of Quality Technology*, **41**, 198-212.

2. Chu, Y. X., Hu, S. J., Hou, W. K., Wang, P. C., and Marin, S. P. (2004) Signature Analysis for Quality Monitoring in Short Circuit GMAW. *Welding Journal*, 336s-343s.

3. Daubechies, I. (1992). *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA.

4. Davidian, M., and Giltinan, D.M. (1995) *Nonlinear Models for Repeated Measurements Data*, Chapman and Hall, London, UK.

5. Demidenko E. (2004) *Mixed Models: Theory and Applications*, Wiley, New York, NY.

6. Ding, Y., Zeng, L., Zhou S., (2006) Phase I analysis for monitoring nonlinear profiles in manufacturing processes. *Journal of Quality Technology*, **38**, 199-216.

7. Donoho, D. L., and Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *Journal of American Statistical Association*, **90**, 1200–1224.

8. Ertoz, L., Steinbach, M., and Kumar, V. (2003) Finding clusters of different sizes, shapes and densities in noisy high dimensional data. Presented at the SIAM International Conference on Data Mining, San Francisco, CA, USA, May 2003.

9. Fan, J. (1996) Test of Significance Based on Wavelet Thresholding and Neyman's Truncation. *Journal of the American Statistical Association*, **91**, 674-688.

10. Fraley, C., and Raftery, A. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, **41**, 578–588.

11. Gardner, M., Lu, J., Gyurcsik, R., Wortman, J., Hornung, B., Heinisch, H., Rying, E., Rao, S., Davis, J. and Mozumder, P. (1997) Equipment fault detection using spatial signatures. *IEEE Transaction on Components, Packaging, and Manufacturing Technology, Part C*, **20**, 294-303.

12. Jensen, W. A., Birch, J. B., and Woodall, W. H. (2008). Monitoring Correlation Within Linear Profiles Using Mixed Models. *Journal of Quality Technology*, **40**, 167–183.

13. Jensen, W. A., and Birch, J. B. (2009) Profile monitoring via nonlinear mixed models. *Journal of Quality Technology*, **41**,18-34.

14. Jeong, M. K., Lu, J. C., and Wang, N. (2006) Wavelet-based SPC procedure for complicated functional data. *International Journal of Production Research*, **44**, 729-744.

15. Jin, J. and Shi, J. (1999) Feature-preserving data compression of stamping tonnage information using wavelets. *Technometrics*, **41**, 327–339.

16. Jin, J., and Shi, J. (2001). Automatic feature extraction of waveform signals for in-process diagnostic performance improvement. *Journal of Intelligent Manufacturing*, **12,** 140-145.

17. Johnson, N. L., and Kotz, S. (1970). *Distributions in statistics. Continuous univariate distributions*-1 : Wiley, New York, NY.

18. Kang, L., and Albin, S. L. (2000) On-line Monitoring When the Process Yields a Linear Profile. *Journal of Quality Technology*, **32**, 418-426.

19. Kim, K., Mahmoud, M. A., and Woodall, W. H. (2003) On the Monitoring of Linear Profiles. *Journal of Quality Technology*, **35**, 317-328.

20. Kothari, R., and Pitts, D. (1999) On finding the number of clusters. *Pattern Recognition Letters*, **20**, 405–416.

21. Mahmoud, M. A., Parker, P. A., Woodall, W. H., and Hawkins, D. M. (2007) A Change Point Method for Linear Profile Data. *Quality and Reliability Engineering International*, **23**, 247-268.

22. Mahmoud, M. A. and Woodall, W. H. (2004) Phase I Analysis of Linear Profiles with Calibration Applications. *Technometrics*, **46**, 380-391.

23. Mallat, S. G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **11**, 674–693.

24. Mallat, S. (1999). *A Wavelet Tour of Signal Processing*, Academic Press, Burlington, MA.

25. Mosesova, S. A., Chipman, H. A., MacKay, R. J., Steiner, S. H. (2006) Profile monitoring using mixed-effects models. *Technical report RR-06-06*, University of Waterloo.

26. Pinheiro, J.C., and Bates, D.M. (2000) *Mixed-Effects Models in S and S-PLUS*, Springer- Verlag, New York, NY.

27. Ramsay, J. O., and Silverman B. W. (1997) *Functional Data Analysis*, Springer-Verlag, New York, NY.

28. Shiau, J. H., Huang, H., Lin S., and Tsai M. (2009) Monitoring nonlinear profiles with random effects by nonparametric regression. *Communications in Statistics—Theory and Methods*, **38**, 1664–1679.

29. Sullivan, J. H., and Woodall, W. H. (1996) A Control Chart for Preliminary Analysis of Individual Observations. *Journal of Quality Technology*, **28**, 265-278.

30. Sullivan, J. H., and Woodall, W. H. (2000) Change-Point Detection of Mean Vector or Covariance Matrix Shifts Using Multivariate Individual Observations. *IIE Transactions-Quality and Reliability Engineering*, **32**, 537-549.

31. Sullivan, J. H. (2002) Estimating the Locations of Multiple Change Points in the Mean. *Computational Statistics*, **17**, 289-296.

32. Vargas, J. A. (2003) Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology*, **35**, 367-376.

33. *Walker*, E., and *Wright*, S. P. (2002) Comparing Curves Using Additive Models. *Journal of Quality Technology,* **34**, 118-129.

34. Williams, J. D., Woodall, W. H., and Birch, J. B. (2007) Phase I analysis of nonlinear product and process quality profiles. *Quality and Reliability Engineering International,* **23**, 925–941.

35. Worsley, K. J. (1979) On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, **74**, 365-367.

36. Zamba, K. D., Hawking, D. M. (2006) A Multivariate Change-point Model for Statistical Process Control. *Technometrics*, **48**, 539-549.

37. Zhang, H., and Albin, S. (2007) Determining the number of operational modes in baseline multivariate SPC data. *IIE Transactions-Quality and Reliability Engineering*, **39**, 1103-1110.

38. Zou, C., Zhang, Y., Wang, Z. (2006) A Control Chart Based on a Change-Point Model for Monitoring Profiles. *IIE Transactions-Quality and Reliability Engineering*, **38**, 1093-1103.

39. Zou, C., Tsung, F., and Wang, Z. (2008) Monitoring Profiles Based on Nonparametric Regression Methods. *Technometrics*, **50**, 512-526.

40. Zou, C., Qiu, P., and Hawkins, D. (2009) Nonparametric Control Chart for Monitoring Profiles Using Change Point Formulation and Adaptive Smoothing. *Statistica Sinica,* **19**, 1337-1357.

CHAPTER IV

Parametric Risk-adjusted Modeling and Monitoring of Binary Survival Profiles with Categorical Operational Covariates

4.1 Introduction

Statistical monitoring for effective detection of the deteriorated mortality rate of surgical outcomes has increasingly attracted researchers' attention. Such detection can be further used to assist root cause identification and decision-making for surgical operation improvement. In order to more effectively detect the performance anomalies that go beyond the natural variability of surgical operations, the risk factor of each patient, which reflects the patient's health condition prior to surgery, must be taken into account. Ignoring patients' risk factors may lead to misjudgments about surgical performance since a surgical outcome depends on not only surgical operation performance, but also patients' risk factors before surgery. For example, patients with high risk factors may inadvertently result in a low rate of successful surgical outcomes. In the related literature, the monitoring charts that account for patients' risks are known as the "risk-adjusted" control charts.

Since Treasure et al. (1997) and Waldie (1998) investigated cases in which poor performance of cardiac surgery centers remained undetected for a long time, tremendous research interests in improving Phase II monitoring of surgical outcomes have emerged.

The relevant research on this topic can be divided into two groups, depending on the types of surgical outcome data used in the monitoring. The first group of monitoring methods focused on each patient's binary survival status at a specified time period after surgery. The second group of monitoring methods used continuous measures of each patient's survival time, or a fixed right censored time if a patient survives beyond a specified time period after surgery.

In the first group of monitoring methods, various risk-adjusted control charts have been developed. Steiner et al. (2000) introduced a risk-adjusted cumulative sum (RA-CUSUM) chart to monitor the binary survival status of any given patient during a thirty-day period after surgery. They adjusted the patient's risk using a logistic regression model and utilized a CUSUM chart to monitor the log-likelihood score corresponding to each operation. Cook et al. (2003) proposed a simpler control chart for monitoring binary surgical outcomes. They developed a risk-adjusted Shewhart p-chart with variable control limits based on grouped patients. Spiegelhalter et al. (2003) proposed a more general monitoring approach, known as the resetting sequential probability ratio test (RSPRT) chart. They showed that the RA-CUSUM chart is a special case of the risk-adjusted RSPRT chart. Instead of directly monitoring the binary survival status of patients, Grigg and Farewell (2004a) proposed a risk-adjustment method to monitor the number of operations between two unsuccessful operations (two deaths). To evaluate the efficacy of the existing monitoring methods, Grigg and Farewell (2004b) compared the performance of the existing RA control charts in detecting changes based on the binary outcomes. Grigg and Spiegelhalter (2007) developed a risk-adjusted exponentially weighted moving

average (RA-EWMA) chart, which, in addition to monitoring, can be used to estimate the risk of unsuccessful surgery for each patient.

In the second group of monitoring methods, the control charts are developed based on continuous measures of the exact survival time or right censored time after surgery. Biswas and Kalbfleisch (2008) proposed an RA-CUSUM chart to monitor continuous survival time based on the Cox model. Sego et al. (2009) used location-scale regression models to monitor survival time, in which the corresponding observations are considered as censored data if a patient survives beyond thirty days after surgery. They proposed a risk-adjusted survival time CUSUM chart (RAST-CUSUM) based on the log-likelihood score of each operation. Gandy et al. (2010) extended the RAST-CUSUM to a more general setting, in which they considered general alternatives and a head start of the CUSUM chart. Steiner and Jones (2010) proposed an updated EWMA chart for monitoring risk-adjusted survival time.

