Cluster-Based Profile Monitoring in Phase I Analysis

Yajuan Chen

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Statistics

Jeffery B Birch

Pang Du

Inyoung Kim

William H. Woodall

1/31/2014

Blacksburg VA

Keywords: Cluster, Mixed Model, Phase I, Phase II, Robust, T^2 Statistic

Cluster-Based Profile Monitoring in Phase I Analysis

Yajuan Chen

ABSTRACT

Profile monitoring is a well-known approach used in statistical process control where the quality of the product or process is characterized by a profile or a relationship between a response variable and one or more explanatory variables. Profile monitoring is conducted over two phases, labeled as Phase I and Phase II. In Phase I profile monitoring, regression methods are used to model each profile and to detect the possible presence of out-of-control profiles in the historical data set (HDS). The out-of-control profiles can be detected by using the T^2 statistic. However, previous methods of calculating the T^2 statistic are based on using all the data in the HDS including the data from the out-of-control process. Consequently, the ability of using this method can be distorted if the HDS contains data from the out-of-control process. This work provides a new profile monitoring methodology for Phase I analysis. The proposed method, referred to as the *cluster-based profile monitoring* method*,* incorporates a cluster analysis phase before calculating the T^2 statistic.

Before introducing our proposed cluster-based method in profile monitoring, this cluster-based method is demonstrated to work efficiently in robust regression, referred to as *cluster-based bounded influence regression or CBI*. It will be demonstrated that the CBI method provides a robust, efficient and high breakdown regression parameter estimator. The CBI method first represents the data space via a special set of points, referred to as anchor points. Then a collection of single-point-added ordinary least squares regression estimators forms the basis of a metric used in defining the similarity between any two observations. Cluster analysis then yields a main cluster containing at least half the observations, with the remaining observations comprising one or more minor clusters. An initial regression estimator arises from the main cluster, with a group-additive DFFITS argument used to carefully activate the minor clusters through a bounded influence regression frame work. CBI achieves a 50% breakdown point, is regression equivariant, scale and affine equivariant and distributionally is asymptotically normal. Case studies and Monte Carlo results demonstrate the performance advantage of CBI over other popular robust regression procedures regarding coefficient stability, scale estimation and standard errors.

The cluster-based method in Phase I profile monitoring first replaces the data from each sampled unit with an estimated profile, using some appropriate regression method. The estimated parameters for the parametric profiles are obtained from parametric models while the estimated parameters for the nonparametric profiles are obtained from the p-spline model. The cluster phase clusters the profiles based on their estimated parameters and this yields an initial main cluster which contains at least half the profiles. The initial estimated parameters for the population average (PA) profile are obtained by fitting a mixed model (parametric or nonparametric) to those profiles in the main cluster. Profiles that are not contained in the initial main cluster are iteratively added to the main cluster provided their T^2 statistics are "small" and the mixed model (parametric or nonparametric) is used to update the estimated parameters for the PA profile. Those profiles contained in the final main cluster are considered as resulting from the in-control process while those not included are considered as resulting from an out-of-control process. This cluster-based method has been applied to monitor both parametric and nonparametric profiles. A simulated example, a Monte Carlo study and an application to a real data set demonstrates the detail of the algorithm and the performance advantage of this proposed method over a non-cluster-based method is demonstrated with respect to more accurate estimates of the PA parameters and improved classification performance criteria.

When the profiles can be represented by m $p \times 1$ vectors, the profile monitoring process is equivalent to the detection of multivariate outliers. For this reason, we also compared our proposed method to a popular method used to identify outliers when dealing with a multivariate response. Our study demonstrated that when the out-of-control process corresponds to a sustained shift, the cluster-based method using the successive difference estimator is clearly the superior method, among those methods we considered, based on all performance criteria. In addition, the influence of accurate Phase I estimates on the performance of Phase II control charts is presented to show the further advantage of the proposed method. A simple example and Monte Carlo results show that more accurate estimates from Phase I would provide more efficient Phase II control charts.

Acknowledgments

I would like to express my wholehearted gratitude to my advisor Dr. Jeffery B. Birch for his invaluable guidance, advice and countless hours of editing to help me complete this dissertation. His encouragement, support and thoughtfulness are highly appreciated. As a professor, Dr. Birch is an expert in his statistical areas and as a person, Dr. Birch is a very kind, organized and positive man, I have learned a lot from him during my graduate study and he is the guide in my future life. Also, I would like to thank Dr. Birch for his excellent work as the director of our graduate program. As a graduate student in the Department of Statistics at Virginia Tech, I feel so lucky to have him here as he cares about every aspect of each student's life and provides the help needed to make each graduate student's experience in our department an enjoyable one.

I would like to thank my committee members: Dr. Pang Du, Dr. Inyoung Kim, and Dr. William H. Woodall. They were very helpful and provided me positive input to further improve this work. I thank Dr. Woodall for his great editing and technical advice for my dissertation, papers and presentations. I thank Dr. Kim for her guide in nonparametric analysis and I thank Dr. Du for his guide in functional data analysis.

Many people on the faculty, staff and among the graduate students of the Department of Statistics assisted and encouraged me in various ways. I would like to thank them. I would especially like to thank our faculty for providing excellent courses and consulting experiences that have prepared me for my future career. Also, I thank all my friends in Blacksburg, whose friendship have supported me and have brought me a lot of good memories in my life.

I thank my parents and my brothers. They have been very supportive and encouraging in my life. I want to especially thank my brother Chuanwen, who give me much help whenever it was needed. Further, I thank my parents-in-law for helping us to take care of Ella so I have time to finish this dissertation. Last but not least, I would like to thank my sweet husband Yi and my lovely daughter Ella. I thank them for coming into my life, and for bringing me such

joy and happiness. I thank Yi for his love, support, patience and help during my graduate studies.

Contents

List of Tables

List of Figures

Acronyms

ARL	Average Run Length
ARL_1	Out-of-Control ARL
ARL_0	In-Control ARL
ARL_C	ARL based on the cluster-based T^2 control chart
ARL_NCB	ARL based on the non-cluster-based T^2 control chart
BI	Bounded Influence
blup	Best Linear Unbiased Predictor
CBI	Cluster-based Bounded Influence Regression
eblup	Estimated Best Linear Unbiased Predictor
FCC	Fraction Correctly Classified
FNR	False Negative Rate
FPR	False Positive Rate
HDS	Historical Data Set
HKB	Hawkins DM, Bradu D, Gordon VK
hip	High Influence Point
LMM	Linear Mixed Model
LMS	Least Median Squares
LTS	Least Trimmed Squares
L-W	Laird-Ware
MCD	Minimum Covariate Determinant
MLE	Maximum Likelihood Estimator
MRPM	Model Robust Profile Monitoring
MVE	Minimum Volume Ellipsoid
M1S	Mallows 1-step
OLS	Ordinary Least Square
PA	Population Average
PH	Pendleton and Hocking
POS	Probability of Signal
PS	Profile Specific

REMLE	Restricted Maximum Likelihood Estimator
REWLS	Robust and Efficient Weighted Least Square
RMCD	Reweighted Minimum Covariance Determinant
SPC	Statistical Process Control
S1S	Schweppe's 1-step

Nomenclature

$\boldsymbol{\beta}$	A $p \times 1$ unknown parameter vector
$\hat{\boldsymbol{\beta}}$	A $p \times 1$ estimated parameter vector
$\hat{\boldsymbol{\beta}}_M$	The $p \times 1$ estimated parameter vector by using M regression
$\hat{\boldsymbol{\beta}}_{CBI}$	The $p \times 1$ estimated parameter vector by using CBI regression
$\hat{\boldsymbol{\beta}}_i^P$	The $p \times 1$ estimated parameter vector for the i^{th} profile
$\hat{\boldsymbol{\beta}}_{LMM}$	The $p \times 1$ estimated parameter vector for the PA profile by LMM approach
$\hat{\boldsymbol{\beta}}_{REWLS}$	The $p \times 1$ estimated parameter vector by using REWLS regression
β_0	The intercept parameter for the in-control PA
β_1	The slope parameter for the in-control PA
β_2	The quadratic parameter for the in-control PA
β_0'	The intercept parameter for the out-of-control PA
β_1'	The slope parameter for the out-of-control PA
β_2'	The quadratic parameter for the out-of-control PA
$\boldsymbol{b}_i$	A vector of random effects that represent the eblups for the i^{th} profile
b_{0i}	The random intercept effect for the i^{th} profile
b_{1i}	The random slope effect for the i^{th} profile
b_{2i}	The random quadratic effect for the i^{th} profile
$f(.)$	True PA profile function
$f_i(.)$	True i^{th} profile function
$\boldsymbol{G}$	The $q \times q$ covariance matrix for the random effects $\boldsymbol{b}$
$\boldsymbol{H}$	Hat matrix of $\boldsymbol{X}$
h_{ii}	The i^{th} diagonal element of $\boldsymbol{H}$
$MVE_1(\boldsymbol{Z}_y)$	The MVE center estimator for $\boldsymbol{Z}_y$

$MVE_2\left(\boldsymbol{Z}_y\right)$	The MVE scale matrix estimator for $\boldsymbol{Z}_y$
m_1	Number of the in-control profiles
m	Number of the total profiles in HDS
n	Number of observations
$\boldsymbol{R}_i$	A $n_i \times n_i$ covariance matrix for the random error $\boldsymbol{\varepsilon}_i$
$\boldsymbol{r}$	A $n \times 1$ vector of residuals
r_i	The residual for the i^{th} observation
rs_i	The absolute scaled residual for the i^{th} observation
T_i^2	Hotelling's T^2 statistic for or the i^{th} time period
$T_{MVE,i}^2$	Hotelling's T^2 statistics for or the i^{th} time period based on MVE
$T_{MCD,i}^2$	Hotelling's T^2 statistic for or the i^{th} time period based on MCD
$T_{P1,i}^2$	Hotelling's T^2 statistics based on parametric fitted value
$T_{P2,i}^2$	Hotelling's T^2 statistics based on parametric eblups
$T_{NP2,i}^2$	Hotelling's T^2 statistics based on P-spline fitted value
$T_{NP1,i}^2$	Hotelling's T^2 statistics based on P-spline eblups
$\boldsymbol{t}_i$	A vector of random effects represents the coefficients for spline component
$\hat{\boldsymbol{V}}$	Estimator of variance-covariance matrix
$\hat{\boldsymbol{V}}_D$	The successive difference estimator of variance-covariance matrix
$\boldsymbol{V}_i$	The $n_i \times n_i$ covariance matrix for the response vector $\boldsymbol{y}_i$
$\hat{\boldsymbol{V}}_{MVE}$	Estimator of variance-covariance matrix based on MVE
$\hat{\boldsymbol{V}}_{MCD}$	Estimator of variance-covariance matrix based on MCD
$\hat{\boldsymbol{V}}_P$	The pooled sample estimator of variance-covariance matrix
$\boldsymbol{W}$	The $n \times n$ diagonal weight matrix
w_i	The i^{th} diagonal element of $\boldsymbol{W}$
$\boldsymbol{X}$	A design matrix

$\boldsymbol{x}_i^T$	The i^{th} row of design matrix $\boldsymbol{X}$
$\boldsymbol{X}_i$	The $n_i \times p$ matrix of explanatory variables for the i^{th} profile
x_{ij}	The fixed regressor for the j^{th} observation from the i^{th} profile
$\hat{\boldsymbol{\mu}}_i$	A $n \times 1$ estimated mean vector for the i^{th} time period
$\hat{\boldsymbol{\mu}}_{MCD}$	An estimated mean vector for the i^{th} time period based on MVE
$\hat{\boldsymbol{\mu}}_{MVE}$	An estimated mean vector the i^{th} time period based on MVE
$\hat{\boldsymbol{y}}_{PS,i}^{P}$	A vector of parametric fitted value for the i^{th} profile
$\hat{\boldsymbol{y}}_{PS,i}^{p-s}$	A vector of p-spline fitted value for the i^{th} profile
$\hat{\boldsymbol{y}}_{PA}^{p-s}$	A vector of p-spline fitted value for the PA profile
$\hat{\boldsymbol{y}}_{PA}$	A vector of parametric fitted value for the PA profile
$\boldsymbol{y}_i$	The $n_i \times 1$ response vector for the i^{th} profile
y_{ij}	The response value for the j^{th} observation from the i^{th} profile
$\boldsymbol{Z}$	The $n \times k$ matrix containing only the k regressor variables
$\boldsymbol{Z}_{\boldsymbol{y}}$	The $n \times p$ matrix formed by augmenting the vector $\boldsymbol{y}$ to $\boldsymbol{Z}$
$\boldsymbol{Z}_i$	The $n_i \times q$ matrix of explanatory variables for the i^{th} profile
$\boldsymbol{z}_i^T$	The i^{th} row of matrix $\boldsymbol{Z}$
$\boldsymbol{z}_{y,i}^T$	The i^{th} row of matrix $\boldsymbol{Z}_y$
$\xi_i(.)$	True i^{th} profile smooth function
$\boldsymbol{\varepsilon}_i$	The random error term for the i^{th} profile
ε_{ij}	The random error for the j^{th} observation from the i^{th} profile
λ	The smooth parameter for the penalized regression

Chapter 1. Introduction and Motivation

In statistical analysis, the observed data often does not fully conform to statistical model assumptions. For example, as stated in Hampel et al. (1986) "routine data are thought to contain 1% to 10% gross errors". Because of these errors, robust estimation plays an important role in statistical analysis. For example, in a regression study, the abnormal data points can severely distort the estimates and the true relationship between the covariates and the response. Consequently, the model's prediction ability is similarly distorted. In other applications, such as in statistical process control (SPC), robust statistics are also provided so that the control limits based on these statistics are not distorted by the abnormal measurements in the historical data set (HDS).

1.1 Robust Estimation in Regression

It is known that the regular ordinary least squares (OLS) estimator lacks resistance to as little as one unusual observation. The corresponding coefficients and their standard errors, predictions, diagnostics, hypothesis tests, and other numerical measures can all become very misleading due to a single anomalous observation. Robust procedures are designed to capture the general trend of the data in the presence of unusual data.

Most of the robust regression methodologies were provided by the early 1980's. For example, M regression (Huber (1981)), and bounded influence (BI) (Huber (1981)) regression work well in the presence of low leverage outliers and at least one high influence point, respectively. However, they are unable to combat a small percentage of outliers. Repeated sampling based methods, such as Least median squares (LMS) (Rousseeuw (1984)) regression and least trimmed squares (LTS) (Ruppert and Carroll (1980)) regression, on the other hand, are examples of high breakdown estimators as they possess the ability to provide reasonable parameter estimates with as much as 50%

of the data being contaminated. Poor efficiency and numerical/computational sensitivity with large datasets has typically led to their primary use as an initial estimator feeding into other robust procedures such as M or BI estimators. Examples include Mallows 1-step (M1S) regression (Simpson et al. (1992)) and Schweppe's 1-step (S1S) regression (Coakley and Hettmansperger (1993)), which are one-step adjustments of LTS that increase efficiency versus the LTS estimator. However, two virtually identical LTS estimates may yield dramatically different M1S (or S1S) estimators (Lawrence (2003)), thereby illustrating a potential negative issue with repeated sampling based methods. Another high breakdown one-step estimation method is due to Gervini and Yohai (2002). Their robust and efficient weighted least square estimate (REWLS) procedure attains full asymptotic efficiency with the assumption of normally distributed random errors. However, the REWLS, on the average, fails to correctly identify the good and bad high leverage points when the error term is not ideally normally distributed (Lawrence (2003)). A robust, efficient, high breakdown robust regression methodology was proposed by Lawrence (2003), called the cluster-based bounded influence regression (CBI) method, which combined a suitable clustering method with the bounded influence regression method. In this research, a revised version of this method is presented and evaluated in Chapter 2.