All of the afore-mentioned research focuses on Phase II (prospective) monitoring, where it is assumed that the parameters of the risk-adjustment model are known or can be accurately estimated from historical data collected from a stable process. The Phase I (retrospective) control, however, is crucially needed in practice for checking the quality of historical data, and for obtaining accurate estimates of the model parameters, based on which the patient's risk factor can be correctly adjusted for Phase II monitoring. Despite the importance of Phase I control, very little work has been done in the literature on risk-adjusted control charts for Phase I control. Furthermore, for constructing risk-adjusted control charts in Phase I, since each sample represents an individual operation for each patient, it would be impossible to fit a risk-adjustment

model for each patient based on individual observations. Therefore, it is necessary to check whether individual observations can be grouped together, which can then be adjusted by the same risk-adjustment model.

Most of the previous research considered patient's risk factors as the only continuous covariates in risk-adjustment models. However, there are often other variables that may also significantly affect surgical outcomes. For example, in addition to the preexisting health condition of a patient, certain operational variables such as surgeons, surgical procedures, and the types of surgery operations may also influence surgical outcomes. Generally, the performance of experienced surgeons may be different from that of inexperienced surgeons. As a result, the parameters of the risk-adjustment model would be different for surgeons with different levels of experience or skills. Hence, ignoring such important variables in the risk-adjustment model may result in an inaccurate estimation of the risk-adjustment model. In this chapter, we focus on the Phase I control of binary surgical outcomes that are affected by heterogeneous risk factors and operational variables. Generally, these operational variables are often recorded as categorical covariates. Therefore, a logistic regression model, which includes dummy variables for categorical covariates, is first employed in this chapter to represent a unified risk-adjustment model. Change-point models have been widely used in various Phase I control chart applications for continuous responses (See for example, Sullivan and Woodall 2000, Zamba and Hawkins 2006, and Mahmoud et al., 2007). This chapter intends to extend such work for binary responses with the focus on the Phase I control of risk-adjustment models. Specifically, the Phase I control chart is constructed via a likelihood ratio test derived from a change-point model (LRTCP) based on the risk-

adjustment logistic regression. The proposed Phase I risk-adjusted control chart is the first to include both continuous and categorical covariates in the risk-adjustment model.

The rest of the chapter is organized as follows. A motivating example is presented in the next section that illustrates the importance of including categorical covariates in risk-adjustment models. Then a general risk-adjustment model that includes categorical operational covariates is introduced. The detailed development of the proposed Phase I risk-adjusted control charts using LRTCP is also given. After that, the proposed control charts are examined through a case study of evaluating cardiac surgery performance. The effect of ignoring the categorical surgeon covariate on modeling and monitoring is also analyzed and discussed. The following section is devoted to studying and comparing, through Monte Carlo simulations, the performance of the proposed Phase I risk-adjusted LRTCP control charts with and without considering categorical operational covariates.

4.2 A Motivating Example

In this section, a motivating example is presented to illustrate the potential drawbacks of model fitting when categorical operational covariates are ignored. We analyzed a cardiac surgery dataset from a single surgical center in the U.K. This is the same dataset that was used by various authors including Steiner et al. (2000), Sego et al. (2009), Steiner et al. (2010), etc. The dataset covers a seven-year period from 1992 to the end of 1998, and includes surgery information for each patient such as the surgeon's code, the type of surgery operation, survival time, patient's age, and patient's Parsonnet score. The Parsonnet score, commonly used in cardiac surgery, is a weighted score of pre-operative variables that reflect the health condition of a patient before surgery (Parsonnet et al.

1989). This score is often used to adjust the risk associated with heterogeneous patients. Similar to Steiner et al. (2000), we selected the first two years of data, years 1992 and 1993, as the initial Phase I observations of three experienced surgeons. The analysis of this portion of the dataset indicated that the risk-adjusted mortality rate of patients operated by surgeon 1 differs from that of the others. Therefore, surgeon 1 was labeled as group 1 and the rest as group 2. Two separate models were fitted for these two groups of surgeons. Figure 4-1 shows the fitted mortality rates obtained by fitting a logistic regression model of the thirty-day mortality rate on the Parasonnet score as the risk factor. In Figure 4-1, the dashed and dotted lines represent the two fitted models corresponding to group 1 and group 2 of surgeons, respectively; and the solid line represents the fitted model when the surgeon covariate is ignored (combining two groups of surgeons together). As can be seen from Figure 4-1, the fitted model without considering the surgeon covariate is different, especially from that of group 2. Therefore, the model that ignores the effect of surgeons may not accurately reflect the patients' risk in the surgical performance by the surgeons in group 2.

In addition, ignoring the important categorical operational covariates (e.g., different surgeons) in fitting a risk-adjustment model may affect the control chart performance in Phase I as well as in Phase II. The major reason is that this may lead to higher variances of parameter estimators, and result in a poorer performance in detecting possible changes. Therefore, it is vital to consider important categorical operational covariates in the risk-adjustment model for examining historical data in Phase I control. A naive approach to account for the effect of a categorical operational covariate is to fit a risk-adjustment model for each level of the categorical variable. However, as the number

of levels in the categorical operational covariates increases, the number of models and control charts required for monitoring the process also quickly increases. This results in not only more computing efforts, but an increase in the overall Type I error rate associated with the control charts. To tackle this problem, we propose to incorporate the categorical variables into one model by using dummy variables.

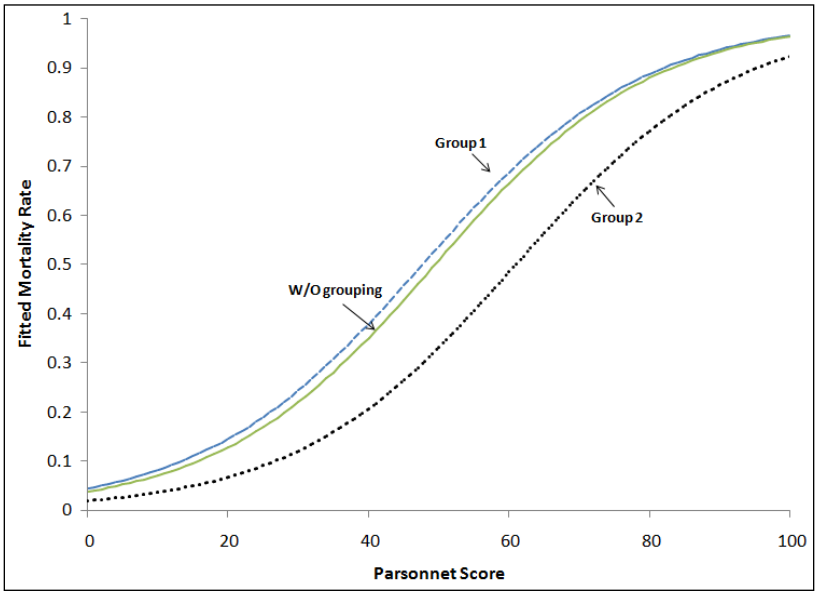

Figure 4-1. Fitted risk-adjustment models based on each surgeon's group and all surgeons.

4.3 Phase I Risk-Adjusted Control Charts with Categorical Variables

4.3.1 A Risk-adjustment Model with Categorical Variables

Consider a process with a binary outcome Y_i , where $Y_i = 1$ if the process fails and $Y_i = 0$ otherwise, and i=1,2,… represents the time index. Suppose θ_i denotes the process failure probability that is a function of a set of risk factors denoted by the vector $\mathbf{x}_i$. The risk-adjustment model is represented as

$$\theta_i = g(\boldsymbol{\beta}, \mathbf{x}_i), \tag{4-1}$$

where $\boldsymbol{\beta}$ is a vector of parameters and the function $g(\cdot)$ denotes a risk-adjustment function. Since the outcome is binary, a logit function is an appropriate choice for $g(\cdot)$. The logit function based risk-adjustment model can be written as

$$\theta_i = \frac{\exp(\sum_{r=0}^{p} \beta_r x_{ir})}{1+\exp(\sum_{r=0}^{p} \beta_r x_{ir})} = \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}^T \mathbf{x}_i)}, \tag{4-2}$$

where $\mathbf{x}_i = (x_{i0}, x_{i1}, ..., x_{ip})^T$ represents the i^{th} observation vector of p risk factors, and $x_{i0} = 1$; $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_p)$ denotes the vector of risk factors' coefficients; and the superscript T denotes the transpose operator.

We extend the logistic regression model in (4-2) to include the categorical covariates. Suppose there are K categorical operational covariates, each with c_k levels, k=1,2,…,K. As mentioned earlier, one way to consider the categorical variables is to group the dataset based on the levels of the categorical variables and fit a risk-adjustment model for each group. This requires an exhaustive model fitting effort, in which a total of $\prod_{k=1}^{K} c_k$ models would need to be fitted. In addition to the increased computational complexity, this approach is also problematic in Phase I control because the number of required control charts is equal to $\prod_{k=1}^{K} c_k$, which results in an excessive increase of the overall Type I error rate of the control procedure. An alternative approach is to account for categorical variables in a unified model by incorporating dummy variables. In this case, the categorical variable k is represented in the model by $(c_k - 1)$ dummy variables. The risk-

adjustment model with dummy variables, which can represent different levels of the categorical operational covariates, can be written as

$$\theta_i = \frac{\exp\left(\sum_{k=1}^{K}\sum_{j=2}^{c_k}\gamma_{jk}d_{ijk} + \sum_{l=0}^{p}\beta_l x_{il}\right)}{1+\exp\left(\sum_{k=1}^{K}\sum_{j=2}^{c_k}\gamma_{jk}d_{ijk} + \sum_{l=0}^{p}\beta_l x_{il}\right)} = \frac{\exp(\sum_{k=1}^{K}\boldsymbol{\gamma}_k^T\mathbf{d}_{ik} + \boldsymbol{\beta}^T\mathbf{x}_i)}{1+\exp(\sum_{k=1}^{K}\boldsymbol{\gamma}_k^T\mathbf{d}_{ik} + \boldsymbol{\beta}^T\mathbf{x}_i)}, \quad (4\text{-}3)$$

where $\mathbf{d}_{ik} = \left(d_{i2k}, d_{i3k}, ..., d_{ic_k k}\right)^T$ is the vector of dummy variables, and $d_{ijk}, (i = 1,...,m;\ j = 2,...,c_k, k = 1,...,K)$ corresponds to observation *i* at level *j* of the k^{th} categorical operational covariate, which is equal to 1 when the k^{th} categorical variable is at level *j*, and equal to 0 otherwise. By convention, if the k^{th} categorical variable is at its first level for observation *i*, $d_{ijk} = 0$, for all $j(j = 2,...,c_k)$. $\boldsymbol{\gamma}_k = \left(\gamma_{2k}, \gamma_{3k}, ..., \gamma_{c_k k}\right)^T$ is also the coefficient vector of the k^{th} dummy variables with c_k categorical levels, and γ_{jk} corresponds to level *j* of the k^{th} categorical operational covariate. The vectors $\boldsymbol{\gamma}_k$ and $\boldsymbol{\beta}$ can be estimated based on historical data using the maximum likelihood (ML) approach (McCullagh and Nelder 1989, pp.115-117).

4.3.2 Phase I Risk-Adjusted Control Charts Based on a Change-Point Model

Suppose there are *m* historical observations available, the i^{th} of which denoted as $\left[Y_i\ \mathbf{d}_{i1}^T\ \mathbf{d}_{i2}^T\ ...\ \mathbf{d}_{iK}^T\ \mathbf{x}_i\right]$, where $i = 1,2,...,m$ indicates the observations order. Let θ_i represent the failure probability for observation *i*. In this section, we present a monitoring approach for examining the historical data and determining the baseline model in Phase I control. This baseline model can be further used to monitor process outcomes for incoming observations in Phase II monitoring. To develop the Phase I control chart, a likelihood

ratio test procedure derived from a change-point model (LRT_{CP}) is utilized in our approach.

The change-point models have been effectively applied in Phase I control. For example, the LRT_{CP} model was developed for univariate normally distributed data (Sullivan and Woodall 1996) and for multivariate normally distributed data (Worsley 1979, Sullivan and Woodall 2000, and Zamba and Hawkins 2006). Sullivan (2002) developed a change-point model based on a clustering approach. Zou et al. (2006), and Mahmoud et al. (2007) applied a change-point model to linear profiles monitoring. The previous research (see for example, Sullivan and Woodall 1996, and 2000) has shown that the LRT_{CP} control chart outperforms other Phase I methods such as the Shewhart and multivariate T^2 control charts when sustained changes occur in the process. Furthermore, the LRT_{CP} model can also provide an estimate of the time when changes occur in the process. These estimates can further be used to assist in identifying the root causes of process changes. For binary data monitoring, Gurevich and Vexler (2005) developed an LRT_{CP} model by using the risk-adjustment logistic regression model with both known and unknown parameters wherein the response risk is adjusted by the continuous risk covariates. Shang et al. (2011) used a similar LRT_{CP} combined with the EWMA chart for the purpose of Phase II monitoring of binary profiles. Nevertheless, there is little research on developing the LRT_{CP} for Phase I control of risk-adjusted binary outcomes with the presence of both categorical covariates and continuous risk factors. As shown in the motivation example, categorical covariates often exist and play a very important role in constructing a risk-adjustment model. Thus ignoring important categorical covariates may bring about misleading results. Therefore, we extend the LRT_{CP} proposed by

Gurevich and Vexler (2005) to develop a Phase I monitoring method for risk-adjustment models with categorical covariates. Henceforth, the proposed control chart is simply called the risk-adjusted LRT_{CP} (RA-LRT_{CP}) chart.

As defined earlier, the binary outcome Y_i follows a Bernoulli distribution with $P(Y_i = 1) = \theta_i = 1 - P(Y_i = 0)$. Suppose an assignable cause occurs at an unknown time τ, which leads to changes of parameters of the risk-adjustment model. The distribution of Y_i, hence, can be written as

$$Y_i \sim \begin{cases} \text{Bernoulli}\left(\theta_i = \dfrac{\exp\left(\sum_{k=1}^{K} \boldsymbol{\gamma}_{0k}^T \mathbf{d}_{ik} + \boldsymbol{\beta}_0^T \mathbf{x}_i\right)}{1 + \exp\left(\sum_{k=1}^{K} \boldsymbol{\gamma}_{0k}^T \mathbf{d}_{ik} + \boldsymbol{\beta}_0^T \mathbf{x}_i\right)}\right) & i \le \tau \\ \text{Bernoulli}\left(\theta_i = \dfrac{\exp\left(\sum_{k=1}^{K} \boldsymbol{\gamma}_{1k}^T \mathbf{d}_{ik} + \boldsymbol{\beta}_1^T \mathbf{x}_i\right)}{1 + \exp\left(\sum_{k=1}^{K} \boldsymbol{\gamma}_{1k}^T \mathbf{d}_{ik} + \boldsymbol{\beta}_1^T \mathbf{x}_i\right)}\right) & i > \tau, \end{cases}$$

where $\boldsymbol{\gamma}_{0k}$ and $\boldsymbol{\beta}_0$ represent the parameters of the risk-adjustment model before the change; and $\boldsymbol{\gamma}_{1k}$ and $\boldsymbol{\beta}_1$ represent these parameters after the change. Define $\boldsymbol{\psi}^{(sl)}$ as the vector of the risk-adjustment model parameters for observations *s*+1 to *l*. Hence, $\boldsymbol{\psi}^{(0\tau)} = \left[\boldsymbol{\gamma}_{01}^T \boldsymbol{\gamma}_{02}^T \ldots \boldsymbol{\gamma}_{0K}^T \boldsymbol{\beta}_0^T\right]^T$ and $\boldsymbol{\psi}^{(\tau m)} = \left[\boldsymbol{\gamma}_{11}^T \boldsymbol{\gamma}_{12}^T \ldots \boldsymbol{\gamma}_{1K}^T \boldsymbol{\beta}_1^T\right]^T$. If all the data follow an identical distribution, then $\boldsymbol{\psi}^{(0\tau)} = \boldsymbol{\psi}^{(\tau m)}$ for all $\tau = u, u+1, \ldots, m-u$, implying that the process is in-control, where *u* (*u*> the number of coefficients) is the minimum required sample size for estimating the parameters of the risk-adjustment model. Therefore, in order to control the process in Phase I, it suffices to evaluate the following hypotheses:

$$\begin{cases} \mathrm{H}_0 : \boldsymbol{\psi}^{(0\tau)} = \boldsymbol{\psi}^{(\tau m)} \\ \mathrm{H}_\mathrm{a} : \boldsymbol{\psi}^{(0\tau)} \neq \boldsymbol{\psi}^{(\tau m)} \end{cases}, \quad \tau = u, u+1, ..., m-u, \tag{4-4}$$

The value of u is chosen so that at least one outcome with value 0 and one outcome with value 1 exist among sampled data from 1 to u and also from m-u+1 to m. Thus, a likelihood ratio test can be used to test the hypotheses in (4-4).

The log likelihood function of Y_i under the alternative hypothesis in (4-4) can be written as:

$$\begin{aligned} l_a &= \log\left\{ \prod_{i=1}^{\tau} h\left(y_i; \boldsymbol{\psi}^{(0\tau)}, \mathbf{d}_{i1}^T, \mathbf{d}_{i2}^T, ..., \mathbf{d}_{iK}^T, \mathbf{x}_i\right) \prod_{i=\tau+1}^{m} h\left(y_i; \boldsymbol{\psi}^{(\tau m)}, \mathbf{d}_{i1}^T, \mathbf{d}_{i2}^T, ..., \mathbf{d}_{iK}^T, \mathbf{x}_i\right) \right\} \\ &= \sum_{i=1}^{\tau} \left\{ y_i \left(\sum_{k=1}^{K} \boldsymbol{\gamma}_{0k}^T \mathbf{d}_{ik} + \boldsymbol{\beta}_0^T \mathbf{x}_i \right) - \log\left(1 + \exp\left(\sum_{k=1}^{K} \boldsymbol{\gamma}_{1k}^T \mathbf{d}_{ik} + \boldsymbol{\beta}_1^T \mathbf{x}_i \right) \right) \right\} + \\ &\quad \sum_{i=\tau+1}^{m} \left\{ y_i \left(\sum_{k=1}^{K} \boldsymbol{\gamma}_{1k}^T \mathbf{d}_{ik} + \boldsymbol{\beta}_1^T \mathbf{x}_i \right) - \log\left(1 + \exp\left(\sum_{k=1}^{K} \boldsymbol{\gamma}_{1k}^T \mathbf{d}_{ik} + \boldsymbol{\beta}_1^T \mathbf{x}_i \right) \right) \right\} \\ &= \sum_{i=1}^{\tau} \left\{ y_i \ \mathrm{logit}\left(\theta_{i0}\right) + \log\left(1 - \theta_{i0}\right) \right\} + \sum_{i=\tau+1}^{m} \left\{ y_i \ \mathrm{logit}\left(\theta_{i1}\right) + \log\left(1 - \theta_{i1}\right) \right\}, \end{aligned} \tag{4-5}$$

where $h(.)$ is the Bernoulli probability mass function; θ_{i0} and θ_{i1} represent the failure rates for the i^{th} observation calculated by using $\boldsymbol{\psi}^{(0\tau)}$ and $\boldsymbol{\psi}^{(\tau m)}$, respectively; and $\mathrm{logit}\left(a\right) = \log\left\{a/(1-a)\right\}; 0 < a < 1$. Let $\hat{\boldsymbol{\psi}}^{(sl)}$ denote the ML estimate of the model parameters that is obtained by fitting the risk-adjustment model (3) based on observations $s+1$ to l. Therefore, the ML estimates of $\boldsymbol{\psi}^{(0\tau)}$ and $\boldsymbol{\psi}^{(\tau m)}$ under H_a are $\hat{\boldsymbol{\psi}}^{(0\tau)}$ and $\hat{\boldsymbol{\psi}}^{(\tau m)}$, respectively. Furthermore, θ_{i0} and θ_{i1} are functions of $\boldsymbol{\psi}^{(0\tau)}$ and $\boldsymbol{\psi}^{(\tau m)}$, therefore, $\hat{\theta}_i^{(0\tau)}$ and $\hat{\theta}_i^{(\tau m)}$, which are calculated by substituting $\hat{\boldsymbol{\psi}}^{(0\tau)}$ and $\hat{\boldsymbol{\psi}}^{(\tau m)}$ correspondingly into (4-3), are ML estimates of θ_{i0} and θ_{i1}, respectively (Casella and Berger, 2001).