1.2 Robust estimation in SPC

As previously mentioned, statistical data sets frequently contains errors. Not surprisingly, such data anomalies can occur in the statistical process control setting. Robust estimation methods have also been proposed in SPC to avoid the misleading results from these errors.

1.2.1 Phase I and Phase II in SPC

The SPC involves two phases, Phase I and Phase II. In Phase I, a HDS is analyzed to determine which data points are from an in-control process and which ones, in any, are from an out-of-control process. Data points determined to be from an out-of-control process are usually removed and the remaining data points from the in-control

process are used to calculate the statistics needed for computing control limits used in Phase II analysis. In Phase II, future observations are monitored by using the control limits calculated from Phase I estimates to determine if the process continues to be in-control. The control limits in Phase I will directly affect the performance of Phase II analysis. Accurate control limits in Phase I are desirable for the Phase II analysis. This research also focuses on the estimation in Phase I and how these estimates affect performance in Phase II.

1.2.2 Robust estimation in Phase I

Recall that the purpose of the Phase I analysis is to examine the HDS and obtain the control limits that are sufficiently accurate for Phase II monitoring. However, like the estimates of regression analysis, the statistics used for control limits obtained from the HDS can be "pulled" in the direction of the multivariate outliers if the HDS contains data from an out-of-control process. Robust estimation techniques are used to obtain the control limits that are not unduly influenced by unusual data points. Consequently, the control limits will be more accurate and effective in Phase II analysis.

In most previous studies, products and processes were characterized by either univariate quality control data or multivariate quality control data. Robust estimation methods for univariate quality control data (such as those based on a median or trimmed mean) are straightforward and have received attention in past research (Rocke (1989); Tatum (1997); de Mast and Roes (2004); Cali Manning and Adams (2005)). Robust methods for multivariate quality control data are not as straightforward, nor as easily implemented.

When dealing with multivariate quality control data, it is assumed that the HDS consists of m time ordered vectors that are independent of each other. Frequently the Hotelling's T^2 statistics is used to determine if a multivariate data point results from an out-of-control process. In particular, if each vector is of dimension p and if $\hat{\boldsymbol{\mu}}_i$ denotes

a vector containing p elements for the i^{th} time period, then the Hotelling's T^2 statistic for the i^{th} time period is defined as

$$T_i^2 = \left(\hat{\boldsymbol{\mu}}_i - \bar{\boldsymbol{\mu}}\right)^T \hat{\boldsymbol{V}}^{-1} \left(\hat{\boldsymbol{\mu}}_i - \bar{\boldsymbol{\mu}}\right),\ i = 1, 2, ..., m. \tag{1.1}$$

where $\bar{\boldsymbol{\mu}} = \dfrac{\sum_{i=1}^{m} \hat{\boldsymbol{\mu}}_i}{m}$ and $\hat{\boldsymbol{V}}$ is an estimator of the variance-covariance matrix $\boldsymbol{V}$ of $\hat{\boldsymbol{\mu}}_i$ (see section 3 for more details). The Hotelling's T^2 is an example of a statistic that in not robust to outlying observations.

One commonly used robust T^2 statistic for multivariate data results by replacing the moment-based estimator of $\boldsymbol{V}$, the one typically used, by an estimator based on the minimum volume ellipsoid (MVE) estimator (Vargas (2003)). The MVE estimator, first proposed by Rousseeuw (1984), has been frequently used for the detection of the multivariate outliers. The MVE estimator seeks to find the ellipsoid of minimum volume that covers a subset of at least half of the total data points. One well known algorithm for MVE estimator is provided by Rousseeuw and Leroy (1987) is an approximate method using a sub-sampling procedure. However, the problem of this sub-sampling algorithm is that it lacks repeatability and results in estimates with poor efficiency. An exact method to calculate the MVE estimator was later proposed by Cook et al. (1993) to avoid the repeatability problem. However, this exact method is only computationally feasible for small datasets (Cook et al. (1993)). Other computationally feasible methods to find an approximate MVE have been proposed. For example, Hawkins (1994) proposed a feasible solution algorithm (FSA). Also, methods to find the MVE based on a heuristic search algorithms were proposed by Woodruff and Rocke (1993). The T^2 statistic for the i^{th} time period based on MVE is denoted by $T^2_{MVE,i}$ (Vargas (2003))

$$T^2_{MVE,i} = \left(\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_{MVE}\right)^T \hat{\boldsymbol{V}}^{-1}_{MVE} \left(\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_{MVE}\right), \tag{1.2}$$

where $\hat{\boldsymbol{\mu}}_{MVE}$ is the MVE estimator of multivariate mean and $\hat{\boldsymbol{V}}_{MVE}$ is the MVE estimator of multivariate variance-covariance matrix.

Another frequently used robust T^2 statistic for multivariate data is based on the minimum covariate determinant (MCD) estimator which was also proposed by Rousseeuw (1984). The MCD estimator is obtained by finding the half set of the data points that gives the minimum value of the determinant of the variance-covariance matrix. Similar to the MVE estimator, there are both approximate methods and exact methods to obtain the MCD estimates. For example, MCD estimates can be computed via the exact method provided by Cook et al. (1993). The sub-sampling approach of Rousseeuw and Leroy (1987) can be used to get an approximate MCD estimate which would have the same repeatability issue. The feasible solution algorithm of Hawkins (1993) can be implemented for the MCD, as shown by Hawkins (1994). An improved version of the feasible solution algorithm for the MCD was proposed by Hawkins and Olive (1999). The T^2 statistic for the i^{th} time period based on MCD is denoted by $T^2_{MCD,i}$ (Vargas (2003))

$$T^2_{MCD,i} = \left(\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} - \hat{\boldsymbol{\mu}}_{MCD}\right)^T \hat{\boldsymbol{V}}_{MCD} \left(\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_{MCD}\right), \tag{1.3}$$

where $\hat{\boldsymbol{\mu}}_{MCD}$ is the he MCD estimator of multivariate mean and $\hat{\boldsymbol{V}}_{MCD}$ is the MCD estimator of multivariate covariance matrix. Estimators based on the MVE and MCD are powerful in detecting a reasonable number of outliers as demonstrated by Jensen et al. (2007) and Yanez et al. (2010)). Other robust estimators have been proposed for the multivariate SPC setting. Yanez et al. (2010) proposed a T^2 statistic using S estimators based on the biweight function for the location and dispersion parameters when monitoring multivariate individual observations. They showed that this method outperforms the MVE estimators for a small number of observations. Other robust estimators defined using trimming, proposed by Alfaro and Ortega (2008) and Chenouri et al. (2009) and referred to as reweighted minimum covariance determinant (RMCD)

estimators, have been shown to provided highly robust and efficient estimators of the mean vector and covariance matrix.

1.3 Profile Monitoring in SPC

Another more recent approach to SPC occurs when the product or process can be characterized by a profile or a relationship between a response variable and one or more explanatory variables instead of univariate or multivariate vectors. The profile monitoring process in Phase I is first to represent the profiles in the HDS by some proper modeling technique, and use some appropriate method to identify those profiles from the in-control process, and those, if any, from the out-of-control process. As a final step, these in-control profiles are used to obtain the control limits for future profile monitoring in real-time during Phase II.

Further details concerning the profile monitoring literature will be given in Chapter 3. This dissertation will focus on monitoring the profiles using the mixed model where the mixed model is first fit to the profiles in the HDS using the mixed model technique to estimate the population average (PA) profile and the proper variance-covariance matrix. See chapter 3 for more details concerning the mixed model applied to profile monitoring.

1.4 Motivation

When using the mixed model technique for the profile monitoring, the first step is to estimate the PA profile and use this estimate in calculating the Hotelling's T^2 for each profile to determine whether this profile results from the in-control process. However, in the typical mixed model analysis, the estimated PA profile is based on all profiles in the HDS including the profiles from the out-of-control process. For example, if there is large amount of profiles that from the out-of-control process or small amount which are far away from the in-control process, the estimated PA profile based on the HDS would likely be "pulled" in the direction of the out-of-control process. Additionally, the corresponding variance-covariance matrix, needed in computing the

T^2 statistic for each profile, will be similarly distorted. Consequently, the T^2 statistics will be misleading and the in-control limits used in Phase I will be unable to properly separate those profiles belonging to the in-control process from those belonging to the out-of-control process.

In this research, a new profile monitoring methodology in Phase I which incorporates a cluster method will be utilized to obtain the estimated PA profile. This new cluster-based method will be demonstrated to be more robust to the profiles from the out-of-control process than the existing non-cluster-based method (see (Jensen et al. (2008)) for a thorough discussion of the non-cluster-based method).

Further, it is known that the performance of the Phase I analysis can be measured in terms of correctly identifying the unstable process or, equivalently, the presence of profiles from the out-of-control process in the HDS. An important criterion used to measure the success of a Phase I method at detecting an unstable process is the probability of signal (POS), the probability of detecting at least one profile from the out-of-control process in the HDS. However, the POS only measures the ability of detecting the presence of the profile from the out-of-control process in the HDS and does not give any information about whether the classification of profiles into the two categories of in-control and out-of-control is correctly specified.

A simple example is presented to illustrate that the POS is not sufficient to measure the performance of Phase I analysis. In this example, it assumed that there are total m=12 profiles in the HDS of which nine are from the in-control process while the other three are from the out-of-control process. The in-control profiles were generated from the linear mixed model

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})x_{ij} + (\beta_2 + b_{2i})x_{ij}^{\ 2} + \varepsilon_{ij},\ i = 1,2,...,m_1,\ j = 1,2,...,n, \qquad (1.4)$$

and the out-control profiles were generated as

$$y_{ij} = (\beta_0' + b_{0i}) + (\beta_1' + b_{1i})x_{ij} + (\beta_2' + b_{2i})x_{ij}^{\ 2} + \varepsilon_{ij}, \tag{1.5}$$

$$i = m_1 + 1, ..., m, \ \ j = 1, 2, ..., n,$$

where the random effects are defined as

$$\begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix} \sim MN\left(\mathbf{0} \ , \begin{bmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{bmatrix} \right),$$

$$\boldsymbol{\varepsilon} \sim MN\left(0, \ \ \sigma^2 \boldsymbol{I}\right),$$

(here MN represents the multivariate normal distribution) and with fixed effects $\boldsymbol{\beta}^T = \left(12.5, \ \ -7, \ \ 2\right)$ for the profiles from the in-control process and $\boldsymbol{\beta}^T = \left(21.875, \ \ -14.5, \ \ 3.5\right)$ for the profiles from the out-of-control process. Additionally, $m_1 = 9, \ m = 12, \ n = 8, \ \ \sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 0.5$ and $\sigma^2 = 4$. Thus, profiles 1 through 9 represent profiles from the in-control process and profiles 10, 11, and 12 represent profiles from the out-of-control process.

The 12 true profiles, based on the actual parameter values and random effects, are plotted in Figure 1.1 where the blue curves represent the profiles from the in-control process while the red curves represent the profiles from the out-of-control process. It is difficult to distinguish the profiles from the in-control and out-of-control process by looking only at the plot.

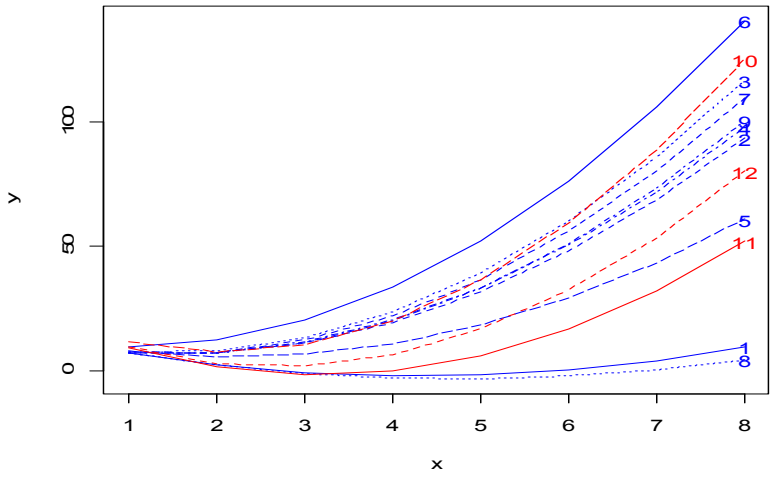

Figure 1.1: The plot of 12 true profiles

Using the T^2 statistic, both the existing non-cluster-based method and the proposed cluster-based method signaled, indicating that both methods detected a change in the process. However, the non-cluster-based method signaled due to misclassifying the 6th profile as the out-of-control process. The cluster-based method, on the other hand, correctly classified the 10th, 11th and 12th profiles as from the out-of-control process and classified the other nine profiles as from the in-control process. The estimates of the PA parameters from the non-cluster-based method (Jensen et al. (2008)) are $\hat{\boldsymbol{\beta}}^T = (16.261, \ -9.709, \ 2.178)$ while the estimates of the PA parameters from the proposed method are $\hat{\boldsymbol{\beta}}^T = (14.486, \ -7.764, \ 2.027)$. Compared to the true PA parameters, $\boldsymbol{\beta}^T = (12.5, \ -7, \ 2)$, the estimates of the non-cluster-based method (Jensen et al. (2008)) are severely distorted while the proposed method provided PA estimates much closer to the true values, as expected.

Chapter 2. Cluster-Based Bounded Influence Regression

Recall that robust regression estimation plays an important role in statistical analysis. In this chapter, a new robust and efficient regression method, called the cluster-based bounded influence (CBI) regression (Lawrence (2003)) will be reviewed. Additionally, the CBI regression algorithm will be updated by using the modern R package and compared to other existing robust regression methods.

2.1 Review of Robust Regressions

The detection of observations not conforming to a given statistical model is a common goal of the data analyst. Many methods have been proposed to aid in the detection of such nonconforming observations or "outliers". For example, in a recent paper by Fan et al. (2012a), a hierarchical clustering method was employed that greatly improves the ability of certain multivariate control chart techniques at detecting the presence of multivariate outliers. Detecting unusual observations in the multiple regression setting is a far more complicated process however and many techniques have been introduced (see section 2) for this purpose. As in the Fan et al. (2012a) paper, the use of clustering methodology can improve the ability of a technique to identify unusual data points in the multiple regression setting. The use of clustering to improve the properties of the bounded-influence regression method is demonstrated in this chapter.

In building a linear regression model, a single unusual observation can dramatically influence ordinary least squares (OLS) estimation. With OLS, a single low leverage outlier can have a dramatic effect on the estimation of the general trend, especially concerning the intercept. However, a single high influence point, or hip, can have a dramatic effect on any or all parameter estimates. The joint influence of several hips can have an even greater deleterious impact on parameter estimates. These

coefficients and their standard errors, along with predictions, diagnostics, hypothesis tests, and other numerical measures can each become very misleading without a thorough exploratory data analysis accompanying it.

This chapter focuses on the study of robust, high breakdown linear regression modeling. As this discipline is extremely computationally intensive, much of the published work in this area has occurred since the early 1980's. Of course, some ideas were proposed much earlier, but generally limited in actual application. Methods such as M regression (Huber and Ronchetti (2009)), and bounded influence (BI) (Huber and Ronchetti (2009)) regression work well in the presence of low leverage outliers and at most one hip respectively. However, they are unable to combat a small percentage of outliers. Least median of squares (LMS) (Rousseeuw (1984)) regression and least trimmed squares (LTS) (Ruppert and Carroll (1980)) regression, on the other hand, are examples of high breakdown estimators as they possess the ability to provide parameter estimates with as much as 50% of the data being contaminated. Poor efficiency and numerical/computational sensitivity with large datasets has typically led to their primary use as an initial estimator feeding into other robust procedures such as M or BI estimators. Examples include Mallows 1-step (M1S) regression (Simpson et al. (1992)) and Schweppe's 1-step (S1S) regression (Coakley and Hettmansperger (1993)), which are one-step adjustments of LTS that increase efficiency versus the LTS estimator. However, two virtually identical LTS estimates may yield dramatically different M1S (or S1S) estimators (Lawrence (2003)), thereby illustrating a potential negative issue with repeated sampling based methods. Another high breakdown one-step estimation method is due to Gervini and Yohai (2002). Their robust and efficient weighted least square estimate (REWLS) procedure attains full asymptotic efficiency with the assumption of normally distributed random errors. However, according to the Monte Carlo study in section 2.5, the REWLS, on the average, fails to correctly identify the good and bad high leverage points when the error term is not ideally normally distributed.