Under H_0, the corresponding log likelihood function would be

$$
\begin{aligned}
l_0 &= \log\left\{\prod_{i=1}^{\tau} h\left(y_i; = \boldsymbol{\psi}^{(0m)}, \mathbf{d}_{i1}^T, \mathbf{d}_{i2}^T, ..., \mathbf{d}_{iK}^T, \mathbf{x}_i\right)\right\} \\
&= \sum_{i=1}^{m}\left\{ y_i\left(\sum_{k=1}^{K} \boldsymbol{\gamma}_{0k}^T \mathbf{d}_{ik} + \boldsymbol{\beta}_0^T \mathbf{x}_i\right) - \log\left(1 + \exp(\sum_{k=1}^{K} \boldsymbol{\gamma}_{0k}^T \mathbf{d}_{ik} + \boldsymbol{\beta}_0^T \mathbf{x}_i)\right)\right\} \qquad (4\text{-}6) \\
&= \sum_{i=1}^{m}\left\{ y_i \text{ logit}\left(\theta_{i0}\right) + \log\left(1-\theta_{i0}\right)\right\}.
\end{aligned}
$$

Thus, the ML estimates of $\boldsymbol{\psi}^{(0m)}$ and θ_{i0} under H_0 are $\hat{\boldsymbol{\psi}}^{(0m)}$ and $\hat{\theta}_i^{(0m)}$, respectively. Replacing θ_{i0} and θ_{i1} in (4-5) and (4-6) by their ML estimators, after simplification, the negative of the log likelihood ratio for each τ $(\tau = u, u+1, ..., m-u)$ can be expressed as

$$
\Lambda(\tau) = l_a - l_0 = \sum_{i=1}^{\tau}\left\{ y_i \log\left(\hat{R}_{i1}\right) + \log\left(\frac{1-\hat{\theta}_i^{(0\tau)}}{1-\hat{\theta}_i^{(0m)}}\right)\right\} + \sum_{i=\tau+1}^{m}\left\{ y_i \log\left(\hat{R}_{i2}\right) + \log\left(\frac{1-\hat{\theta}_i^{(\tau m)}}{1-\hat{\theta}_i^{(0m)}}\right)\right\}, \quad (4\text{-}7)
$$

where $\hat{R}_{i1}$ and $\hat{R}_{i2}$ denote two odds ratio between $\hat{\theta}_i^{(0\tau)}$ and $\hat{\theta}_i^{(0m)}$, and between $\hat{\theta}_i^{(\tau m)}$ and $\hat{\theta}_i^{(0m)}$, respectively. The odds ratio of *a* and *b* is defined as $\frac{a/(1-a)}{b/(1-b)}; 0 < a, b < 1$.

The RA-LRT_{CP} chart plots the values of $\Lambda(\tau)$ $(\tau = u, u+1, ..., m-u)$ against time indexes, and each $\Lambda(\tau)$ is compared with an upper control limit (UCL). If for all $\tau, \tau = u, u+1, ..., m-u$, $\Lambda(\tau) < \text{UCL}$, it will be concluded that the historical data have been collected from an in-control process, that is, the process performance has been consistent during the time of data collection and the model parameters estimated from the historical dataset can be used to construct the control charts for Phase II monitoring. Otherwise, the process is considered out-of-control, indicating that the process

performance has significantly changed at some time index τ. The value of τ that maximizes $\Lambda(\tau)$ can provide an ML estimate of the change-point, i.e., $\hat{\tau} = \underset{\tau=u,u+1,\ldots,m-u}{\arg\max}\ \Lambda(\tau)$. If the RA-LRT$_{CP}$ chart detects a change, the dataset can be separated into two groups based on the estimated change-point. Then, the RA-LRT$_{CP}$ chart can be applied to each group to further check whether a change-point exists in each group. Using this iterative approach, multiple change-points can also be detected. Furthermore, the estimated change-points may assist practitioners in identifying the root causes of the change for further improvement. It should be noted that the proposed RA-LRT$_{CP}$ chart is also applicable when no categorical operational covariates exist. In this case, to derive equation (4-7), all terms related to categorical variables are deleted, and model (4-2) is used for estimating model parameters. The rest is the same as the RA-LRT$_{CP}$ chart with categorical operational covariates.

4.3.3 Determining the UCL Values for the RA-LRT$_{CP}$ Chart

If the process is in-control, for each fixed τ, the asymptotic distribution of the monitoring statistic $\Lambda(\tau)$ is independent of the values of $\boldsymbol{\gamma}_{0k}$'s and $\boldsymbol{\beta}_0$ and follows a Chi-square distribution with the degrees of freedom equal to the number of coefficients in the logistic regression model (Myers et al., 2010, p. 112). However, since the values of $\Lambda(\tau)$ across all different τ values are autocorrelated, determining the exact distribution of $\Lambda(\tau)$ under H_0 is intractable. Therefore, a Monte-Carlo simulation is used to determine the UCL as follows. At First, the whole Phase I data are used to fit the logistic regression model in (4-3). Then, the obtained model is used to generate *m* random observations, and calculate $\Lambda(\tau)$'s from which $\kappa = \underset{\tau=u,u+1,\ldots,m-u}{\max} \Lambda(\tau)$ is obtained. This

procedure is repeated *r* times and the values of κ are recorded in each repetition. Finally, the UCL can be determined as the $100(1-\alpha)^{\text{th}}$ percentile of the recorded κ's, where α is the desired Type-I error rate.

In order to simplify the UCL calculation for practitioners, the simulated UCL's of the RA-LRT$_{CP}$ chart for different sample sizes (*m*), different numbers of regression coefficients (*v*), and $\alpha = 0.05, 0.01$, obtained from 1,000 replications are reported in Tables 4-1 and 4-2, respectively.

Table 4-1. Simulated UCL values for $\alpha = 0.05$

Sample size (*m*)	Number of regression coefficients (*v*)								
	2	3	4	5	6	7	8	9	10
500	4.20	4.95	5.91	7.20	8.09	9.20	10.08	10.67	11.50
1000	4.62	5.64	6.80	7.52	8.74	10.70	11.69	12.45	13.42
1500	4.89	6.42	7.56	8.87	9.57	11.04	11.83	13.05	13.64
2000	5.41	6.59	7.79	9.10	9.77	11.17	12.12	13.09	13.90
2500	5.73	6.77	7.95	9.28	9.81	11.39	12.14	13.44	14.02
3000	6.00	7.03	8.09	9.36	10.16	11.43	12.22	13.47	14.35
3500	6.07	7.14	8.25	9.47	10.22	11.72	12.52	13.52	14.46
4000	6.27	7.17	8.48	9.51	10.38	11.79	12.63	13.66	14.52
4500	6.27	7.27	8.59	9.57	10.53	11.81	12.80	13.67	14.87
5000	6.29	7.49	8.65	9.69	10.66	11.85	12.83	13.74	14.92

Table 4-2. Simulated UCL values for $\alpha = 0.01$

Sample size (*m*)	Number of regression coefficients (*v*)								
	2	3	4	5	6	7	8	9	10
500	6.20	6.37	8.19	8.99	10.54	11.77	12.91	13.05	14.94
1000	6.48	7.85	9.43	10.01	10.79	12.60	13.61	15.02	16.01
1500	6.99	8.31	9.65	10.86	11.77	13.30	14.56	15.52	16.13
2000	7.42	8.43	9.76	11.08	11.80	13.36	14.67	15.80	16.48
2500	7.46	8.56	10.12	11.12	11.89	13.76	15.07	15.85	16.50
3000	8.01	8.61	10.23	11.33	12.44	13.80	15.14	15.93	16.76
3500	8.22	9.16	10.33	11.42	12.64	13.87	15.24	16.12	17.17
4000	8.30	9.39	10.35	11.45	12.82	13.95	15.50	16.47	17.44
4500	8.31	9.50	10.40	11.57	12.88	14.06	15.58	16.72	17.48
5000	8.33	9.90	10.57	11.64	12.93	14.25	15.91	16.91	17.55

4.4 A Case Study: Phase I Control of Cardiac Surgical Outcomes

In this section, the proposed RA-LRT$_{CP}$ chart is applied to the cardiac surgery data discussed in the introduction. Following Steiner et al. (2000) and Sego et al. (2009), we use the first two years of data (corresponding to 1992 and 1993) as the Phase I data. It should be pointed out that Steiner et al. (2000) and Sego et al. (2009) assumed that the first two years of data were collected from an in-control process. They used the data to estimate the parameters of the risk-adjustment model in order to monitor the rest of the Phase II data from 1994 to 1998. In contrast, we apply the proposed control chart to check whether the cardiac surgery process was in-control during Phase I data collection. If a change is detected, the out-of-control data are removed and a risk-adjustment model is fitted to the remaining Phase I data. Clearly, the fitted model can be further exploited to implement an RA-LRT$_{CP}$ chart for Phase II monitoring.

In Phase I data, there are a total of six surgeons each designated as either a trainee or an experienced surgeon. The cardiac surgery data corresponding to the three experienced surgeons are chosen. This is because trainee surgeons operated on only relatively simple cases. Furthermore, during an operation by a trainee surgeon, an experienced surgeon has always been present to take over the operation if serious difficulties occur. In this case the operation performance may not consistently reflect the performance of individual trainee surgeons. There are 1,112 records in Phase I data related to the selected experienced surgeons. The numbers of patients operated by surgeons 1, 2, and 3 are 565, 286, and 261, respectively. The numbers of dead patients are 54, 27, and 18 (99 in total) corresponding to surgeons 1, 2, and 3, respectively. The Parsonnet score is used as the risk factor in the model since it has been shown that it can effectively reflect the patients'

risks prior to operation such as hypertension, diabetic status, and renal function (Steiner et al., 2000). The Parsonnet scores of patients operated by the selected surgeons range from 0 to 69.