The CBI method was introduced by (Lawrence et al. (2013)) as a new regression methodology that obtains competitive, robust, efficient, high breakdown regression parameter estimates. Additionally, this method provides an informative summary regarding possible multiple outlier structure.

A simple example below gives the comparison of the CBI regression method to several existing robust procedures when the data has more than one high leverage point. The data set has 11 observations with observations 1-8 generated from the linear model

$$y_i = 100 - 4x_i + \varepsilon_i$$

where $\varepsilon_i \sim N(\mu = 0, \sigma^2 = 25)$, and with the regressor variable generated via $x_i \sim U[10,20]$. Observations 9-11 were arbitrary added to reflect a mild influence point and two hips, respectively.

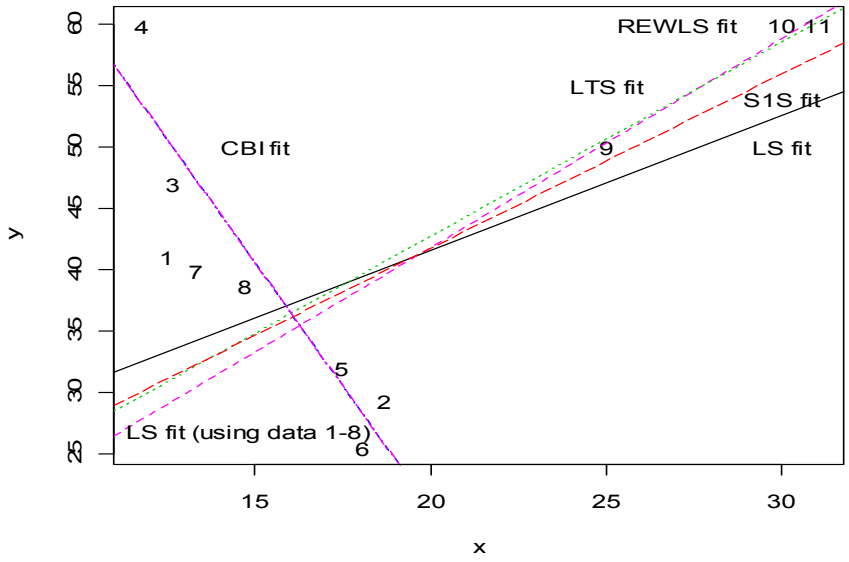

Figure 2.1: The fitted line of the different robust methods

The data are plotted in Figure 1.1 where the outlier (9) and the two hips (10, 11) are clearly seen. Regarding the collection of fits also displayed in Figure 1.1, only the

proposed method (CBI) detects the correct trend of the uncontaminated data. Each of the other estimators was dramatically misled by the joint influence of these three arbitrary points, resulting in a positive slope estimate when the true underlying slope is negative.

2.2 Review of Selected Robust Regression Methods

As the basis for linear regression analysis, the statistical model is restricted to be of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,$$

with the response variable, y_i, being explained as a linear function of the k regressor variables, x_{ji}, $j = 1,2 \ldots k,$ plus a random error component, ε_i, for each of the n observations, $i = 1,2 \ldots n$.

Given the computational nature of the proposed method, clarity in notation becomes quite important and, therefore, this paper offers sufficient detail. The linear model also can be written matrix form as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \text{, where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}),$$

or element wise as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

There are $p = k + 1$ unknown parameters that form the $p \times 1$ parameter vector $\boldsymbol{\beta}$, which is to be estimated by the $p \times 1$ vector $\widehat{\boldsymbol{\beta}}$**.** This subsequently yields the estimated fits as $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

Further, the $n \times 1$ vector of residuals is computed as $\boldsymbol{r} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, with r_i representing the residual for the i^{th} observation. Also, define $\boldsymbol{Z}$ as the $n \times k$ matrix

containing only the $k$regressor variables, with $\boldsymbol{Z}_y$ representing the $n \times p$ matrix formed by augmenting the vector $\boldsymbol{y}$ to $\boldsymbol{Z}$. To accommodate reference to individual observations, let the i^{th} row of $\boldsymbol{X}$ be denoted by the $1 \times p$ row vector $\boldsymbol{x}_i^T$ and the $1 \times k$ row vector $\boldsymbol{z}_i^T$ denote the i^{th} row of $\boldsymbol{Z}$. When the response variable is included, the notation for i^{th} row of $\boldsymbol{Z}_y$ is $\boldsymbol{z}_{y,i}^T$. Consider the objective function

$$\min_{\forall \boldsymbol{b}} \sum_{i=1}^{n} r_i^2,$$

for the OLS estimator, which may be written as

$$\min_{\forall \boldsymbol{b}} \sum_{i=1}^{n} \rho(r_i),$$

with $\rho(t) = t^2$. In robust regression, the function ρ can be selected to either down weight or bound any argument rising from unusual observations. This becomes the basis for M regression (Huber and Ronchetti (2009)) which has the objective function

$$\min_{\forall \boldsymbol{b}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{b}}{\hat{\sigma}}\right),$$

where the ρ-function is chosen to be bounded and odd-symmetric, $\boldsymbol{b}$ represents an arbitrary point in the p-dimension estimation space, and where $\hat{\sigma}$ is some appropriately chosen estimate of σ. The choice for $\hat{\sigma}$ is generally limited to robust measures of scale. One such estimator that is frequently used is the median absolute deviation (MAD), where

$$MAD = 1.4826 \operatorname*{med}_{\forall i} \left| r_i - \operatorname*{med}_{\forall i} r_i \right|.$$

Taking derivatives with respect to $\boldsymbol{b}$ leads to solving p "altered normal equations",

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{b}}{\hat{\sigma}}\right) \boldsymbol{x}_i = 0,$$

where $\psi(t) = \frac{d\rho(t)}{dt}$ and $\widehat{\boldsymbol{\beta}}_M$ is the solution for $\boldsymbol{b}$. These altered normal equations form a system of nonlinear equations that may be solved by a number of popular numerical methods including (1) Newton-Raphson and (2) iteratively reweighted least squares (IRLS), the later used in this paper. At convergence, IRLS produces the M regression parameter estimator

$$\widehat{\boldsymbol{\beta}}_M = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y},$$

where $\boldsymbol{W}$ is the $n \times n$ diagonal "weight matrix", with diagonal elements denoted as w_i. Each weight, w_i, determines how much emphasis the regression will place on a particular observation. A large weight (near 1) should indicate a good observation. An outlier or a hip, on the other hand, should get a reduced weight or perhaps even a zero weight. In M regression the i^{th} weight is calculated as $w_i = \frac{\psi(r_i/\hat{\sigma})}{r_i/\hat{\sigma}}$, a function of the i^{th} residual. Typically, the larger is the residual, the smaller is the weight.

A single hip will "pull" the fitted M regression line toward it to make the corresponding residual small, thus that weight will be large. This means that M regression can be dominated by a single hip. One solution to this problem is to use bounded influence (BI) regression. Here, the name refers to "bounding" the influence that the point $\boldsymbol{x}_i^T$ has in the regressor-space. One altered normal equation form, called the Schweppe form (Staudte (1990)), is written as

$$\sum_{i=1}^{n} u(\boldsymbol{x}_i) \psi\left(\frac{y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}}{\hat{\sigma} u(\boldsymbol{x}_i)}\right) \boldsymbol{x}_i = 0.$$

Here, $u(\boldsymbol{x}_i)$ is chosen so that the effect of a large $\boldsymbol{x}_i^T$ is reduced if $(y_i, \boldsymbol{x}_i^T)$ is a hip. One choice is to have $u(\boldsymbol{x}_i) = \pi_i = \frac{1-h_{ii}}{\sqrt{h_{ii}}}$, where h_{ii} is the i^{th} diagonal element of the so-called hat matrix, $\boldsymbol{H}$, with $\boldsymbol{H} = \boldsymbol{X}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}$. The π_i value is referred to as the BI weight. The BI regression estimator can be obtained in exactly the same manner as the M-estimator via IRLS, as

$$\widehat{\boldsymbol{\beta}}_{BI} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}.$$

However, the i^{th} weight now has the form $w_i = \psi\left(\frac{r_i^*}{\pi_i}\right)/\frac{r_i^*}{\pi_i}$, where r_i^* is the scaled residual $r_i/\widehat{\sigma}$. Specifically, the BI weight depends on both the residual and the location of $\boldsymbol{x}_i^T$ in the regressor-space.

While M and BI estimators provide an improvement over OLS if the data has an outlier or hip, respectively, they cannot provide protection against data with even modest amounts of contamination. Ruppert and Carroll (1980) introduced LTS to combat this situation, defining the objective function as

$$\min_{\forall \boldsymbol{b}} \sum_{i=1}^{h} r_{[i]}^2,$$

representing the sum of the h smallest squared residuals where h is generally taken to be $[(n + p + 1)/2]$, with [.] denoting the greatest integer function. Since this objective function is not differentiable, no closed-form expression exists for the LTS estimator. However, algorithms are available that give the exact LTS estimator for the location model, the exact LTS estimator for the regression model based on small data sets, and a relatively accurate LTS estimator for large data sets. The algorithmic details may be found in Rousseeuw and Van Driessen (2006). Historically, methods like LTS (and its

predecessor LMS) had involved repeated sampling computational methods incorporating probabilistic arguments.

One problem with high breakdown estimators such as LTS is poor efficiency due to large variability associated with estimated coefficients. The remedy for this poor efficiency is to use the LTS estimator, or another high breakdown estimator, as an initial estimator $\widehat{\boldsymbol{\beta}}_0$, with the generalized M estimator form to obtain a one-step generalized M estimator. The S1S estimator is one such estimator and results from solving the "altered" normal equations

$$\sum_{i=1}^{n} w_i \psi\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0}\right) \boldsymbol{x}_i = 0.$$

A Gauss-Newton approximation using a first-order Taylor series expansion about the initial estimate $\widehat{\boldsymbol{\beta}}_0$ yields a one-step improvement of the form

$$\widehat{\boldsymbol{\beta}}_{S1S} = \widehat{\boldsymbol{\beta}}_0 + (\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\psi(\underline{\boldsymbol{r}})\hat{\sigma}_0.$$

None of the above estimators achieve full efficiency at the normal distribution while simultaneously maintaining a breakdown bound close to 50%. Gervini and Yohai (2002) proposed an adaptive one-step estimation method that attains full asymptotic efficiency at the normal error distribution while at the same time has a high breakdown bound and small maximum bias. Their method, referred to as the REWLS estimator, is a weighted LS estimator computed from an initial high breakdown estimate $\widehat{\boldsymbol{\beta}}_0$, and a robust scale estimate $\hat{\sigma}_0$ such as MAD. However, rather than deleting those observations whose absolute scaled residuals are greater than a given value, the procedure will keep a number N of observations, corresponding to the smallest values of the absolute scaled residual $rs_i = \frac{|r_i(\widehat{\boldsymbol{\beta}}_0)|}{\hat{\sigma}_0}, i = 1, \ldots n$. The N has the property that in large samples under normality it will have $N/n \to 1$, which means a vanishing fraction of observations will

be deleted and full efficiency will be attained (Maronna et al. (2006)). The REWLS estimator can be obtained as

$$\widehat{\boldsymbol{\beta}}_{REWLS} = \begin{cases} \widehat{\boldsymbol{\beta}}_0 + (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} & if\ \hat{\sigma}_0 > 0 \\ \widehat{\boldsymbol{\beta}}_0 & if\ \hat{\sigma}_0 = 0 \end{cases},$$

where W is the diagonal matrix with

$$w_i = \begin{cases} 1 & if\ rs_i \leq rs_N \\ 0 & otherwise \end{cases}.$$

2.3 Cluster-Based Bounded Influence Regression

The CBI regression methodology offers a new philosophical approach to the robust regression arena and consists of two primary phases, the cluster phase and the regression phase. First, an initial high-breakdown regression estimator is produced via a sophisticated clustering algorithm. Second, refinement of this initial regression estimator is investigated and possibly implemented under a carefully structured use of BI regression. The rationale behind this second phase is to allow for a possible improvement in efficiency, especially when the level of data contamination does not come close to approaching 50%. The CBI regression method has been named cluster-based bounded influence regression, or CBI for short, to reflect the nature of its two phases computation process.

The cluster phase begins with high-breakdown location and scale estimation of the p dimensional regressor-response space. A special set of points, referred to as the set of anchor points, is computed that together represent the general trend of the data. Each observation is then characterized by the OLS regression fit that would occur if this individual observation is augmented to the anchor points. High breakdown location and scale estimation of this set of n OLS coefficients provides the foundation for the construction of the similarity matrix (technically, a distance matrix). The desire for a

tight, compact sphere of similar coefficients exhibiting a common trend description is the basis for the selection of complete linkage hierarchical clustering (Lawrence (2003)) as the default method and clustering is performed until an initial main cluster of at least $[(n+p+1)/2]$ observations are formed. Two aspects worth mentioning are that (1) the OLS sensitivity to a single point is being exploited to our advantage in evaluating the data, and (2) the anchor points serve to alleviate repeated sampling (as required by other 50% breakdown point estimators such as LTS) and the use of minimal sized elemental subsets that must be in general position (i.e. no singularity issues).

A simple OLS fit to this main cluster is used as the basis for the possible adjustment of the anchor set metric to more directly relate to the general trend. A revised similarity matrix is constructed, with a second cluster analysis yielding a revised, final main cluster and g minor clusters. The determination of this cluster classification structure completes the cluster phase.

To begin the regression phase, the initial CBI estimator is simply the OLS estimate of the main cluster observations. A high breakdown scale estimate is then computed. High breakdown BI leverage weights are computed from the regressor-space only. Using only the main cluster, a BI regression updates the initial CBI estimator. To this point, the minor clusters have not been utilized in the computation of the CBI regression estimator and their observations are said to be inactive. The activation process for these remaining observations has two primary stages. First, a $DFFITS_{+I}^2$ statistic is computed for each of the minor clusters, where $I = 1,2,\dots,g$. A candidate minor cluster is one such that $DFFITS_{+I}^2 < \delta$ for the cutoff value δ. Then, a single $DFFITS_{+J}^2$ statistic, denoted by J, is computed for the union of all candidate minor clusters. If $DFFITS_{+J}^2$ is "small enough", then the final CBI estimator is determined from this activation process (provided at least one minor cluster observation obtained a nonzero weight). Otherwise, the minor clusters do not play an active role (i.e. all observations possess a zero weight) and there is no further update to the current CBI regression

estimator. A final CBI scale estimate is computed once the final CBI regression estimator has been determined.

The detailed algorithm consisting of ten interrelated steps for the CBI estimator is presented below. Steps 1 through 3 represent the cluster phase and steps 4 through 10 represent the regression phase. Notation is introduced as needed.

Step1

Perform minimum volume ellipsoid, MVE, estimation (see Rousseeuw and Leroy (2003)) of $\boldsymbol{Z_y}$; determine the $(2p+1) \times p$ anchor point matrix, $\boldsymbol{\Omega}$. These points include $\boldsymbol{MVE_1(Z_y)}$, the MVE location vector for $\boldsymbol{Z_y}$, and the end points of the ellipsoid of constant distance $\chi^2_{0.975,p}$ from $\boldsymbol{MVE_1(Z_y)}$ based on the $\boldsymbol{MVE_2(Z_y)}$ metric, the MVE scale matrix estimator for $\boldsymbol{Z_y}$, the pair of end points is determinate by the expression $\boldsymbol{MVE_1(Z_y)} \pm \sqrt{\lambda_i \chi^2_{0.975,p}}\, \boldsymbol{e}_i$, where λ_i and $\boldsymbol{e}_i$ is the i^{th} eigenvalue and eigenvector of $\boldsymbol{MVE_2(Z_y)}$, respectively.

Step 2

Determine the $n \times p$ base regression estimator matrix $\boldsymbol{B}$. The i^{th} row of $\boldsymbol{B}$, denoted by the $1 \times p$ vector $\boldsymbol{b}_i$, is defined as the estimator that results from an OLS regression analysis of the set of anchor points supplemented by the addition of the i^{th} observation in the dataset. Perform an MVE estimation of $\boldsymbol{B}$, treating each row of $\boldsymbol{B}$ as an observation in p dimensions.

Step 3

Using $\boldsymbol{MVE_2(B)}$ as the distance metric, compute a $n \times n$ similarity matrix $\boldsymbol{S}$ whose elements are defined to be

$$s_{ij} = \left(\boldsymbol{b}_i - \boldsymbol{b}_j\right)^T \left(\boldsymbol{MVE_2(B)}\right)^{-1} \left(\boldsymbol{b}_i - \boldsymbol{b}_j\right).$$

Perform a cluster analysis on the dataset given the similarity matrix S and using complete linkage to obtain the tightest cluster of $\boldsymbol{b}_i$ vectors. The initial main cluster, C_0, is defined at the first instance of which a single cluster consists of at least $h = [(n + p + 1)/2]$ observations. The remaining observations fall into one of g minor clusters that are labeled as $C_1, C_2, \ldots, C_g$.

Step 4

Compute the OLS estimate $\widehat{\boldsymbol{\beta}}_0$ using the data points in C_0. A preliminary estimate of scale, $\hat{\sigma}_0$, is defined to be the MAD of all n residuals $\boldsymbol{r}(\widehat{\boldsymbol{\beta}}_0)$ where

$$r_i(\widehat{\boldsymbol{\beta}}_0) = y_i - \boldsymbol{x}_i{}^T\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}.$$

Determine the set of observations, H, such that

$$H = \{i: |r_i(\widehat{\boldsymbol{\beta}}_0)| \leq \hat{\sigma}_0 4.685\sqrt{2pn}/(n-2p)\}.$$

Step 5

Using the data points in H, compute the $p \times 1$ mean vector $\boldsymbol{m}_H(\boldsymbol{Z})$, of the regressor data in H,and $p \times p$ covariance matrix $\boldsymbol{V}_H(\boldsymbol{Z})$, using standard moments estimators, of the regressor data in H, define the $p \times 1$ robust regressor distance vector $\boldsymbol{d}$ containing the p elements

$$d_i = (\boldsymbol{z}_i - \boldsymbol{m}_H(\boldsymbol{Z}))^T(\boldsymbol{V}_H(\boldsymbol{Z}))^{-1}(\boldsymbol{z}_i - \boldsymbol{m}_H(\boldsymbol{Z})).$$

Step 6

Mimic step 1 to step 3 by replacing the MVE statistics with the weighted mean and covariance estimates for the data to get the new initial main cluster, C_0, and g minor clusters C_1, C_2 ..C_g.The weight for the i^{th} data point is define as

$$w_i = \begin{cases} 1, & i \in H \\ 0, & i \notin H \end{cases},$$

Compute the initial CBI estimator, $\widehat{\boldsymbol{\beta}}_1$, using WLS and subsequently updated the scale estimate $\hat{\sigma}_1$ as MAD of all n new residuals.

Step 7

Determine the $h \times 1$ BI leverage weight vector, $\boldsymbol{\pi}$, whose elements are defined as

$$\pi_i = \begin{cases} 1, & i \in C_0 \\ \min\left(1, \dfrac{\chi^2_{0.975,p-1}}{d_i}\right), & i \notin C_0 \end{cases}$$

Perform BI regression using only the main cluster, C_0, to obtain, at convergence of IRLS, the estimate $\widehat{\boldsymbol{\beta}}_2$.

Step 8

Let I represent any minor cluster and m_I be the size of I, and let $\boldsymbol{\pi}_{(C_0,C_I)}$ be the sub-vector set of $\boldsymbol{\pi}$ that corresponds only to the C_0 and C_I observations. Perform the BI regression with these new data points and leverage weight vector $\boldsymbol{\pi}_{(C_0,C_I)}$ to obtain the estimate $\widehat{\boldsymbol{\beta}}_{+I}$ at convergence. A $DFFITS^2_{+I}$ statistic is then computed via

$$DFFITS^2_{+I} = \frac{\sum_{i=1}^{n}\left(\hat{y}_{i,+I}(\widehat{\boldsymbol{\beta}}_{+I}) - \hat{y}_i(\widehat{\boldsymbol{\beta}}_2)\right)^2}{m_I \widehat{\sigma}_1^2},$$

where $\hat{y}_{i,+I}(\widehat{\boldsymbol{\beta}}_{+I})$ represent fits when using both C_0 and C_I observations and $\hat{y}_i(\widehat{\boldsymbol{\beta}}_2)$ represents fits when using just C_0 observations. This statistic is computed for each of the g minor clusters.

Step 9

Define the scalar δ to represent the maximum allowable $DFFITS_{+I}^2$ statistic. Then, let J represent the union of all activation candidate minor sets, i.e.

$$J = \bigcup_{\forall I} C_I \,|(DFFITS_{+I}^2 \leq \delta \text{ and } \exists_{\mathrm{i}\in\mathrm{I}}|\mathrm{w_i} > 0).$$

Provided that $J \neq \emptyset$, then with $\widehat{\boldsymbol{\beta}}_2$, $\hat{\sigma}_1^2$ and $\boldsymbol{\pi}_{(C_0,C_J)}$ as inputs to obtain the BI regression estimate $\widehat{\boldsymbol{\beta}}_{+J}$ and $DFFITS_{+J}^2$. The default value of δ is 4.

Step 10

$$\widehat{\boldsymbol{\beta}}_{CBI} = \begin{cases} \widehat{\boldsymbol{\beta}}_{+J}, & if\left(DFFITS_{+J}^2 \leq \delta \text{ and } \exists_{\mathrm{j}\in\mathrm{J}}\left|\mathrm{w_j} > 0\right)|J \neq \emptyset \\ \widehat{\boldsymbol{\beta}}_2, & otherwise \end{cases}$$

The CBI scalar estimator is then updated as the MAD of new residuals. The final CBI weights for the individual observations are simply the observations weights at convergence of BI regression used to compute $\widehat{\boldsymbol{\beta}}_{CBI}$.

Three scale estimators are provided by the CBI procedure, specifically $\hat{\sigma}^2{}_{CBI}$, $\hat{v}^2{}_{CBI}$ and $\hat{v}_w{}^2{}_{CBI}$. $\hat{\sigma}_{CBI}$ is the MAD of the CBI residuals. Given the CBI scale estimate $\hat{\sigma}_{CBI}$, the BI leverage weight vector $\boldsymbol{\pi}$, and $\widehat{\boldsymbol{\beta}}_{CBI}$, a robust mean square error that mimics the robust ANOVA scale estimate introduced by Birch (1992) is found via

$$\hat{v}^2{}_{CBI} = \frac{\frac{n^2}{n-p}\hat{\sigma}^2{}_{CBI}\sum_{i=1}^{n}\psi^2\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_{\mathrm{CBI}})}{\pi_i\hat{\sigma}_{CBI}}\right)}{\sum_{i=1}^{n}\psi'\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_{\mathrm{CBI}})}{\pi_i\hat{\sigma}_{CBI}}\right)}.$$

Using the effective sample size, $n_w = \sum_{i=1}^{n} w_i$ (Birch (2010)), a modified version of the robust analysis of variance scale estimate then becomes $\hat{v}_w{}^2$

$$\hat{v}_w{}^2{}_{CBI} = \frac{\frac{n_w{}^2}{n_w - p}\hat{\sigma}^2{}_{CBI}\sum_{i=1}^{n}\psi^2\left(\frac{r_i(\hat{\boldsymbol{\beta}}_{\mathrm{CBI}})}{\pi_i\hat{\sigma}_{CBI}}\right)}{\sum_{i=1}^{n}\psi'\left(\frac{r_i(\hat{\boldsymbol{\beta}}_{\mathrm{CBI}})}{\pi_i\hat{\sigma}_{CBI}}\right)}.$$

Once the CBI estimate is obtained, the BI based analysis of variance methods of Birch (1992) and Birch and Agard (1993) can be used to perform inference on any single parameter or any subset of parameters.

Many theoretical properties of the CBI estimator have been studied and proved by Lawrence (2003). For example, it has been demonstrated that the CBI regression estimator belongs to the family of high breakdown regression estimators; with a breakdown point approaching 50% as $n \to \infty$. It was further showed that the CBI estimator is asymptotically normally distributed. That is,

$$\sqrt{n}\left((\hat{\boldsymbol{\beta}}_{CBI} - \boldsymbol{\beta}\right) \xrightarrow{Law} N[\mathbf{0}, \boldsymbol{M}^{-1}\boldsymbol{Q}\boldsymbol{M}^{-1}],$$

where the $\boldsymbol{M}$ and $\boldsymbol{Q}$ is defined as

$$\boldsymbol{M} = E_F\left[\left(w + \frac{d\omega(\boldsymbol{x}, r)}{dr}r\right)\boldsymbol{x}\boldsymbol{x}^T\right],$$

$$\boldsymbol{Q} = E_F[\omega^2(\boldsymbol{x}, r)r^2\boldsymbol{x}\boldsymbol{x}^T],$$

$$\mathrm{w} = \omega(\boldsymbol{x}, r).$$

The function $\omega(\boldsymbol{x}, r)$, the weight function is nonnegative, bounded and measurable in $(\boldsymbol{x}, r)$. The CBI regression estimator has also been shown to achieve regression equivariance, scale equivariance and affine equivariance properties (see Rousseeuw and Leroy (2003) for definitions of these equivariance properties). These equivariance properties also impact the following Monte Carlo simulation study by the fact that the values defined for the regression coefficients and scale will not impact the final Monte Carlo results; i.e., these values are themselves arbitrary and meaningless. Overall, the theoretical foundation for the CBI methodology strongly supports its inclusion in the class of high breakdown regression estimators.

Reflection on the development of the CBI algorithm yields an interesting and diverse discussion onto itself. Motivation initially stemmed from an interest in how iteration breaks down M and BI estimators and a curiosity about joint influence diagnostics in general. The joint influence aspect itself led to the inclusion of some sort of clustering mechanism to identify these various subgroups of problematic observations. Many forms of the initial similarity matrix construct were considered, including one based on the altered hat matrix. Further, initial strategies were more spatially oriented and were utilizing single-linkage clustering to take advantage of the chaining property that is often considered a detrimental property of the method but could track a regression trend under this alternative use. In fact, such a CBI version was proposed early in its development (Lawrence (2003)).

A major breakthrough in the development of the CBI algorithm occurred with the introduction of the anchor set. Ironically, this thought arose during development of a closed-form computation method for a multivariate C_P statistic in a completely different research area. However, it was clear that this anchor set could alleviate the random subsampling with elemental sets issues (faced with the leading high-breakdown estimators) as it was large enough to fit the regression model without any singularity issues. Further, it had a direct implementation into the clustering framework. The exploitation of the OLS breakdown property would form the basis of this new paradigm. Common regression estimates would indicate common trends (either general trend or common deviant trend that would reflect joint influence) and, very importantly, there is no spatial requirement directly involved. Joint influence can involve observations scattered across the response-regressor space. As a direct consequence, clustering moved from single-linkage to complete-linkage to more appropriately capture what are effectively similar regression estimates.

Iteration has both beneficial and detrimental aspects, so the CBI algorithm had to be robust to such negative effects. Earlier versions of CBI allowed for minor clusters to be added sequentially. From the research, it was deemed more prudent to assess them

individually, then together, to avoid estimator drift due to iteration as well as to further bolster the robustness versus joint influence of several minor clusters.

Overall, while the technical and computational details of the CBI algorithm have evolved during the development process, the general philosophy and intent have remained steadfast. The goal was to take an efficient low-breakdown point method, BI regression, and improve the breakdown point while not making a huge sacrifice regarding efficiency. A more thorough discussion of the motivation of each step of the CBI algorithm may be found in Chapter 5 of Lawrence (2003).

2.4 Case Studies and Comparison

Two well-known datasets are used to illustrate and compare the CBI method to several other robust techniques, (1) the Pendleton and Hocking (1981) (PH) data, and (2) the Hawkins et al. (1984) (HKB) data.

The PH dataset has three regressors $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}$ and $n = 26$ observations. The parameters to be estimated are $\boldsymbol{\beta}^T = (20\ \ 3\ \ -2\ \ 0)$. Three low-leverage outliers were artificially created and inserted as observations 11, 17 and 18. One hip was inserted as observation 24.

The CBI cluster phase of the PH data resulted in a main cluster of 19 observations (four more than *h* =15) and five minor clusters. A summary of the entire CBI regression analysis is provided as Table 2.1 and Figure 2.2. The final CBI fitted equation is

$$\hat{y}_i = 25.615 + 2.719x_{1i} - 2.136\ \ x_{2i} - 0.194x_{3i}.$$

It is clear (p-value = 0.331) that $\boldsymbol{x_3}$ is not significant in the presence of $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, a correct decision for this case study. The intercept, $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are each statistically significant (p-values of 0.038, 0.000and 0.000, respectively) terms, as they should be.

According to the CBI weight plot in Figure 2.2, four observations received zero weight, these being the three outliers and the one hip.

Table 2.1: Summary of the CBI regression analysis of the PH dataset

Cluster History				
Step	Clusters			n=26
Initial	$C_0 = \{2\ ,5\ ,7\ ,8\ ,9\ ,12:16\ ,19:23\}$			h=15
Final	$C_0 = \{1::5\ ,7\ ,8,10,12:16,19,21:23,25\ ,26\}$ $C_1 = \{6\ ,9\ ,20\} C_2 = \{11\}$ $C_3 = \{17\} C_4 = \{18\}$ $C_5 = \{24\}$			Initial OLS :
				intercept: 26.987
				X_1: 2.601
				X_2: -2.108
				X_3: -0.173
Minor Sets	$\text{DFFITS}_{+I}{}^2$			Activate
C_1	1.9153			YES
C_2	0			NO
C_3	0			NO
C_4	0			NO
C_5	0			NO
Candidate J	$\text{DFFITS}_{+J}{}^2$			Activate
C_1	1.9153			YES
Parameter Estimate				
Parameter	Estimate	Sd.Error	t	P -value
intercept	25.615	13.677	1.873	0.038
X_1	2.719	0.695	3.909	0.000
X_2	-2.136	0.321	-6.638	0.000
X_3	-0.194	0.441	-0.440	0.331
Scale	$\hat{\sigma}_{CBI} = 0.516 \quad \hat{v} = 0.306 \quad \hat{v}_w = 0.254$			

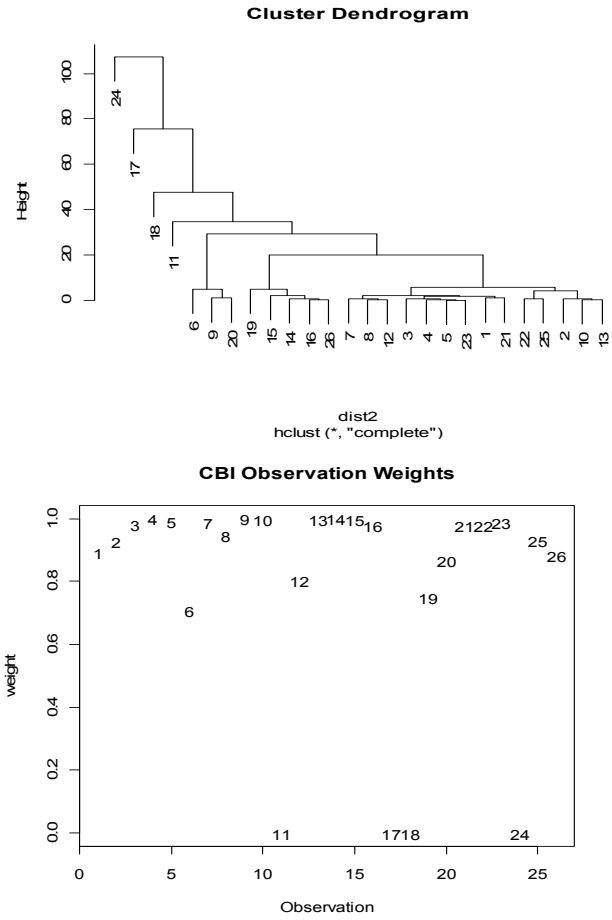

Figure 2.2: Cluster dendrogram and final observation weights of PH dataset

Other competing regression methods are applied to the PH dataset and the corresponding estimates are given by Table 2.2.