Since the performance of the surgeons may be different from one another over time, surgeons should be included in the risk-adjustment model as a categorical operational covariate. In order to demonstrate the important role of the surgeon covariate in Phase I control, two risk-adjustment models are used to develop the RA-LRT$_{CP}$ chart; i.e., one model without the surgeon covariate and the other with the surgeon covariate. Henceforth, the RA-LRT$_{CP}$ charts based on the former and latter models are called the RA-LRT$_{CP1}$ chart and the RA-LRT$_{CP2}$ chart, respectively.

To construct a RA-LRT$_{CP1}$ chart, the risk-adjustment model in (4-2) is fitted to the data. Since the Parsonnet score is the only risk factor in this study, the risk-adjustment model in (4-2) is reduced to $\theta_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)}$ with x_i as the Parsonnet score of the patient *i*, and the parameters are estimated as $\hat{\boldsymbol{\beta}}_{\text{cp1}} = [\hat{\beta}_0 \quad \hat{\beta}_1]^T = [\text{-}3.471 \quad 0.073]^T$. Furthermore, the RA-LRT$_{CP2}$ chart can be developed based on the risk-adjustment model in (3), which considers the surgeon covariate. Because the categorical covariate considered in this study consists of three surgeons, two dummy variables are included in the risk-adjustment model, i.e., $\theta_i = \frac{\exp(\beta_0 + \beta_1 x_i + \gamma_2 d_{i2} + \gamma_3 d_{i3})}{1+\exp(\beta_0 + \beta_1 x_i + \gamma_2 d_{i2} + \gamma_3 d_{i3})}$, where d_{i2} and d_{i3} are dummy variables corresponding to surgeons 2 and 3, respectively, i.e., if patient *i* is operated by surgeon 2, $d_{i2} = 1, d_{i3} = 0$, if patient *i* is operated by surgeon 3, $d_{i2} = 0, d_{i3} = 1$, and if he/she is operated by surgeon 1, $d_{i2} = 0, d_{i3} = 0$. In this case, the

parameters are estimated as $\hat{\boldsymbol{\beta}}_{cp2} = [\hat{\beta}_0 \;\; \hat{\beta}_1]^T = [\text{-}3.376 \;\; 0.073]^T$ and $\hat{\boldsymbol{\gamma}} = [\hat{\gamma}_2 \;\; \hat{\gamma}_3]^T = [\text{-}0.079 \;\; \text{-}0.324]^T$, respectively. $\hat{\beta}_0 = \text{-}3.376$ can be interpreted as the logit of mortality probability for a healthy patient with the Parsonnet score equal to zero, which is operated by surgeon 1. $\hat{\beta}_1 = 0.073$ is the effect of the Parsonnet score on the mortality logit when the Parsonnet score changes one unit. $\hat{\gamma}_2 = \text{-}0.079$, and $\hat{\gamma}_3 = \text{-}0.324$ indicate the difference between the performance of surgeon 1 and surgeon 2, and surgeon 1 and surgeon 3 in terms of mortality logit, respectively.

The RA-LRT_{CP1} and RA-LRT_{CP2} control charts are constructed using the likelihood ratio test of (4-7) based on the fitted models. The UCLs for the RA-LRT_{CP1} and RA-LRT_{CP2} charts, which are obtained from the simulation procedure described earlier, are 5.99 and 6.86, respectively. These control charts are depicted in Figure 4-2. As can be seen from Figure 4-2, the RA-LRT_{CP1} chart indicates that the process is in-control, while the RA-LRT_{CP2} chart that accounts for the surgeon covariate shows out-of-control signals. According to the change-point estimator of the RA-LRT_{CP2} chart, the time of change in the process is estimated to be $\hat{\tau} = 400$. Clearly, this contradiction in the results is due to the fact that the performance of surgeons is not the same over time. To further explore this, the data before and after the estimated change-point are used to separately fit a risk-adjustment model without the surgeon covariate, and to fit three separate models for each of the three surgeons. The plots of mortality rate versus Parsonnet score are shown in Figure 4-3. It can be seen in Figure 4-3 that the mortality rate of surgeon 1 decreases after the estimated change-point 400. For surgeon 2, this rate also decreases for patients 401 to 1,112 with high Parsonnet scores. On the other hand, an increase in

mortality rate is evident for surgeon 3 after patient 400, even for patients with relatively low Parsonnet scores. However, if the surgeon covariate is ignored and all data are combined, the decrease in mortality rates corresponding to surgeons 1 and 2 is compromised by increased mortality rates of patients operated by surgeon 3, and consequently, as shown in Figure 4-3(d), the mortality rate curves do not significantly change when all patients are combined without considering the surgeon covariate. This is the reason that the RA-LRT$_{CP1}$ chart did not detect the change.

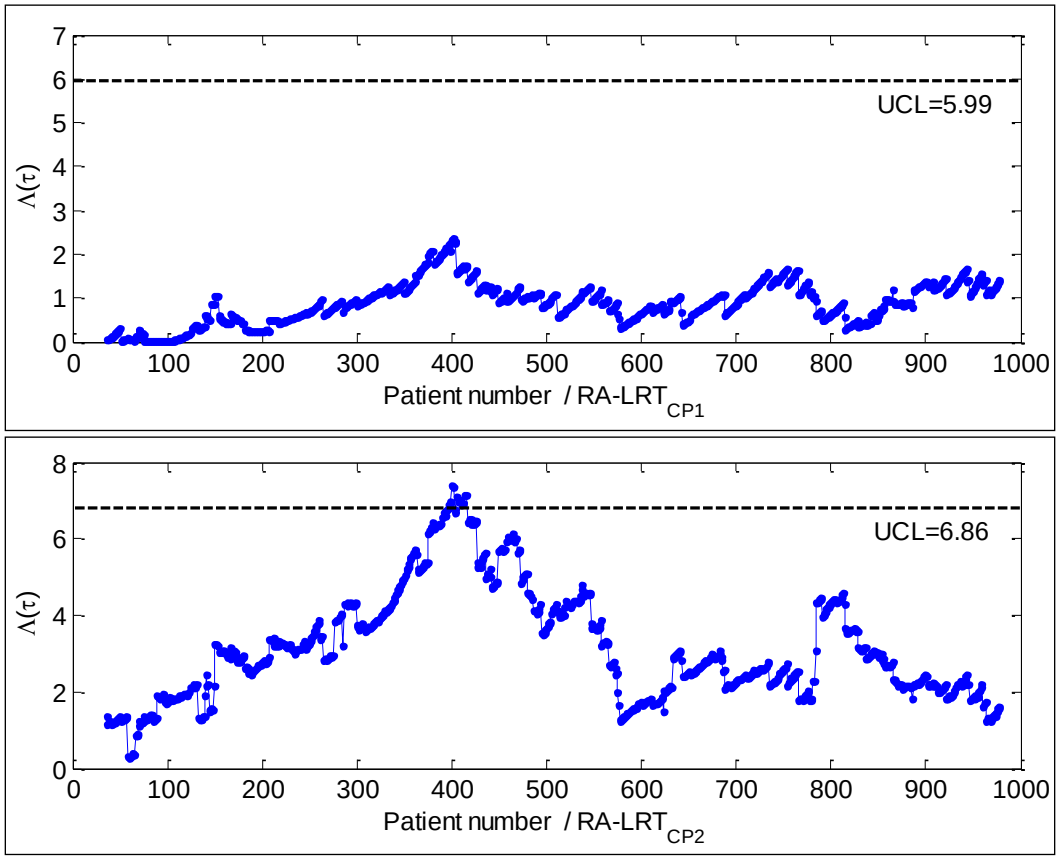

Figure 4-2. RA-LRT$_{CP1}$ (top panel) and RA-LRT$_{CP2}$ (bottom panel) control charts of surgical data.

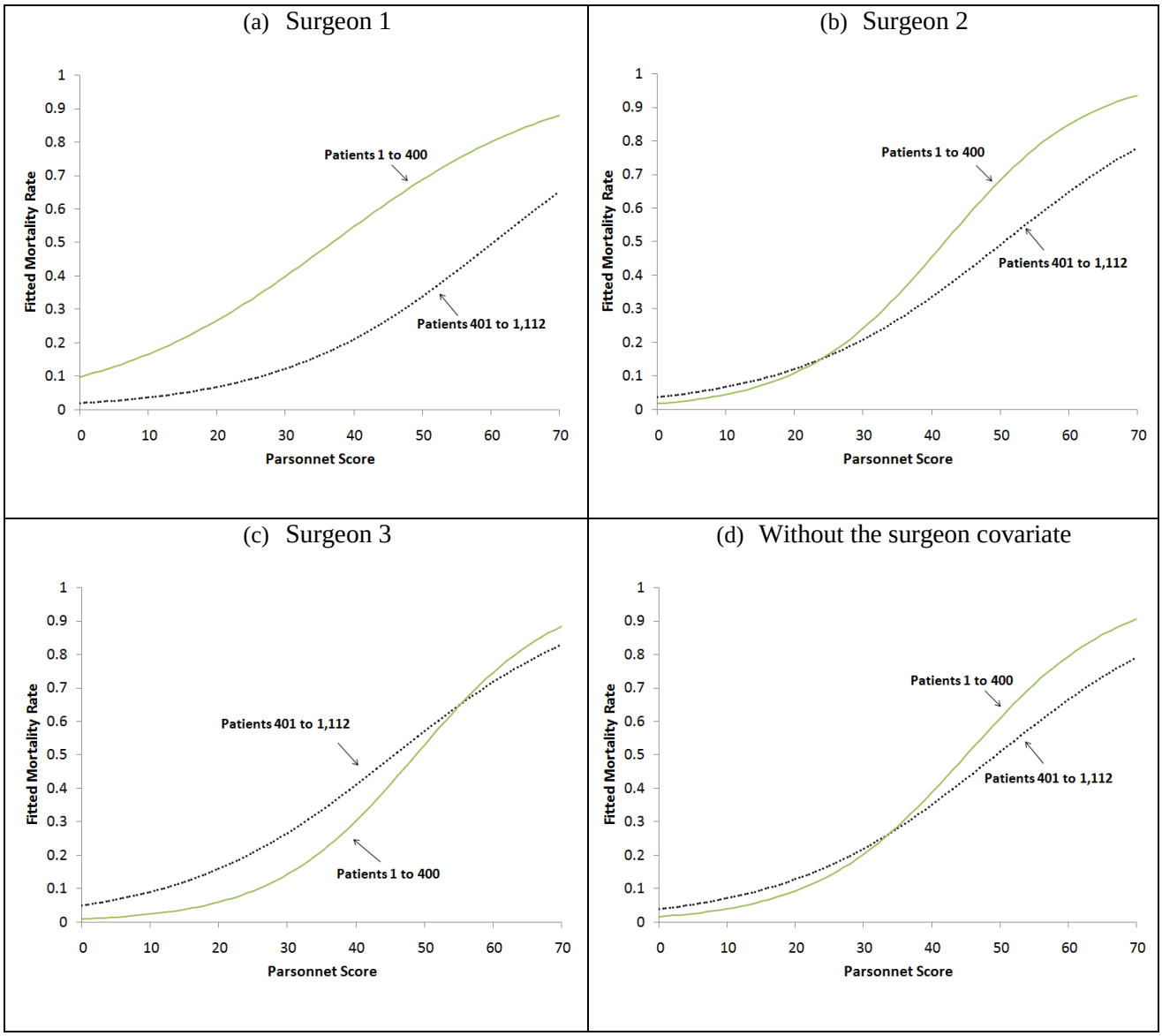

Figure 4-3. Mortality rate plots for different surgeons and the model without the surgeon covariate before and after the change

After receiving a signal from the RA-LRT$_{CP2}$ chart, the Phase I data are divided into two segments based on the estimated change-point 400. Each segment is examined by similar procedures to check if additional out-of-control observations are detected. The results, however, indicate no additional change-point in the Phase I data. Therefore, the first data segment is discarded and the second data segment from patients 400 to 1,112, which reflects the current state of the cardiac surgery operations, is used to fit a baseline

model for implementing the Phase II monitoring. The parameters of the fitted model using the second data segment are estimated as $\hat{\boldsymbol{\beta}}_{cp2} = [\text{-}2.957 \quad 0.065]^T$ and $\hat{\boldsymbol{\gamma}} = [\text{-}0.323 \quad \text{-}0.955]^T$. To check the significance of each coefficient, the Wald hypothesis test is conducted (Myers et al., 2010, pp.109-112). The *p*-values corresponding to all parameters but $\hat{\gamma}_2 = \text{-}0.323$ are less than 0.05. This indicates that surgeon 2 performs as well as surgeon 1, and the data corresponding to surgeons 1 and 2 should be combined into one group. In other words, the number of levels of the surgeon covariate is reduced to 2. The parameters of the new risk-adjustment model with one dummy variable are estimated as $\hat{\boldsymbol{\beta}}_{cp2} = [\text{-}3.050 \quad 0.064]^T$ and $\hat{\gamma} = \text{-}0.849$. This model can be used as a baseline model for monitoring the cardiac surgery process in Phase II. Furthermore, it can help assess whether including the surgeon covariate in the risk-adjustment model is essential in terms of the goodness of fit criterion. This is done by comparing model (4-2) and model (4-3) using likelihood inference (Myers et al., 2010, p.112). The test statistic can be calculated as $-2\log(L_1/L_2) = 5.7061$, in which $-2\log(L_1/L_2)$ follows a Chi-square distribution if the surgeon covariate is not significant. The estimated likelihood functions L_1 and L_2 are obtained for model (2) and model (3), respectively, based on the data set after the change-point of 400. Since the test statistic is larger than the critical value $\chi^2_{1,0.05} = 3.8415$, there is a significant evidence to suggest that the risk-adjustment model which includes the surgeon covariate is better.

If the RA-LRT$_{CP1}$ chart were to be used to control the Phase I data, the entire data would be employed to estimate the parameters of the baseline model since no out-of-control point was detected by the RA-LRT$_{CP1}$ chart. In this case, the estimated parameters

are $\hat{\boldsymbol{\beta}}_{cp1} = [-3.471 \quad 0.073]^T$ which is very different from the baseline model estimated by the RA-LRT_{CP2} chart. Figure 4-1, presented in the introduction, is based on these two estimated baseline models. Such an obvious difference between these two models, as shown in Figure 4-1, implies that ignoring the surgeon covariate in the risk-adjustment model may produce misleading results in both the mortality rate modeling and Phase II monitoring.

Generally, two approaches can be used for Phase II monitoring. One is to construct a set of control charts corresponding to each group of surgeons (i.e., at one level of the categorical covariate). In this way, each control chart at Phase II is built with a different risk-adjusted baseline model that is estimated based on each group's data at Phase I, i.e., observations 400 to 1,112 of each surgeons' group. The other approach is to construct a single control chart using all groups of surgeons' data, in which a categorical covariate is added to the risk-adjustment model to represent different group levels of surgeons. Such a risk–adjustment model is estimated at Phase I based on all groups of surgeons' data. The advantage of the first approach is that it is more explicit than the second approach in identifying which group of surgeons' performance has changed after receiving an alarm at Phase II monitoring. On the other hand, the overall Type-I error rate corresponding to the first approach increases as the number of groups (i.e., the levels of the categorical covariate) increases, while the second approach has a fixed Type-I error rate. We recommend using the second approach for Phase II monitoring because of its simplicity in requiring only one chart construction. For the purpose of identifying which level of the categorical covariate has changed, we suggest using a post-alarm analysis similar to the one illustrated in Figure 4-3.

4.5 Performance Evaluation of the RA-LRT$_{CP}$ Charts

In this section, the performance of the proposed RA-LRT$_{CP1}$ and RA-LRT$_{CP2}$ charts for binary surgical outcomes are further compared through Monte Carlo simulations. The probability of detecting a change in the mortality rate, and the average of the estimated change time ($\bar{\hat{\tau}}$) are two criteria used for the performance comparison.

To generate the simulated data, the risk-adjustment model presented in (4-3) is utilized. Parsonnet score is the sole risk-adjustment factor and the surgeon is the categorical operational covariate with two levels. The estimated parameters of the baseline model from the cardiac surgery example, as obtained in the previous section, are used in the risk-adjustment model. The model that is used to generate the simulation data in this section can be written as

$$\theta_i = \frac{\exp\left(-3.050 + 0.064x_{i1} - 0.849d_{i1}\right)}{1+\exp\left(-3.050 + 0.064x_{i1} - 0.849d_{i1}\right)}. \tag{4-8}$$

Based on the risk-adjustment model in (8), the logit of the mortality rate is denoted by $\text{logit}(\theta_i^j),\ j=1,2,$ where $\text{logit}(\theta_i^1) = -3.899 + 0.064x_{i1}$ and $\text{logit}(\theta_i^2) = -3.050 + 0.064x_{i1}$ correspond to two levels of the surgeon covariate, respectively.

In order to simulate a random binary outcome, the level of the categorical operational covariate d_{i1} is first randomly chosen based on an assigned prior probability of each level, denoted by $\pi_j, j=1,2$. Then, a random Parsonnet score x_{i1} is generated independently from the empirical distribution of Parsonnet scores obtained from the

Phase I surgery dataset. The reason for using the empirical distribution is that Parsonnet scores in the surgery dataset follow no known probability distribution. After that, the randomly generated x_{i1} and d_{i1} are substituted into (4-8) to calculate the mortality rate θ_i. Finally, the obtained mortality rate θ_i is used to generate a binary outcome through a Bernoulli distribution. In each replication, *m* binary outcomes are randomly generated using this procedure.

To assess the performance of the RA-LRT$_{CP1}$ and RA-LRT$_{CP2}$ charts in detecting changes in the mortality rate, two out-of-control scenarios with different magnitudes are examined.

Scenario 1: The logit of the mortality rate corresponding to each surgeon increases, i.e., $\text{logit}\left(\theta_{i1}^{j}\right)=\text{logit}\left(\theta_{i0}^{j}\right)+\delta_j;\ \delta_j>0,\ j=1,2,\ i>\tau$. The expected change in the logit of the overall mortality rate can be calculated as $\delta=\pi_1\delta_1+\pi_2\delta_2$.

Scenario 2: The logit of the mortality rate corresponding to each surgeon changes in the opposite directions, while the expected change in the logit of the overall mortality rate $\pi_1\delta_1+\pi_2\delta_2$ is positive, i.e., $\text{logit}\left(\theta_{i1}^{j}\right)=\text{logit}\left(\theta_{i0}^{j}\right)+\delta_j;\ \delta_1>0,\ \delta_2<0,\ j=1,2,\ i>\tau$.

In both scenarios, δ_1 and δ_2 are set to be proportional to the standard errors of $(\hat{\beta}_1+\hat{\gamma}_1)$ and $\hat{\beta}_1$, respectively, and π_1 is equal to 0.5. The UCL is set so that the estimated Type I error rate α is approximately equal to 0.05. The value of α is estimated by the proportion of simulation runs where at least one of $\underset{\tau=u,u+1,\ldots,m-u}{\Lambda(\tau)}$ values is plotted beyond

UCL. The 95th percentile of the $\max_{\tau=u,u+1,\ldots,m-u} \Lambda(\tau)$ values obtained from 1000 simulation runs when the process is in-control is chosen as the UCL. Furthermore, to investigate the sensitivity of the control charts performance, the simulations are carried out for different values of simulation parameters. We compare the performance under two sample sizes of m=1000 and 2000, and two different change point positions of $\tau = 0.5m$ and $0.75m$, with 1000 replications in each case.