Table 2.2: Robust analysis of parameter estimate summary of PH dataset

Parameter	OLS	LTS	S1S	REWLS	BI	CBI	LS without outliers
Intercept	8.205	10.961	40.96	8.931	17.954	25.615	24.270
X_1	3.560	3.384	1.974	3.523	3.120	2.719	2.791
X_2	-1.640	-1.712	-2.538	-1.697	-1.971	-2.136	-2.112
X_3	0.334	0.483	-0.781	0.4337	0.052	-0.196	-0.156

The estimated coefficients resulting from the different estimation methods described in Section 1 for the PH data reveal some interesting results, especially as they

relate to the CBI algorithm. First, it is seen that the BI estimator has coefficient estimates very close to the true parameter vector. The CBI estimator began with estimates based on the final main cluster and then improved upon them through the minor cluster activation process. It is interesting to note that the estimated coefficients using the OLS method for the 22 good observations is nearly identical to those obtained by the CBI method. Thus, the CBI estimator is actually closer to the observed trend of the data than is the BI estimator.

We note that the PH data had no troublesome jointly influential observations. Consider next the HBK data which has a cluster of ten hips (as observations 1 through 10) and another cluster of four good high leverage points (observations 11 through 14). Since the true parameters were not reported by Hawkins et al. (1984), the goal in analyzing this dataset was to ascertain the ability of the robust methods to distinguish between the outliers and the non-outliers occurring at the high leverage points.

The CBI method applied to the HBK data resulted in a weight for each observation (Figure 2.3). Figure 2.3 shows that the first ten observations received zero weight, the ideal case. The four good leverage observations, on the other hand, all have weights greater than zero, as they should be; especially the observations 11, 12 and 14 received very high weights each close to 1.

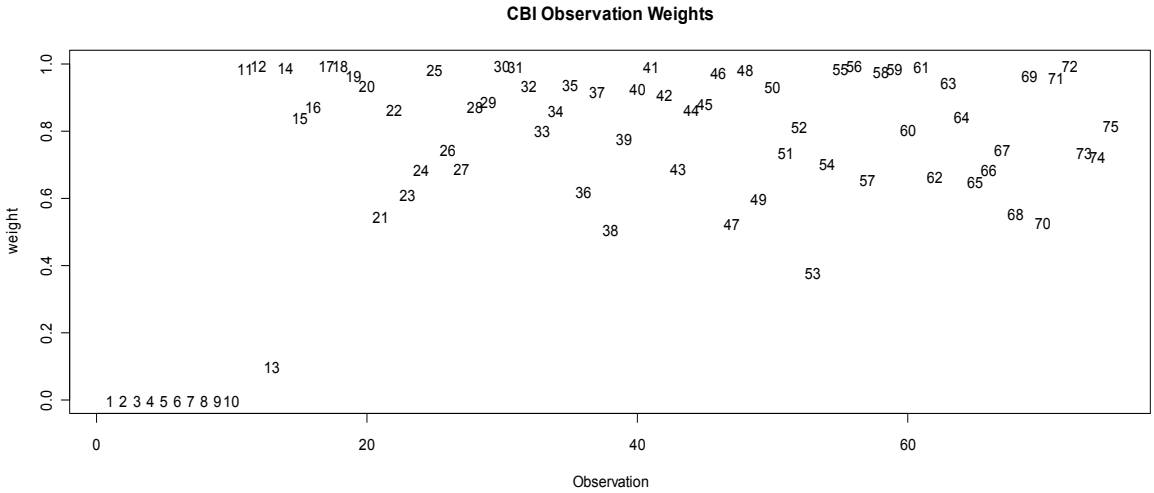

Figure 2.3: The final CBI regression observation weights of HBK dataset

The final CBI fitted equation is

$$\hat{y}_i = -0.224 + 0.097\ x_{1i} + 0.045x_{2i} - 0.064x_{3i}.$$

A summary of the CBI regression analysis is provided as Table 2.3.

Table 2.3: CBI analysis of parameter estimate summary of HBK dataset

Parameter Estimate				
Parameter	Estimate	Sd. Error	t	P -value
intercept	-0.224	0.169	-1.326	0.190
X_1	0.097	0.107	0.901	0.371
X_2	0.045	0.061	0.736	0.464
X_3	-0.064	0.055	-1.165	0.249
Scale	$\hat{\sigma}_{CBI} = 0.867 \quad \hat{v} = 0.890\ \hat{v}_w = 0.646$			

A comparison of the CBI results to other competing regression methods is given in Table 2.3. It is seen that the REWLS estimate provide the same result as the OLS estimate without hips, this result is not surprising because the REWLS method took advantage of the fact that the hips in this case have larger residuals as determined by its initial LTS estimate. The CBI estimates, while not identical, are very close to the OLS estimates based on the good data points.

Table 2.4: Robust analysis of parameter estimate summary of HBK dataset

Parameter	**OLS**	**LTS**	**S1S**	**REWLS**	**BI**	**CBI**	**OLS without hips**

Intercept	-0.388	-0.612	-0.004	-0.180	-0.934	-0.224	-0.180
X_1	0.239	0.255	0.041	0.081	0.144	0.097	0.081
X_2	-0.335	0.048	0.021	0.039	0.192	0.045	0.039
X_3	0.383	-0.106	-0.082	-0.051	0.184	-0.064	-0.051

The CBI estimates are close to the trend of the data for both case studies and the weight plots also show that it can correctly identify the outliers and hips for the case studies considered here. Results from a small Monte Carlo study are presented in the next section to further evaluate the ability if the competing regression methods to detect multiple outliers, especially those occurring at high leverage points.

2.5 Monte Carlo Study

In this Monte Carlo study, the simulated dataset utilized the original regressor values of the HBK dataset, but generated a new response vector while maintaining observations 1 through 10 as a high influence cluster. Specifically, the $n = 75$ observations were generated by the linear model

$$y_i = \begin{cases} \varepsilon_i, & i \in (1{:}10) \\ 0.2 - 0.15x_{1i} + 0.1x_{3i} + \varepsilon_i, & i \notin (1{:}10) \end{cases}.$$

With the random errors generated from the following distributions

$$\varepsilon_i \sim \begin{cases} N(\mu = 10, \sigma^2 = 0.385^2), & i \in (1{:}10) \\ N(\mu = 0, \sigma^2 = 0.5^2), & i \notin (1{:}10) \end{cases}.$$

The results of this Monte Carlo study are provided in Table 2.5. Here, the parameters to be estimated are $\boldsymbol{\beta}^T = (0.2 - 0.15 \;\; 0.1)$ and ${\sigma_\varepsilon}^2 = 0.25$. The number of Monte Carlo repetitions was 2000.

According to the characteristics of the estimators in Table 2.5, it is seen that CBI estimator had overall better performance. For example, consider $\hat{E}[\hat{\boldsymbol{\beta}}]$, the simulated

expected coefficient vector for each estimation method. We see that S1S and CBI were similar, with little exhibited bias. LTS and REWLS, on the other hand, were very close to each other, demonstrated a moderate bias. OLS and BI were severely biased as expected. All simulated scale estimates, $\hat{E}[\hat{\sigma}^2]$, overestimated, on the average, the true scale parameter of 0.25. On the other hand, the simulated robust scale estimate, $\hat{E}[\hat{v}^2]$ for the BI procedure severely underestimated the scale parameter. This led to the smallest expected standard errors of the BI coefficients, results based on the average of the simulated coefficient standard errors using the average square root of the diagonal value of $\hat{v}^2 * (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$ matrix. Among the scale estimates, the robust scale estimate based on the effective sample size, $\hat{v}_w{}^2$, for the CBI procedure had the smallest bias, on the average.

Between CBI and S1S, the CBI coefficients had the smaller standard error and were more stable, both in terms of the observed range as well as with respect to the IQR. The REWLS improved the stability of LTS and had smaller standard error for its coefficients. Both OLS and BI exhibited very tight distributions for each of the four coefficients was of little consequence given the extreme bias that was exhibited.

Table 2.5: Simulation results for Monte Carlo study

(*The crossed cells are not applicable*)

	OLS	**LTS**	**S1S**	**REWLS**	**BI**	**CBI**	
$\hat{\mathrm{E}}[\hat{\boldsymbol{\beta}}]$	0.029	−0.104	0.209	−0.105	−0.424	0.218	
	−0.019	−0.122	−0.156	−0.125	−0.090	−0.147	
	−0.307	0.069	−0.005	0.063	0.119	−0.006	
	0.456	0.188	0.105	0.189	0.295	0.093	
$\hat{E}[\hat{\sigma}^2]$	3.478	0.363		0.312	0.358	0.329	$\hat{E}[\hat{v}_w{}^2]$
$\hat{E}[\hat{v}^2]$					0.016	0.469	0.276
$\hat{\mathrm{E}}[\mathrm{se}[\hat{\boldsymbol{\beta}}]]$	0.345		0.211	0.114	0.024	0.177	0.141
	0.217		0.075	0.069	0.015	0.095	0.076
	0.128		0.073	0.053	0.013	0.073	0.057
	0.107		0.074	0.042	0.009	0.065	0.051
	0.689 −0.310 0.379 0.127	2.333 −1.168 1.165 0.665	2.316 −0.570 1.746 0.298	1.848 −1.077 0.7 0.630	0.619 −0.735 −0.115 0.080	1.903 −0.759 1.144 0.189	

$\widehat{\boldsymbol{\beta}}$	0.401	0.977	1.066	0.660	0.354	0.642
Range	−0.223	−0.597	−0.693	−0.372	−0.251	−0.466
Min	0.179	0.329	0.373	0.373	0.102	0.176
Max	0.078	0.192	0.094	0.111	0.080	0.090
IQR	0.235	0.901	0.869	0.556	0.264	0.637
	−0.423	−0.400	−0.473	−0.137	0.141	−0.307
	−0.188	0.501	0.396	0.419	0.405	0.329
	0.046	0.183	0.102	0.125	0.048	0.083
	0.202	1.019	0.945	0.504	0.265	0.543
	0.364	−0.313	−0.549	−0.029	−0.405	−0.198
	0.566	0.706	0.396	0.475	0.140	0.345
	0.039	0.193	0.113	0.193	0.005	0.073

The average observation weights are denoted as $\overline{\boldsymbol{w}}$, and the standardized average weight $\overline{ws}_i$ is defined as

$$\overline{ws}_i = \frac{\overline{w}_i - Min(\overline{\boldsymbol{w}})}{Max(\overline{\boldsymbol{w}}) - Min(\overline{\boldsymbol{w}})}.$$

Table 2.6: Standardized average weight for observations 1-14

Observation	REWLS	BI	CBI	Observation	REWLS	BI	CBI
1	0.014	0.971	0.002	8	0.096	0.981	0.000
2	0.144	0.991	0.002	9	0.131	0.985	0.004
3	0.159	0.991	0.005	10	0.186	0.992	0.006
4	0.000	0.964	0.001	11	0.243	0.000	0.614

5	0.107	0.987	0.005	12	0.243	0.000	0.533
6	0.132	0.986	0.002	13	0.248	0.000	0.677
7	0.152	0.990	0.003	14	0.231	0.000	0.646

Considering the result in Table 2.6, the CBI, on average, was more likely to identify the hips. For example, it gave almost 0 weights on the average to all the hips and weights greater than 0.5 to all the good leverage points. The REWLS, ended with the low weights to all the bad and good leverage points. The BI, on the other hand, mistakenly attributed the weights, provided very high weight for the first ten bad leverage points and 0 weights for the four good leverage points.

2.6 Chapter Summary

In this chapter, a robust and efficient regression methodology, called the cluster-based bounded influence regression is reviewed and updated by using the modern software package R. Both the case studies and the Monte Carlo study show that this regression methodology is competitive with methods such as LTS (Ruppert and Carroll (1980)), S1S (Coakley and Hettmansperger (1993)) and REWLS (Gervini and Yohai (2002)) when the data is highly contaminated but also be able to compete with the efficient M and BI regression methods (Huber and Ronchetti (2009)) when the data has few or no problematic observations. Specifically, the first case study shows that the CBI outperformed the other high breakdown procedures under the low contamination situation. The Monte Carlo study, on the other hand, shows that the CBI is one of the two procedures (S1S and CBI) that provide unbiased regression coefficients. Between

the unbiased procedures, the CBI has the smaller standard errors of the regression coefficients and has more stable of the coefficient estimates.

Chapter 3. Profile Monitoring Literature

Profile monitoring is a well-known approach in SPC that is widely used when the quality of the product or process is characterized by a profile or the relationship between a response variable and one or more explanatory variables. There are two initial steps in the profile monitoring procedure during Phase I analysis (see section 3.1). The first step is to represent each profile using model building theory and the second one is to detect the presence of profiles from the out-of-control process, those profiles caused by special variability, using quality control methodology.

3.1 Phase I and Phase II

Profile monitoring and statistical process control in general, is conducted over two phases, labeled as Phase I and Phase II. In Phase I, a HDS is analyzed to determine which profiles among the data represent the process when in-control and those profiles that represent the process when not in-control. The profiles representing the in-control process are then used to establish the control limits for monitoring new profiles as they become available in real time, the Phase II component of profile monitoring. However, the control limits established in Phase I may undermine the resulting performance in Phase II if the HDS contains anomalies such as trends, step changes and other types of instability. Thus, it is important to calculate the control limits using the stable process data contained in the HDS in Phase I. The goal in Phase I analysis is to separate the stable process data from the unstable process data in the HDS, remove the data from the out-of-control process and then use the data from the in-control-process to estimate the control limits. The performance of the Phase I analysis can be measured in terms of correctly identifying the unstable process in the HDS, which is usually represented as the POS, the probability of detecting at least one profile from the out-of-control process in the HDS.

Phase II consists of monitoring future profiles with the control limits obtained from Phase I analysis to determine the on-going stability of the process. Performance of Phase II is often measured by the average length run, which is the average number of samples taken until the first out-of-control signal. More details for the difference of analyses between Phase I and Phase II can be obtained by referring to Sullivan (2002), Mahmoud and Woodall (2004), and Montgomery (2009). In this chapter, the focus is on Phase I, detecting the unstable process data and using the stable process data to establish the control limits for Phase II. The impact of proper estimates for successful Phase II control charts is discussed in Chapter 5.

3.2 Profile Monitoring Literature Review

In past SPC applications, univariate or multivariate quality characteristics were typically used to represent the quality of a process or product. Recently however, it is becoming more common to use a profile, a response variable and one or more explanatory variables, to characterize the quality of a process or product. Monitoring these estimated profiles using quality control techniques is referred to as profile monitoring. Woodall et al. (2004) ,Woodall (2007) and Noorossana et al. (2012) presented an introduction and literature reviews on this subject.

In Phase I, one needs to fit each profile first using some appropriate modeling technique. In some applications, the profile can be represented adequately by a linear regression function. Croarkin (1982), Stover and Brill (1998), Kang and Albin (2000), Kim et al. (2003), Mahmoud and Woodall (2004), Wang and Tsung (2005), Gupta et al. (2006) and Zhang et al. (2009) all considered of the use of linear profiles. In many other cases, profiles may not be well-modeled by a linear regression function. Nonlinear profile applications were studied by Jin and Shi (2001), Lada et al. (2002), Walker and Wright (2002), Ding et al. (2006), Gupta et al. (2006), Williams et al. (2007a) and Williams et al. (2007b).

However, neither linear nor nonlinear profile monitoring methods discussed above incorporate the situation when the data within the profile are correlated rather than independent. For example, in the repeated measures situations, the subjects are repeatedly measured and the responses within the same subject are very likely to be correlated. Thus, in longitudinal studies, it may be incorrect to assume that the data from the same subject are independent. The mixed model is preferred for cases such as repeated measures situations or longitudinal studies when the data are grouped or clustered. Jensen et al. (2008) and Jensen and Birch (2009) proposed the use of the linear and nonlinear mixed model to monitor linear and nonlinear profiles in order to account for the correlation structure within a profile.

One of the basic assumptions for using parametric fixed and/or mixed models is that the response variable can be adequately modeled by a well-defined parametric function of both fixed effects and random effects. That is, the underlying relationship between response and explanatory variables is parametric. However, this assumption is not always satisfied in practical applications. For example, Härdle (1992), Fan and Gijbels (1996), Green and Silverman (1994), and Ramsay and Silverman (2002) among others, provide data examples where it is not possible to be adequately describe the profile with any parametric model. In these cases, nonparametric regression modeling techniques are proposed to monitor profiles based on a nonparametric regression method. Qiu (2010) also used nonparametric regression profile monitoring when the data within each profile are correlated. Other nonparametric regression profile monitoring methods are presented by Reis and Saraiva (2006), Jeong (2006) and Chicken et al. (2009). There are many existing smoothers that can be used to fit the nonparametric regression model. Different smoothers have different strengths in one aspect or another. For example, smoothing splines may be good for handling sparse data, while local polynomial smoothers may be computationally advantageous for handling dense designs (Wu and Zhang (2006)). The four most popular smoothers include local polynomial smoothers (Fan and Gijbels (1996), Doruska (1998)), regression splines (Eubank (1988) and Eubank (1999)), smoothing splines (Wahba (1990), Green and Silverman (1994), Wang (2011)), and penalized splines (Ruppert et al. (2003)).

A combination of parametric and nonparametric methods to represent the profiles was introduced by Abdel-salam (2009). In this work, monitoring profiles via a procedure referred to as model robust profile monitoring (MRPM), a semiparametric procedure, which combines the parametric fit to the profiles with the nonparametric profile fits via an appropriate linear combination was considered. The resulting MRPM fit can be "better" than either the parametric or nonparametric fits, especially when the parametric model has been misspecified.

After correct representation of the profiles has been achieved in Phase I, the second step is to detect the data from the out-of-control process and obtain the in-control limits necessary for Phase II. One current method for Phase I monitoring based on mixed models (Jensen, et al. (2008) and Jensen and Birch (2009)) for detecting the profiles from the out-of-control process is to compare each estimated profile specific (PS) curve to the estimated PA curve using the T^2 statistic (to be discussed in detail in the next section). Some authors, for example, Kang and Albin (2000), Kim et al. (2003), and Mahmoud and Woodall (2004), under the assumption that the parametric model was correctly specified, utilized the T^2 statistic to determine profiles from the out-of-control process based on the estimated parameters. Jensen, et al. (2008) proposed the use of the T^2 statistic approach to determine profiles from the out-of-control process in the parametric mixed model and extended it by using the T^2 statistic based on the estimated best linear predictors (eblups) on the eblups of each profile.

3.3 Multivariate T^2 Statistics

In order to develop the T^2 statistic to monitor profiles, first consider the general framework of the multivariate T^2 statistic. Given a sample of m independent observation vectors to be monitored, $\hat{\boldsymbol{\mu}}_i, i = 1, 2, ...m$, each of dimension p the general form of the T^2 statistic in Phase I for observation i is

$$T_i^2 = \left(\hat{\boldsymbol{\mu}}_i - \bar{\boldsymbol{\mu}}\right)^T \hat{\boldsymbol{V}}^{-1} \left(\hat{\boldsymbol{\mu}}_i - \bar{\boldsymbol{\mu}}\right),\ i = 1, 2, ..., m. \tag{3.1}$$

here $\bar{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{m} \hat{\boldsymbol{\mu}}_i}{m}$ and $\hat{\boldsymbol{V}}$ is an estimator of the variance-covariance matrix $\boldsymbol{V}$ of $\hat{\boldsymbol{\mu}}_i$. The observation is considered abnormal if the T_i^2 value exceeds the upper control limit.

There are several candidates for the variance-covariance matrix estimator. In the profile monitoring literature two estimators are commonly used. The first one is the pooled sample variance-covariance matrix, $\hat{\boldsymbol{V}}_p$, computed as

$$\hat{\boldsymbol{V}}_p = \frac{1}{m-1} \sum_{i=1}^{m} \left(\hat{\boldsymbol{\mu}}_i - \bar{\boldsymbol{\mu}}\right)\left(\hat{\boldsymbol{\mu}}_i - \bar{\boldsymbol{\mu}}\right)^T \tag{3.2}$$

The T^2 statistic based on the sample mean vector and pooled sample variance-covariance is widely used, but it is ineffective in cases containing single moderately abnormal observations (Vagas (2003)). The T^2 statistics are approximately chi-square distributed for large sample sizes while it is proportional to beta distribution for small sample sizes (Mason and Young (2002)).

The second variance-covariance estimator, $\hat{\boldsymbol{V}}_D$, is known as the successive difference estimator, which was first introduced by Hawkins and Merriam (1974) and used by Holmes and Mergen (1993). To obtain $\hat{\boldsymbol{V}}_D$, let

$$\hat{\boldsymbol{d}}_i = \hat{\boldsymbol{\mu}}_{i+1} - \hat{\boldsymbol{\mu}}_i, \; i = 1, 2, ..., m-1. \tag{3.3}$$

Then stack the transpose of these $m-1$ successive difference vectors into the $(m-1) \times p$ matrix $\hat{\boldsymbol{D}}$, that is

$$\hat{\boldsymbol{D}} = \begin{bmatrix} \hat{\boldsymbol{d}}_1^T \\ \hat{\boldsymbol{d}}_2^T \\ \vdots \\ \hat{\boldsymbol{d}}_{m-1}^T \end{bmatrix} \tag{3.4}$$

The formula of the successive difference estimator is given as

$$\hat{\boldsymbol{V}}_D = \frac{\hat{\boldsymbol{D}}^T \hat{\boldsymbol{D}}}{2(m-1)} \tag{3.5}$$

Sullivan and Woodall (1996) showed that $\hat{\boldsymbol{V}}_D$ is effective in detecting sustained step changes in the production process that occurs in Phase I data. The asymptotic distribution of the T^2 statistic based on successive difference is a chi-square distribution with appropriate degrees of freedom. Small sample properties of the T^2 statistic based on $\hat{\boldsymbol{V}}_D$ can be found in Williams et al. (2006) .

3.4 Profile Monitoring for Mixed Model

The mixed model is a model that contains both fixed effect terms and at least two random effect terms including the error term. In the literature on the mixed model, a collection of data on each experimental unit forms a "profile" or a "cluster" or a "subject", depending on the particular application. The term "profile" will be used here. The mixed model is flexible and capable of fitting a large variety of datasets. There are several advantages of the mixed model over the fixed model including allowing the modeling of the correlation structure within profiles and the interpretation of profiles as a random sample from a population distribution.

The general form of the mixed model is introduced here for the case where the profiles can be expressed by only one regressor and with balanced data for each profile. The model can be easily extended to deal with more than one regressor and the unbalanced data case. The general mixed model can be written as (Abdel-Salam (2009))

$$y_{ij} = f\left(x_{ij}\right) + \xi_i\left(x_{ij}\right) + \varepsilon_{ij},\ i = 1,2,...,m,\ j = 1,2,...,n. \tag{3.6}$$

where $f\left(x_{ij}\right)$ represents the mean response function for all profiles, the PA, $\xi_i\left(x_{ij}\right)$ represents the random effects for the i^{th} profile, where $\xi_i\left(x_{ij}\right)$ follows some appropriate distribution. For example, $\xi_i\left(x_{ij}\right)$ can be expressed as

$$\xi_i\left(x_{ij}\right) = b_{0i} + b_{1i}x_{ij}, \tag{3.7}$$

a random simple linear regression model. The random variables b_{0i} and b_{1i}, the random intercept and random slope, respectively, can be assumed to be jointly normally distributed as $\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim MN(\boldsymbol{0}, \boldsymbol{G})$, a multivariate normal distribution with $\boldsymbol{G}$ as the 2×2 variance-covariance matrix of b_{0i} and b_{1i}. In (3.6), ε_{ij} is the error term, distributed as $\boldsymbol{\varepsilon} \sim MN\left(\boldsymbol{0}, \boldsymbol{R}\right)$, where $\boldsymbol{R}$ is the $n \times n$ variance-covariance matrix for the n error terms.

The two components in (3.6), f and ξ_i, may be both parametric. If so, (3.6) is referred to as a parametric mixed model. If both components are nonparametric then (3.6) is referred to as a nonparametric mixed model. Or, if one component is parametric and the other is nonparametric, then (3.6) is referred to as a semiparametric mixed model.

3.4.1 Linear Mixed Models and its Parametric Estimation

A linear mixed model (LMM) is a mixed model of form (3.6) where both $f\left(x_{ij}\right)$ and $\xi_i\left(x_{ij}\right)$ can be expressed as linear functions of the parameters. The LMM is also an extension of the linear fixed model in which the LMM incorporates at least one additional random effect term other than the error term. An introduction to the LMM can

be found in Verbeke and Lesaffre (1996), Verbeke and Molenberghs (2000), Pinheiro and Bates (2000), Schabenberger and Pierce (2002), and Demidenko (2004).

Suppose the true mixed model (the so-called "true model") for the i^{th} profile can be written as (3.6) for arbitrary functions f and ξ_i .The model in (3.6) can be approximated by a linear mixed model, referred to as the Laird-Ware (L-W) model, for the i^{th} profile as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i,\ i = 1,2,...,m\ \ j = 1,2,...,n_i, \tag{3.8}$$

where $\boldsymbol{y}_i$ is the $n_i \times 1$ response vector for the i^{th} profile, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $n_i \times p$ and $n_i \times q$, respectively, matrices of explanatory variables, $\mathbf{b}_i$ is a $q \times 1$ vector of random effects for the i^{th} profile with $\boldsymbol{b}_i \sim MN(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{G}$ is a $q \times q$ covariance matrix. $\boldsymbol{\varepsilon}_i$ is the random error term for the i^{th} profile with $\boldsymbol{\varepsilon}_i \sim MN\left(\mathbf{0}, R_i\right)$. The interpretation of (3.8) is that the term $\boldsymbol{X}_i\boldsymbol{\beta}$ represents the PA curve at the regressor values in $\boldsymbol{X}_i$ while the $\boldsymbol{Z}_i\boldsymbol{b}_i$ term represent the random departures from the PA that are specific to the i^{th} profile. Together, the term $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$ represents the profile curve specific to the i^{th} profile, denote by PS_i .

As described above, the L-W model is extremely flexible in that it allows random errors to be independent or correlated. If correlated, $\boldsymbol{R}_i$ is often assumed to be a simple form such as the autoregressive form or the compound symmetry form to reduce the number of unknown covariance parameters that require estimation. Similar structure can be used for $\boldsymbol{G}$, but usually $\boldsymbol{G}$ is restricted to a diagonal matrix resulting in independent random effects. Also, it is assumed that $Cov\left(\boldsymbol{\varepsilon}_i, \boldsymbol{b}_i\right) = \boldsymbol{O}$ where $\boldsymbol{O}$ is the $n_i \times q$ matrix of zeros.

The conditional distribution for the i^{th} profile, based on a fixed value of $\boldsymbol{b}_i$, is

$$\boldsymbol{y}_i \mid \boldsymbol{b}_i \sim MN(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{R}_i) \;\; i = 1, 2, ..., m. \tag{3.9}$$

The marginal expected value of $\boldsymbol{y}_i$ is given as

$$E(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}, \tag{3.10}$$

With marginal variance of $\boldsymbol{y}_i$

$$V(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^T + \boldsymbol{R}_i = \boldsymbol{V}_i. \tag{3.11}$$

From the above assumptions, the marginal distribution of $\boldsymbol{y}_i$ is

$$\boldsymbol{y}_i \sim MN(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i). \tag{3.12}$$

The convenient way to derive an estimator of $\boldsymbol{\beta}$ is to stack the responses and the model matrix for the m individual profiles. Let $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}$, $\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_m \end{bmatrix}$, $n = \sum_{i=1}^{m} n_i$, and $\boldsymbol{Z}$ is the $n \times mq$ block diagonal matrix with Z_i along each diagonal $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & \cdots & O \\ \vdots & \ddots & \vdots \\ O & \cdots & \boldsymbol{Z}_m \end{bmatrix}$.

Model (3.8) can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}.$$

With the stack equation above, the corresponding distributions for $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$ can be written as

$$\boldsymbol{b} \sim MN(\boldsymbol{0}, \boldsymbol{G}),$$
$$\boldsymbol{\varepsilon} \sim MN(\boldsymbol{0}, \boldsymbol{R}),$$

where $\boldsymbol{R} = diag\left(\boldsymbol{R}_i\right)$, and the conditional and marginal distribution for $\boldsymbol{y}$ are

$$\boldsymbol{y} \mid \boldsymbol{b} \sim MN\left(\boldsymbol{X\beta} + \boldsymbol{ZB}, \boldsymbol{R}\right),$$

$$\boldsymbol{y} \sim MN\left(\boldsymbol{X\beta}, \boldsymbol{V}\right),$$

$$\boldsymbol{V} = \text{var}\left(\boldsymbol{y}\right) = \boldsymbol{ZGZ}^T + \boldsymbol{R}.$$

Denote $\hat{\boldsymbol{\beta}}_{LMM}$ as the estimator for the PA parameter vector for the fixed effects and denote $\hat{\boldsymbol{b}}_i$ as eblups of the random effects for the i^{th} profile. Then it can be shown that (Ruppert et al. 2003)

$$\hat{\boldsymbol{\beta}}_{LMM} = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}, \tag{3.13}$$

$$\hat{\boldsymbol{b}}_i = \boldsymbol{GZ}_i^T\boldsymbol{V}_i^{-1}\left(\boldsymbol{y} - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}\right). \tag{3.14}$$

Note, $\boldsymbol{V}$ here is usually unknown and needs to be estimated first. The most commonly used estimators for $\boldsymbol{V}$ include the maximum likelihood estimator (MLE) and the restricted maximum likelihood estimator (REMLE). Ruppert et al. (2003) mentioned that for small sample size REMLE is usually more accurate than MLE, but for large samples there is little difference between the two approaches. By substituting the estimates $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{G}}$ into (3.13) and (3.14), the parameter estimates and eblups can be obtained. Subsequently, the estimated parameter vector and eblups for the i^{th} profile are

$$\hat{\boldsymbol{\beta}}_i^p = \hat{\boldsymbol{\beta}}_{LMM} + \hat{\boldsymbol{b}}_i^*, \tag{3.15}$$

where "*p*" represents the estimated coefficients using the parametric approach, $\hat{\boldsymbol{\beta}}_{LMM} = \left(\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$ and $\hat{\boldsymbol{b}}_i^*$ is a $p\times 1$ vector containing $\hat{\boldsymbol{b}}_i$ for the columns of $\boldsymbol{Z}_i$ that are equal to the columns of X_i and zero otherwise. Consequently, $\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{b}}_i^*$ if $\boldsymbol{X}_i = \boldsymbol{Z}_i$. The estimated fits for PS_i curve and for the PA curve are expressed as

$$\hat{\boldsymbol{y}}_{PS,i}^P = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i^P = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{LMM} + \boldsymbol{Z}_i\boldsymbol{b}_i, \tag{3.16}$$

and

$$\hat{\boldsymbol{y}}_{PA}^P = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{LMM}. \tag{3.17}$$

3.4.2 Nonparametric Mixed Regression and P-spline Estimation

In many applications, the profiles cannot be parametrically modeled by either linear or nonlinear functions. In these cases, a nonparametric regression may provide a better fit to the data than either linear or nonlinear functions. The nonparametric regression method has several advantages over parametric regression in that nonparametric regression is more flexible and offers an exploratory approach to evaluate the data. More details for nonparametric regression can be obtained from Ruppert et al. (2003), and Wegener and Kauermann (2008). In fact, the mixed effects nonparametric method will be used extensively in this dissertation.