The obtained detection probabilities of both control charts for different change magnitudes and scenarios are shown in Figures 4-4. From Figure 4-4, it can be seen that under both scenarios and different parameters of m and τ, the RA-LRT$_{CP2}$ chart outperforms the RA-LRT$_{CP1}$ chart. For instance, under Scenario 1 with $m = 2000$, and $\tau = 0.75m$, the detection probability corresponding to $\delta = 1.6$ of the RA-LRT$_{CP2}$ chart is about 0.78, while it is 0.64 for the RA-LRT$_{CP1}$ chart. Since the RA-LRT$_{CP2}$ chart can distinguish different surgeon groups, it generally provides more precise parameter estimates, and thus leading to a better detection power.

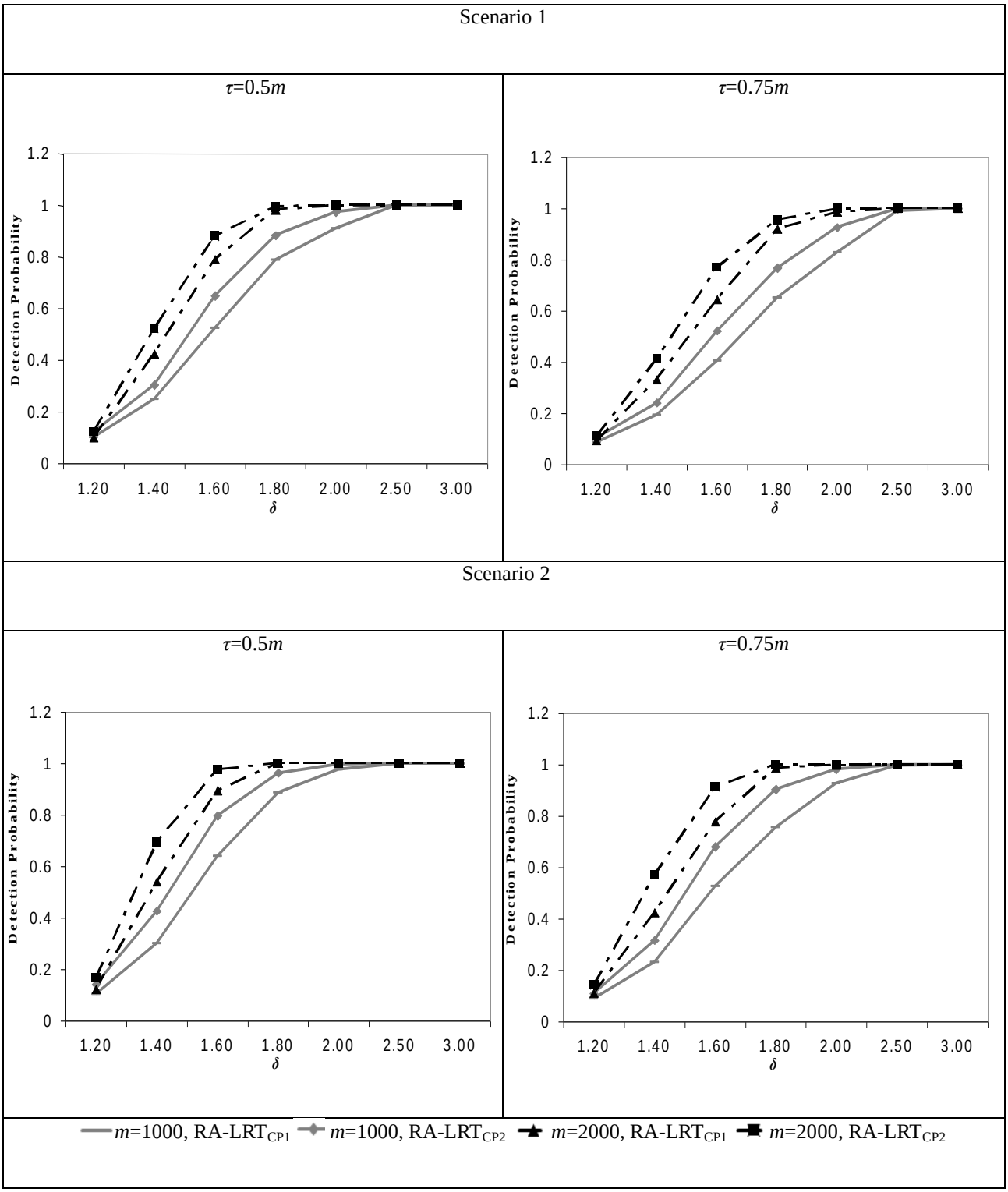

Figure 4-4. Detection probability of "RA-LRT$_{CP1}$" and "RA-LRT$_{CP2}$" charts under different shift scenarios

The performance of both charts improves as the sample size of Phase I data (m) increases. Clearly, the model parameters can be better estimated with a smaller standard

error when *m* is larger, thus resulting in a higher detection power. Another parameter that can be influential in the performance of both control charts is τ. It is obvious from Figure 4-4 that when $\tau = 0.5m$, the performance of both charts is better than when $\tau = 0.75m$. For example, under Scenario 1 with $m = 2000$, and $\tau = 0.5m$, the detection probability corresponding to $\delta = 1.6$ of the RA-LRT$_{CP2}$ chart is about 0.88, while this detection probability decreases to 0.78 for the case with $\tau = 0.75m$. This is due to the fact that when $\tau = 0.5m$, there are equal observations available from each group before and after the change, which makes two groups more distinguishable from each other, leading to better model estimation for both groups.

In order to study the performance of the change-point estimator, the mean values of the change-point estimates obtained from simulations are calculated. These estimated means along with their standard errors (given in the parenthesis) for each of the RA-LRT$_{CP1}$ and RA-LRT$_{CP2}$ charts are reported in Table 4-3. As can be seen from Table 4-3, the change-point estimates in the RA-LRT$_{CP2}$ chart are more accurate and precise than those in the RA-LRT$_{CP1}$ chart. As an example, under Scenario 1 with *m*=2000, $\tau = 1500$, and $\delta = 1.6$, the mean of the estimated change-point is 1477.32 with a standard error of 6.39 for the RA-LRT$_{CP2}$ chart, while these values are 1467.35 and 7.49, respectively, for the RA-LRT$_{CP1}$ chart. The performance of the change-point estimation and the detection probability in both charts are improved when *m* increases. Generally, the higher the detection probability is, the more accurate and precise is the change-point estimator. The estimated change-points whose biases are more than 5% of τ are underlined in Table 4-3. Both RA-LRT$_{CP1}$ and RA-LRT$_{CP2}$ charts produce biases within this range in most cases

considered. Therefore, the change-point estimator performance of both control charts is quite reasonable.

Table 4-3. Mean and standard error of estimated change-point of proposed charts under different scenarios

	δ		1.20	1.40	1.60	1.80	2.00	2.50	3.00
Scenario 1	m=1000	τ=500	CP1*	572.16	520.55	518.57	513.47	505.05	503.25	503.38
				(8.1)	(5.76)	(4.46)	(3.54)	(2.02)	(1.17)	(0.73)
			CP2	544.63	512.10	511.80	509.82	505.77	503.61	503.09
				(8.1)	(5.5)	(4.33)	(2.94)	(1.8)	(0.98)	(0.57)
		τ=750	CP1	583.81	661.30	714.48	738.03	742.72	750.00	751.27
				(9.23)	(7.53)	(5.19)	(3.92)	(3.04)	(1.42)	(0.98)
			CP2	599.30	671.79	716.25	740.20	744.15	750.08	750.26
				(9.23)	(7.34)	(4.62)	(3.54)	(2.43)	(1.14)	(0.82)
	m=2000	τ=1000	CP1	1052.51	1031.52	1017.86	1008.78	1008.44	1004.40	1003.47
				(17.11)	(10.37)	(6.74)	(4.11)	(2.56)	(1.26)	(0.76)
			CP2	1038.49	1017.93	1009.21	1000.03	1006.65	1003.93	1002.07
				(16.19)	(9.23)	(5.63)	(3.16)	(1.9)	(0.85)	(0.57)
		τ=1500	CP1	1249.87	1457.88	1467.35	1480.43	1496.91	1505.24	1504.45
				(18.88)	(11.04)	(7.49)	(5.06)	(3.26)	(1.2)	(0.76)
			CP2	1252.60	1460.51	1477.32	1489.09	1499.79	1504.64	1503.42
				(17.65)	(10.28)	(6.39)	(4.14)	(2.25)	(0.95)	(0.54)
	δ		1.20	1.40	1.60	1.80	2.00	2.50	3.00
Scenario 2	m=1000	τ=500	CP1	575.16	525.06	522.65	512.38	504.82	500.05	497.78
				(7.72)	(5.38)	(4.14)	(3.16)	(2.06)	(0.7)	(0.66)
			CP2	569.29	514.63	511.67	506.14	504.51	499.47	497.34
				(7.62)	(4.87)	(3.57)	(2.5)	(1.52)	(0.66)	(0.63)
		τ=750	CP1	576.91	682.31	717.20	740.65	745.79	748.30	749.66
				(9.11)	(7.08)	(4.84)	(3.29)	(2.21)	(1.17)	(0.76)
			CP2	592.85	683.82	719.21	739.13	745.64	748.47	748.55
				(8.98)	(6.32)	(4.08)	(2.78)	(1.87)	(0.85)	(0.63)
	m=2000	τ=1000	CP1	1069.72	1035.19	1016.15	1010.55	1009.24	1003.74	1002.88
				(16.82)	(9.58)	(5.41)	(3.26)	(1.8)	(0.95)	(0.51)
			CP2	1028.04	1011.45	1008.50	1004.01	1005.04	1002.69	1001.44
				(14.51)	(8.03)	(3.98)	(2.25)	(1.33)	(0.57)	(0.38)
		τ=1500	CP1	1292.85	1467.94	1484.14	1494.26	1498.58	1503.09	1504.73
				(18.12)	(9.87)	(5.6)	(3.6)	(2.62)	(0.82)	(0.54)
			CP2	1306.38	1473.26	1489.42	1495.97	1502.28	1502.98	1502.14
				(15.72)	(8.7)	(4.49)	(2.88)	(1.45)	(0.63)	(0.41)

* CP1 and CP2 represent RA-LRT$_{CP1}$ and RA-LRT$_{CP2}$, respectively.