Before introducing the mixed effects nonparametric regression method, the fixed effect nonparametric regression method is presented in its general form (only one regressor **x** is used for simplicity) as

$$y_i = f\left(x_i\right) + \varepsilon_i \qquad 1 \le i \le n_i, \tag{3.18}$$

where f is some unspecified smooth function with $E(\varepsilon_i)=0$, $\text{cov}(\boldsymbol{\varepsilon})=\sigma_\varepsilon^2 I$ and $E(y \mid x)=f(x)$. There are several methods available to nonparametrically fit the model in (3.18). For example, one can use spline-based smoothers, moving average smoothers or kernel smoothers. The method we focus is the penalized spline (p-spline), a spline-based smoother, which has the attractiveness of being a relative straightforward extension of linear regression modeling (see O'Sullivan et al. (1986); Gray (1994); Eilers and Marx (1996) and Berk (2008).).The main idea of p-spline regression is to fit the function $f(x_i)$ parametrically with a sufficiently flexible spline bases. Instead of simply using parametric estimation, a penalty is imposed to the spline coefficients to achieve a smooth fit. One technical benefit of this approach is that it links this nonparametric method to the L-W model, which can be useful in many applications, see Wand (2003) and Ruppert et al. (2003) for details.

There are many spline bases available in p-spline regression including, for example, truncated polynomial bases, radial bases, and natural cubic bases. The bases used here are the truncated polynomial bases. With the truncated polynomial bases, $f(x_i)$ can be approximated by

$$f(x_i) \approx \beta_0 + \sum_{l=1}^{p} \beta_l x_i^l + \sum_{k=1}^{K} \beta_{pk} (x_i - \kappa_k)_+^p , \tag{3.19}$$

where p is the order of the polynomial and $\kappa_1, \kappa_2, ..., \kappa_K$ are the knots, K, is the total number of knots and $(x_i - \kappa_k)_+^p$ is defined as 0 for $x_i \leq \kappa_k$ and $(x_i - \kappa_k)^p$ otherwise. For example, one common application of p-spline regression is the case where $p=1$ which gives

$$f(x_i) \approx \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} \beta_{1k} (x_i - \kappa_k)_+ , \tag{3.20}$$

where $(x_i-\kappa_1)_+,(x_i-\kappa_2)_+,...,(x_i-\kappa_K)_+$ are the linear spline bases. Claeskens et al. (2009) suggested choosing K according to $K=\min(40,n/40)$. Also, one may use one of several knot selection rules suggested in Ruppert et al.(2003). To keep the approach simple, K will be determined using the above rule of thumb for both the PA profile curve and the m PS profile curves in our presentation. Once K is chosen, the knots κ_k, $k=1,2,...K$ can be selected to cover the range of $\mathbf{x}$ values using quartiles (Wand (2003)).

The ordinary least square fit for the model in (3.20) can be written as

$$\min_{\forall \boldsymbol{\beta}} \|\boldsymbol{y}-\boldsymbol{X\beta}\|^2 \tag{3.21}$$

$$\boldsymbol{X}=\begin{bmatrix} 1 & x_1 & (x_1-\kappa_1)_+ & \cdots & (x_1-\kappa_K)_+ \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n-\kappa_1)_+ & \cdots & (x_n-\kappa_K)_+ \end{bmatrix} \text{ and } \boldsymbol{\beta}^T=(\beta_0,\beta_1,\beta_{11},\cdots\beta_{1K}),$$

where β_{1k} is the spline coefficient for the k^{th} knot. As addressed previously, the p-spline regression imposes a constraint on the spline coefficients. Possible constraints on the $\beta_{11},...,\beta_{1K}$ can be one of the following

(1) $max|\beta_{pk}| < C$

(2) $\sum|\beta_{pk}| < C$

(3) $\sum {\beta_{pk}}^2 < C$

The third constrain is more commonly used for easy implementation. Using the third constrain, our minimization problem can be written as

$$\min_{\forall \boldsymbol{\beta}} \|\boldsymbol{y}-\boldsymbol{X\beta}\|^2$$

subject to (3.22)

$$\boldsymbol{\beta}^T \boldsymbol{D}\boldsymbol{\beta} < C$$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times K} \\ \mathbf{0}_{K\times 2} & I_{K\times K} \end{bmatrix}$$

Using a Lagrange multiplier argument, (3.22) is equivalent to choosing $\boldsymbol{\beta}$ to minimize

$$\left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \right\|^2 + \lambda^2 \boldsymbol{\beta}^T \boldsymbol{D}\boldsymbol{\beta}, \tag{3.23}$$

for some $\lambda \geq 0$. This gives then estimates of mean response at $\boldsymbol{X}$ as

$$\hat{\boldsymbol{f}} = \boldsymbol{X}\left(\boldsymbol{X}^T \boldsymbol{X} + \lambda^2 \boldsymbol{D} \right)^{-1} \boldsymbol{X}^T \boldsymbol{y}, \tag{3.24}$$

Usually, λ is chosen based on criteria such as minimizing either the cross validation error or the generalized cross validation error. Another approach uses the relationship of the p-spline model and the LMM to obtain the λ automatically. To see this relationship, for a simple example, model (3.20) can be rewritten as

$$f\left(x_i\right) = b_0 + b_1 x_i + \sum_{k=1}^{K} \mu_k \left(x_i - \kappa_k \right)_+ + \varepsilon_i$$

Let

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix},\ \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_K \end{bmatrix},\ \boldsymbol{C}_1 = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \text{ and } \boldsymbol{C}_1 = \begin{bmatrix} \left(x_1 - \kappa_1\right)_+ & \cdots & \left(x_1 - \kappa_K\right)_+ \\ \vdots & \ddots & \vdots \\ \left(x_n - \kappa_1\right)_+ & \cdots & \left(x_n - \kappa_K\right)_+ \end{bmatrix}$$

The penalized fitting criteria in (3.23), when divided by σ_ε^2, can be written as

$$\frac{1}{\sigma_\varepsilon^2}\|\boldsymbol{y}-C_1\boldsymbol{b}-C_2\boldsymbol{u}\|^2+\frac{\lambda^2}{\sigma_\varepsilon^2}\|\boldsymbol{u}\|^2$$

Ruppert et al. (2003) showed that this is also the objective function to obtain the ebulps for $\boldsymbol{\mu}$ as a set of random coefficients with $\text{cov}(\boldsymbol{\mu})=\boldsymbol{G}=\sigma_\mu^2\boldsymbol{I}$, where $\sigma_\mu^2=\frac{\sigma_\varepsilon^2}{\lambda^2}$ .The n estimates of mean response can be obtained by

$$\hat{\boldsymbol{f}}=\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{C}+\lambda^2\boldsymbol{D}\right)^{-1}\boldsymbol{C}^T\boldsymbol{y}, \tag{3.25}$$

where $\boldsymbol{C}=\left[\boldsymbol{C}_1,\boldsymbol{C}_2\right]=\boldsymbol{X}$, $\boldsymbol{D}=\begin{bmatrix}\boldsymbol{0}_{2\times2} & \boldsymbol{0}_{2\times K}\\ \boldsymbol{0}_{K\times K} & I_{K\times K}\end{bmatrix}$ and $\lambda^2=\frac{\sigma_\varepsilon^2}{\sigma_\mu^2}$.

The representation of the knot coefficients in p-spline regression as the eblups in the L-W model is useful because it allows smoothing to be done using LMM methodology and software. For example, instead of choosing the smoothing parameter λ by cross validation criteria for the p-spline model, LMM can be used to estimate σ_ε^2 and σ_μ^2 by MLE or REMLE. Then $\hat{\lambda}_{ML}=\sqrt{\hat{\sigma}^2_{\varepsilon\,ML}/\hat{\sigma}^2_{u\,ML}}$ or $\hat{\lambda}_{REML}=\sqrt{\hat{\sigma}^2_{\varepsilon\,REML}/\hat{\sigma}^2_{u\,REML}}$.

Recall the mixed model in (3.6) (Abdel-Salam(2009)) as

$$y_{ij}=f(x_{ij})+\xi_i(x_{ij})+\epsilon_{ij}\, i=1,2,\dots,m\;\; j=1,2,\dots,n.$$

Instead of parametric functions, $f\left(x_{ij}\right)$ and $\xi_i\left(x_{ij}\right)$can be nonparametric functions. Both $f\left(x_{ij}\right)$ and $\xi_i\left(x_{ij}\right)$ can be approximated by p-spline regression. For example, the truncated polynomial bases of order p can be used to approximate $f\left(x_{ij}\right)$ and $\xi_i\left(x_{ij}\right)$ (though other bases can be utilized as well) such as

$$f\left(x_{ij}\right) \approx \beta_0 + \sum_{l=1}^{p} \beta_l x_{ij}^{\ l} + \sum_{k=1}^{K_1} u_{pk} \left(x_{ij} - \kappa_k\right)_+^p \tag{3.26}$$

$$\xi_i\left(x_{ij}\right) \approx b_{i0} + \sum_{l=1}^{p} b_{il} x_{ij}^{\ l} + + \sum_{k=1}^{K_2} t_{ik} \left(x_{ij} - \kappa_k\right)_+^p \quad i = 1,2,3,...,m \quad j = 1,2,3,...,n \tag{3.27}$$

With the relationship between the p-spline regression approximation and the LMM, the approximation for the i^{th} profile can be described succinctly in the LMM framework as

$$\boldsymbol{y}_{PS,i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{u} + X_i\boldsymbol{b}_i + E_i\boldsymbol{t}_i + \boldsymbol{\epsilon}_i i = 1,2,\ldots,m, j = 1,2,\ldots,n, \tag{3.28}$$

where

$$X_i = \begin{bmatrix} 1 & x_{i1} & \cdots & x_{i1}^{\ p} \\ 1 & x_{i2} & \cdots & x_{i2}^{\ p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{in} & \cdots & x_{in}^{\ p} \end{bmatrix}, \quad Z_i = \begin{bmatrix} (x_{i1} - \kappa_1)_+^p & \cdots & \left(x_{i1} - \kappa_{K_1}\right)_+^p \\ \vdots & \ddots & \vdots \\ (x_{in} - \kappa_1)_+^p & \cdots & \left(x_i - n\kappa_{K_1}\right)_+^p \end{bmatrix},$$

$$E_i = \begin{bmatrix} (x_{i1} - \kappa_1)_+^p & \cdots & \left(x_{i1} - \kappa_{K_2}\right)_+^p \\ \vdots & \ddots & \vdots \\ (x_{in} - \kappa_1)_+^p & \cdots & \left(x_{in} - \kappa_{K_2}\right)_+^p \end{bmatrix},$$

$\boldsymbol{\beta}^T = \left[\beta_0, \beta_{1,}, \ldots, \beta_p\right]$, $\boldsymbol{\mu}^T = \left[\mu_0, \mu_{1,}, \ldots, \mu_p\right]$, $\boldsymbol{b}_i^T = \left[b_{i0}, b_{i1,}, \ldots, b_{ip}\right]$, $\boldsymbol{t}_i^T = \left[t_{i0}, t_{i1,}, \ldots, t_{ip}\right]$ and $\boldsymbol{\epsilon}_i$ is the error term with $\boldsymbol{\epsilon}_i \sim MN(\boldsymbol{0}, R_i)$. K_1 and K_2 represent the number of knots chosen for $f\left(x_{ij}\right)$ and $\xi_i\left(x_{ij}\right)$ respectively. It is not necessary for K_1 and K_2 to be equal.

The model in (3.28) can also be written in a stacked matrix notation as follows

$$\boldsymbol{y} = X\boldsymbol{\beta} + ZB + \boldsymbol{\epsilon}, \tag{3.29}$$

where

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix},\ Z = \begin{bmatrix} Z_1 & X_1 & 0 & \cdots & 0 & E_1 & 0 & \ldots & 0 \\ Z_1 & 0 & X_2 & \cdots & 0 & 0 & E_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_m & 0 & 0 & \cdots & X_m & 0 & 0 & \cdots & E_m \end{bmatrix}, B = \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{b} \\ \boldsymbol{t} \end{bmatrix},$$

with

$$\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{K_1} \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_m \end{bmatrix}, \boldsymbol{t} = \begin{bmatrix} \boldsymbol{t}_1 \\ \boldsymbol{t}_2 \\ \vdots \\ \boldsymbol{t}_m \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix},$$

and

$$Cov(B) = G = \begin{bmatrix} \sigma_u^2 I & O & O \\ O & blockdiag_{1 \le i \le m} \Sigma_b & O \\ O & O & \sigma_t^2 I \end{bmatrix},$$

$$\Sigma_b = \sigma_b^2 Corr(\boldsymbol{b}_i).$$

Here, σ_μ^2 controls the amount of smoothing to estimate $f\left(x_{ij}\right)$, σ_t^2 measures the between profile variation and σ_b^2 controls the amount of smoothing required to properly estimate $\xi_i\left(x_{ij}\right)$. $\boldsymbol{\beta}$ and $\boldsymbol{B}$ can be estimated by

$$\widehat{\boldsymbol{\beta}}^{p-s} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y}, \tag{3.30}$$

where the "p-s' represents the estimated coefficients using the p-spline approach.

$$\widehat{\boldsymbol{B}} = \begin{bmatrix} \widehat{\boldsymbol{u}} \\ \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{t}} \end{bmatrix} = \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}), \tag{3.31}$$

where $\boldsymbol{V} = \mathrm{var}\left(\boldsymbol{y}\right) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R}$ and $\boldsymbol{R} = diag\left(\boldsymbol{R}_i\right)$. The estimated PA curve using p-spline (denoted by "p-s" in the formulas below) regression is given by

$$\hat{\boldsymbol{y}}_{PA}^{p-s} = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{p-s} + \boldsymbol{Z}_i\hat{\boldsymbol{u}}, \tag{3.32}$$

and the estimated PS curve for the i^{th} profile is

$$\hat{\boldsymbol{y}}_{PS,i}^{p-s} = X_i\hat{\boldsymbol{\beta}}^{p-s} + Z_i\hat{\boldsymbol{u}} + X_i\hat{\boldsymbol{b}}_i + E_i\hat{\boldsymbol{t}}_i. \tag{3.33}$$

3.5 Detecting the Out-of-control Process

One objective in Phase I is to correctly distinguish between profiles belonging to the in-control process from those profiles belonging to an out-of-control process in order to obtain a properly estimated PA profile curve. In Phase I, correctly identifying the profiles belonging to the in-control and out-of-control processes is equivalent to correctly separating the profiles from the in-control and out-of-control process within the HDS.