References

1. Biswas, P., and Kalbfliesch, J. D. (2008) A risk-adjusted CUSUM in continuous time based on the Cox model. *Statistics in Medicine*, **27**, 3382-3406.

2. Casella, G., and Berger, R. L. (1990) *Statistical Inference*. Brooks/Cole, Florence, KY.

3. Cook, D. A., Steiner, S. H., Farewell, V. T. and Morton, A. P. (2003) Monitoring the evolutionary process of quality: Risk adjusted charting to track outcomes in intensive care. *Critical Care Medicine*, **31**, 1676–1682.

4. Gandy, A., Kvaloy, J. T., Bottle, A., and Zhou, F. (2010) Risk-adjusted Monitoring of Time to Event. *Biometrika*, **97**, 375-388.

5. Grigg, O. A. and Farewell, V. T. (2004a) A risk-adjusted sets method for monitoring adverse medical outcomes. *Statistics in Medicine*, **23**, 1593-1602.

6. Grigg, O., and Farewell, V. T. (2004b) An overview of risk-adjusted charts. *Journal of the Royal Statistical Society, Series A*, **167**, 523-539.

7. Grigg, O., and Spiegelhalter, D. (2007) A simple risk-adjusted exponentially weighted moving average. *Journal of the American Statistical Association*, **102**, 140-152.

8. Gurevich, G. and Vexler, A. (2005) Change Point Problems in the Model of Logistic Regression. *Journal of Statistical Planning and Inference*, **131**, 313-331.

9. Mahmoud, M. A., Parker, P. A., Woodall, W. H., and Hawkins, D. M. (2007) A change point method for linear profile data. *Quality and Reliability Engineering International*, **23**, 247-268.

10. McCullagh, P., and Nelder J. A. (1989) *Generalized Linear Models*. Chapman and Hall London, New York, NY.

11. Myers R. H., Montgomery, D. C., Vining, G. G., and Robinson, T. J. (2010) *Generalized Linear Models: with Applications in Engineering and the Sciences, 2nd edition.* Wiley, New York, NY.

12. Sego, L. H., Reynolds, Jr. M. R., and Woodall, W. H. (2009) Risk adjusted monitoring of survival times. *Statistics in Medicine*, **28**, 1386-1401.

13. Shang, Y., Tsung, F., and Zou, C. (2011) Profile monitoring with binary data and random predictors, *Journal of Quality Technology*, **43**, 196-208.

14. Spiegelhalter, D. J., Grigg, O. A., Kinsman, R. and Treasure, T. (2003) Sequential probability ratio tests (SPRTS) for monitoring risk-adjusted outcomes. *International Journal for Quality in Health Care*, **15**, 1–7.

15. Steiner, S. H., Cook, R. J., Farewell, V. T., and Treasure T. (2000) Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics*, **1**, 441-452.

16. Steiner, S. H., and Jones, M. (2010) Risk-adjusted survival time monitoring with an updating exponentially weighted moving average (EWMA) control chart. *Statistics in Medicine*, **29**,444-454.

17. Sullivan, J. H. (2002) Estimating the locations of multiple change points in the mean. *Computational Statistics*, **17**, 289-296.

18. Sullivan, J. H., and Woodall, W. H. (1996) A control chart for preliminary analysis of individual observations. *Journal of Quality Technology*, **28**, 265-278.

19. Sullivan, J. H., and Woodall, W. H. (2000) Change-Point detection of mean vector or covariance matrix shifts using multivariate individual observations. *IIE Transactions-Quality and Reliability Engineering*, **32**, 537-549.

20. Treasure, T., Taylor, K., and Black, N. (1997) Independent review of adult cardiac surgery-United Bristol. *Bristol: Health Care Trust.*

21. *Waldie,* P. (1998) *Crisis* in the *cardiac* unit. *The Globe and Mail, Canada's National Newspaper* Oct. 27, Section A:3, Column 1.

22. Worsley, K. J. (1979) On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, **74**, 365-367.

23. Yeh, A. B., Huwang, L., and Lee, Y-H. (2009) Profile monitoring for binary response. *IIE Transactions-Quality and Reliability Engineering*, **41**, 931-941.

24. Zamba, K. D., and Hawkins, D. M. (2006) A multivariate change-point model for statistical process control. *Technometrics*, **48**, 539-549.

CHAPTER V

Conclusions and Future Research

5.1 Conclusions

This dissertation has focused on developing new methodologies to model and analyze profile data for the purpose of process monitoring, fault diagnosis, and effective decision making for system performance improvement. We first proposed a new framework to identify informative sensors and extract information of multi-stream waveform signals in distributed sensing systems. Then, the problem of modeling profile variations was studied, and a new method was proposed to characterize both within- and between-profile variations. We used this information for identifying sources of variations and fault diagnosis in a process. Furthermore, in the last part of the dissertation, we addressed the problem of modeling and making inferences about nonlinear profiles in the presence of disturbance covariates. The main research results and new contributions of this dissertation are summarized as follows.

(1) *A new hierarchical non-negative garrote (HNNG) method for identifying informative sensors*. The proposed hierarchical non-negative garrote method consists of two steps. In the first step, the informative sensors and their corresponding signals are identified in a distributed sensing system. In the second step, important individual features are selected within each selected signal. We showed that the proposed method

benefits from the following three useful properties: (a) the "similar effects" property for incorporating highly correlated variables, (b) the applicability in cases when the number of important variables is larger than the sample size, and (c) the simplicity in computation to achieve optimum solutions in both steps of the method by using least angle regression with the order of complexity equal to the ordinary least square method. To evaluate the effectiveness of this method, we conducted a simulation study, and the results indicated that the proposed method outperforms other existing methods in terms of both prediction performance and model sparsity. Moreover, we applied our method to a real case study, where we showed that our method performs the best in predicting drivers' comfort scores among all other methods.

(2) *A new mixed-effect model based on wavelets for characterizing both within- and between profile variations*. In most applications, the total inherent variation of profiles often consists of both within-profile and between-profile variations. Characterizing both types of variations and identifying their variation sources are very useful when making proactive decisions for process improvement. For the purpose of variation modeling, we enhanced the mixed-effect models with the capability of wavelets for modeling local variations. Moreover, we used the multi-resolution property of wavelets to separate within-profile from between-profile variations and for estimating them separately, significantly reducing the computational effort for model fitting. In constructing the mixed-effect model, an LRT-based change-point model was also utilized in order to check the historical profile data and identify potential clusters. Furthermore, a method for effective selection of monitoring features was proposed for improving LRT-CP performance. Several Monte-Carlo simulation scenarios were conducted to evaluate

the effectiveness of this proposed approach. The simulation results indicated that the proposed method can effectively model both within- and between- profile variations and also outperforms the existing method in terms of detecting changes in historical profile data. We also applied the proposed wavelet-based mixed-effect model to profile data collected during valve seat pressing operations in an engine head assembly process, and demonstrated how the proposed method can be used to effectively characterize profile variations and identify sources of variations in the process.

(3) *A new parametric method for risk-adjusted modeling and monitoring of binary survival profiles with the categorical operational covariates.* We proposed a general risk-adjusted control charting scheme for Phase I control of surgical performance, which can account for not only patients' health conditions as described by Parsonnet scores, but also other categorical operational covariates, such as the surgeons' group factor. The proposed Phase I risk-adjusted control chart was developed based on a likelihood ratio test derived from a change-point model. The risk-adjustment model is fitted by incorporating the dummy variables to represent different surgeon groups' performances. To demonstrate the importance of including relevant categorical operational covariates in the risk-adjustment model, a cardiac surgery dataset was analyzed. The discovered data clusters for the mortality rate indicated that the inclusion of the categorical surgeon covariate in the risk-adjustment model can effectively model the heterogeneity of the surgical outcome data. The Monte-Carlo simulations were further conducted to demonstrate that by including the surgeon covariate, the Phase I risk-adjusted chart results in a better detection power of surgical performance change. It is expected that the improved

estimation of model parameters based on the proposed Phase I control will lead to a better Phase II monitoring performance.

5.2 Future Research

Waveform sensing signals have broad applications, which provide great opportunities and challenges for system performance improvement. Multidisciplinary approaches were proposed for waveform signal analysis by integrating domain engineering knowledge with advanced data analysis methods. In this dissertation, some initial efforts have been made and demonstrated in both methodological developments and real-world applications. However, future research is needed in this area, and a few examples of such research topics are listed below.

- In the first problem studied in the dissertation, we considered the case, in which the response variable is continuous. However, the hierarchical NNG can be extended to classification models, where the response variable is categorical and can be suggested as a potential topic for future research.
- In practice, there are many instances where process variables are measured and represented as multi-stream waveform signals. In order to effectively analyze multiple profiles, the inter-relationship among profiles should be considered in addition to between- and within-profile variations. This would make the data analysis very challenging and would require the development of new statistical learning techniques integrated with the domain knowledge.
- In modeling profile variations, it was assumed that within-profile noises are identically and independently distributed. If the independency assumption is not

held, the denosing procedure used for removing the within-profile variation may not perform properly since the estimated within-profile variance would be highly biased. Therefore, developing wavelet-based mixed models for modeling nonlinear profile data in the presence of auto-correlated noises would be an interesting topic for future research.

- Recently, some research has been conducted on constructing Phase II risk-adjusted control charts based on continuous measures, such as patient's survival time. Thus, extending the proposed Phase I risk-adjusted charts to continuous survival profiles would be another potential topic for future research.
- Image data are widely used for system monitoring in various applications, such as hot rolling processes and MRI images and contain rich information which can be extracted and used for monitoring, fault detection and diagnosis purposes. There is extensive research on analysis of image data in the literature; nonetheless, there are many issues surrounding this area, specifically in analyzing the image data as it evolves over time. In order to effectively extract the image data information, both spatial and temporal correlations should be taken into account. Moreover, image data can be considered as a general type of functional data in the sense that they can be represented by a spatial function. Therefore, it would be plausible to adapt and modify the functional data techniques and use them for image data analysis and vice versa.