3.5.1 Detecting the Out-of-control Process Using LMM

Recall (3.15) in the LMM, The estimated parameters and eblups for the i^{th} profile PS_i are contained in $\hat{\boldsymbol{\beta}}_i^P$ as

$$\hat{\boldsymbol{\beta}}_i^P = \hat{\boldsymbol{\beta}}_{LMM} + \hat{\boldsymbol{b}}_i^*,$$

and the estimated parameter vector for the PA is $\hat{\boldsymbol{\beta}}_{LMM}$. Jensen et al. (2008) proposed a parametric approach to determine the unusual profiles based on the distance of the estimated parameter vector from the center of the group of estimated parameter vectors. They introduced a formula for the T^2 statistic based on comparing $\hat{\boldsymbol{\beta}}_i^P$ to the sample mean of $\hat{\boldsymbol{\beta}}_i^P$, $\hat{\boldsymbol{\beta}}_{LMM}$ using the estimated variance covariance matrix, $\hat{\boldsymbol{V}}$, based on the successive difference estimator of $\boldsymbol{V}$. The T^2 statistic for the i^{th} estimated PS curve is then defined as

$$T^2_{P1,i} = \left(\hat{\boldsymbol{\beta}}^P_i - \hat{\boldsymbol{\beta}}_{LMM}\right)^T \hat{\boldsymbol{V}}^{-1} \left(\hat{\boldsymbol{\beta}}^P_i - \hat{\boldsymbol{\beta}}_{LMM}\right),$$

where (3.34)

$$\hat{\boldsymbol{V}} = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \left(\hat{\boldsymbol{\beta}}^P_{i+1} - \hat{\boldsymbol{\beta}}^P_i\right)\left(\hat{\boldsymbol{\beta}}^P_{i+1} - \hat{\boldsymbol{\beta}}^P_i\right)^T.$$

Also, they showed that the distribution of $T^2_{P1,i}$ follows asymptotically a chi-squared distribution with p degrees of freedom for large samples, where " p " the degree of freedom for the chi-squared distribution, is the number of estimated parameters. Since $\sum_{i=1}^{m} \hat{\mathbf{b}}_i = 0$, it follows that **(**Jensen et al. (2008)) (3.34) can be written equivalently as

$$T^2_{P1,i} = \hat{\boldsymbol{b}}_i^T \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{b}}_i$$

and

$$\hat{\boldsymbol{V}} = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \left(\hat{\boldsymbol{b}}_{i+1} - \hat{\boldsymbol{b}}_i\right)\left(\hat{\boldsymbol{b}}_{i+1} - \hat{\boldsymbol{b}}_i\right)^T. \qquad (3.35)$$

As another approach, Abdel-salam (2009) introduced the T^2 statistic based on the fitted values for the i^{th} estimated PS curve and estimated PA profile, $\hat{\boldsymbol{y}}^P_{PS,i}$ and $\hat{\boldsymbol{y}}_{PA}$ respectively, where the fits for both curves are obtained at the same n′ values of the regressors across all m profiles, as

$$T^2_{P2,i} = \left(\hat{\mathbf{y}}^P_{PS,i} - \hat{\mathbf{y}}_{PA}\right)^T \hat{\boldsymbol{V}}^{-1} \left(\hat{\mathbf{y}}^P_{PS,i} - \hat{\mathbf{y}}_{PA}\right), \qquad (3.36)$$

$$\hat{\boldsymbol{V}} = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \left(\hat{\mathbf{y}}^P_{PS,i+1} - \hat{\mathbf{y}}^P_{PS,i}\right)\left(\hat{\mathbf{y}}^P_{PS,i+1} - \hat{\mathbf{y}}^P_{PS,i}\right)^T.$$

3.5.2 Detecting Out-of-control Process Using the P-spline Mixed Model

When using the p-spline regression approach for the nonparametric mixed model, there are two formulas of T^2 statistics to determine the profiles from the out-of-control process. The one T^2 statistic based on the fitted PS curve for the i^{th} profile and the fitted PA curve, denoted as $T^2_{NP2,i}$

$$T^2_{NP2,i} = \left(\hat{\boldsymbol{y}}^{p-s}_{PS,i} - \hat{\boldsymbol{y}}^{p-s}_{PA}\right)^T \hat{\boldsymbol{V}}^{-1} \left(\hat{\boldsymbol{y}}^{p-s}_{PS,i} - \hat{\boldsymbol{y}}^{p-s}_{PA}\right) \; i = 1,2,\cdots,m. \tag{3.37}$$

where "NP2" denotes the second T^2 statistic based on nonparametric regression and $\hat{\mathbf{y}}^{p-s}_{PS,i}$ is the fitted PS curve using p-spline regression (denoted by "p-s") for the i^{th} profile and $\hat{\mathbf{y}}^{p-s}_{PA}$ is the fitted value for the PA profile. $\hat{V}$ is an $\mathrm{n}' \times \mathrm{n}'$ estimated variance-covariance matrix for $\hat{\mathbf{y}}^{p-s}_{PS,i}$ based on successive difference estimator.

The other T^2 statistic is based on the eblups, $\widehat{\boldsymbol{\phi}}_i$, for the random effects, where

$$\widehat{\boldsymbol{\phi}}_i = \begin{bmatrix} \widehat{\boldsymbol{b}}_i \\ \hat{\boldsymbol{t}}_i \end{bmatrix} \quad i = 1,2,\dots,m, \tag{3.38}$$

with $\widehat{\boldsymbol{b}}_i = \left[\hat{b}_{i0}, \hat{b}_{i1}\right]^T$,(using the first order polynomial with ,with p=1) and $\hat{\boldsymbol{t}}_i = \left[\hat{t}_{i1}, \hat{t}_{i2}, \dots, \hat{t}_{iK_2}\right]^T$. Define $T^2_{NP1,i}$ for the i^{th} profile as

$$T^2{}_{NP1,i} = \left(\widehat{\boldsymbol{\phi}}_i - \overline{\boldsymbol{\phi}}\right)^T \hat{V}^{-1} \widehat{\boldsymbol{\phi}}_i - \overline{\boldsymbol{\phi}}\Big),$$

$$\overline{\boldsymbol{\phi}} = \frac{\sum_{i=1}^{m} \widehat{\phi}_i}{m}, \tag{3.39}$$

$$\widehat{\boldsymbol{V}} = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \left(\widehat{\boldsymbol{\phi}}_{i+1} - \widehat{\boldsymbol{\phi}}_i\right) \left(\widehat{\boldsymbol{\phi}}_{i+1} - \widehat{\boldsymbol{\phi}}_i\right)^T$$

Unusual profiles can be determined by comparing $T^2_{NP1,i}$ and $T^2_{NP2,i}$ with a value from chi-squared distribution. The i^{th} estimated PS curve will be marked as outlying if $T^2_{NP,i} \geq \boldsymbol{\chi}^2_{(df,\alpha)}$ for $i = 1, 2, \ldots m$, where α represents the significant level and df represents the degrees of freedom which is equal to $tr\left(H^{PS}_{PS,i}\right)$ where $H^{PS}_{PS,i}$ is the smoother matrix for p-spline regression.

3.6 Chapter Summary

Profile monitoring based on the mixed model in Phase I is presented in this chapter. First, the L-W version of the LMM is used to represent the profiles parametrically. Second, the L-W model is used to represent the profiles nonparametrically using p-spline regression. To detect the presence of profiles from the out-of-control process, T^2 statistics based on the estimated parameters or eblups are calculated for the L-W model for both the parametric and nonparametric models. Finally, the profiles from the out-of-control process can be identified by using the control limits based on the approximate chi-squared distribution the T^2 statistics.

Chapter 4. Cluster-Based Profile Monitoring in Phase I

A new profile monitoring methodology will be introduced for Phase I analysis in this chapter. The proposed method, referred to as the *cluster-based profile monitoring* method, incorporates a cluster analysis phase to cluster the profiles which have the similar behavior before calculating the T^2 statistic. The proposed method will be showed to be robust to the large amount of profiles in HDS from the out-of-control process and the details of the algorithm will be presented in this chapter.

4.1 Motivation

Recall that the goal of the Phase I profile monitoring process is to distinguish between the in-control process and the presence of an out-of-control process by analysis of the HDS. The in-control limits for process monitoring in Phase II analysis then can be estimated based on the in-control profiles. The existing profile monitoring method of Jensen et al. (2008) and Abdel-Salam et al. (2013) estimate the PA profile and PS profiles based on the HDS. The T^2 statistic can be obtained by using the PS fits and the PA fit. For example, in the previous section, it was noted that the T^2 statistic for the i^{th} profile can be calculated as

$$T_i^2 = \left(\hat{\boldsymbol{y}}_{PS,i} - \hat{\boldsymbol{y}}_{PA}\right)^T \hat{\boldsymbol{V}}^{-1} \left(\hat{\boldsymbol{y}}_{PS,i} - \hat{\boldsymbol{y}}_{PA}\right) \quad i = 1, 2, ...m.$$

Usually, the profiles can be represented by either a parametric function or by a nonparametric function or the combination of the parametric and nonparametric functions. If the profiles are represented by a parametric function, the above T^2 statistic

can be modified by using the estimated parameters for each profile. For example, in the LMM, the T^2 statistic can be obtained from (3.34) in Section 3 as

$$T^2_{P1,i} = \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{LMM}\right)^T \hat{\boldsymbol{V}}^{-1} \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{LMM}\right),$$

If the profiles are not well-represented by a parametric function, a p-spline regression, a nonparametric method, can be used to represent the profiles with a nonparametric function, as illustrated in the previous section. For example, in section 3, it was shown that the T^2 statistic for a nonparametric profile fit using p-spline regression can be obtained based on the eblups.

However, one problem with the method discussed above is that the estimated PA profile is based on averaging the fits of all the profiles, including any profiles from out-of control process. Thus, the estimated PA profile will be "pulled" in the direction of the out-of-control process resulting in a biased estimate of the true PA profile. Additionally, the corresponding variance-covariance matrix will be similarly distorted. Consequently, the T^2 statistics will be misleading and the in-control limits will be unable to properly separate those profiles belonging to the in-control process from those belonging to the out-of-control process.

Our goal in this chapter is to propose a new profile monitoring method which is robust to the profiles from the out-of-control process. Recall that in the CBI regression analysis, robust and efficient coefficient estimators are obtained by combining the cluster phase and the bounded influence regression phase. The cluster phase basically obtains a main cluster that contains at least half of the data set and represents the general trend of the data. A starting value for the bounded influence regression is calculated by using the data in the main cluster and iteratively adds data to the main cluster provided that this data is "close" to the data in the main cluster. Data not added is considered as outlying data. The final CBI estimator is obtained using only the final data in the main cluster, the inlying data. The proposed method for profile monitoring also starts with the

cluster phase, but instead of clustering the data points as in CBI regression, the profiles are clustered. The general introduction of the proposed method is given in the following section where details of the algorithm will also be addressed.

4.2 Proposed Cluster-based Profile Monitoring Method

The proposed cluster-based profile monitoring method is designed to provide a robust procedure for the Phase I profile monitoring process. The main idea is to first cluster the profiles to obtain a set of initial main cluster profiles with similar behavior. A cluster-based method has been used previously in the robust regression context to cluster n independent $p \times 1$ vectors by Lawrence (2003). Jobe and Pokojovy (2009) and Fan et al (2012) also proposed using a cluster-based method for use with multivariate control charts. However, clustering in the profile monitoring context is more complex than clustering data points in that the goal now is to cluster estimated curves involving intra-profile correlated data. The first step is to fit a curve, by some appropriate method, to each of m independent $n_i \times 1$ profiles (vectors) where the data within each profile is likely to be correlated. The proposed method thus allows each estimated profile to be represented by a vector of estimated model parameters (and/or by eblups). After each profile is represented with a parameter vector, the estimated variance-covariance matrix estimator, $\hat{\boldsymbol{V}}$, can be calculated by using the estimated parameter vectors. The second step is to calculate the similarity matrix $\boldsymbol{S}$ based on the estimated parameter vectors and $\hat{\boldsymbol{V}}$. Then use an appropriate cluster method to cluster each profile based on $\boldsymbol{S}$.

To obtain a tight, compact sphere of similar parameter estimates, hierarchical clustering with a proper linkage is performed until an initial main cluster containing at least half the profiles is formed. After obtaining an initial main cluster set, denoted by C_{main}, the profiles in C_{main} can be used to obtain an initial estimate of the parameters for the PA profile. These estimated parameters can be used with the previously estimated variance-covariance matrix, $\hat{\boldsymbol{V}}$, to calculate the T^2 statistics for the profiles not

contained in C_{main}. The profiles which have in-control T^2 statistics (that is, T^2 is less than the control limit of the T^2 chart) are then added to C_{main} to obtain a new set of profiles, denoted as C_{new}. The mixed model approach is then used to update the estimated parameters for the PA profile from the profiles in C_{new}. One repeats the above procedure of updating C_{new} by adding the profiles not contained in C_{new} until either the smallest T^2 statistic for the remaining profiles outside of C_{new} is beyond the control limit or all the profiles in the HSD have been added to C_{new}. Upon completion of the algorithm, those profiles contained in C_{new} are labeled as profiles from the "in-control process" and those not included in C_{new} are labeled as profiles from an "out-of-control process". A similar iterative procedure to detect in-control and out-of-control data was also used in a multivariate control chart setting by Shiau and Sun (2009). The proposed algorithm is now outlined in detail.

Step 1: Represent each estimated profile curve by an estimated parameter vector (obtained using some appropriate method) and determine the $m \times p$ parameter matrix $\hat{\boldsymbol{B}}$. The i^{th} row of $\hat{\boldsymbol{B}}$, denoted by the $1 \times p$ vector $\hat{\boldsymbol{\beta}}_i^T$, is defined as the estimated parameter vector for the i^{th} profile. Obtain a robust estimate of the variance-covariance matrix for $\hat{\boldsymbol{B}}$ using an appropriate robust estimator. As an example, the successive difference estimator, $\hat{\boldsymbol{V}}_D$, is use throughout this paper, where

$$\hat{\boldsymbol{V}}_D = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \left(\hat{\boldsymbol{\beta}}_{i+1} - \hat{\boldsymbol{\beta}}_i\right)\left(\hat{\boldsymbol{\beta}}_{i+1} - \hat{\boldsymbol{\beta}}_i\right)^T$$

Step 2: Using $\hat{\boldsymbol{V}}_D$, obtained in step 1, compute a $m \times m$ similarity matrix $\boldsymbol{S}$, where the i, j entry is defined as

$$s_{ij} = \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j\right)^T \hat{\boldsymbol{V}}_D^{-1} \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j\right),$$

where $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\beta}}_j$ are i^{th} and j^{th} rows of $\hat{\boldsymbol{B}}$, respectively.

Step 3: Perform a hierarchical cluster analysis using an appropriate linkage function on the given similarity matrix to obtain the main clusters of $\hat{\boldsymbol{\beta}}_i$ for $i = 1, 2, ..., m$. The initial

main cluster is defined as the first cluster that contains at least half the profiles. Denote the set of indices for the profiles in the main cluster as C_{main}. Stop the cluster process as soon as at least $\lfloor m/2 \rfloor + 1$ profiles are contained in the main cluster. Since new profiles may be added to C_{main} during the iteration process, we denote by C the main cluster at each iteration step. Thus at the end of step 3, $C = C_{main}$.

Step 4: Use the mixed model approach to estimate the parameters for the PA profile using the profiles in C, denoted as $\hat{\boldsymbol{\beta}}_{PA}$. For all profiles not contained in C, compute

$$T_i^2 = \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{PA}\right)^T \hat{\boldsymbol{V}}_D^{-1} \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{PA}\right),$$

where " i " denotes the i^{th} profile not contained in C and add the profiles which have $T_i^2 < \chi^2_{[1-\alpha/m],df=p}$ to C and obtain a new index set C_{new}. Here, $\chi^2_{[1-\alpha/m],df=p}$ is the [1- α/m] quantile of a chi-squared distribution with p degrees of freedom , α is the level of the test and [α/m] is the Bonferroni adjustment for multiple comparisons.

Step 5: If the profiles in C_{new} are different from the profiles in C set $C = C_{new}$ and go back to step 4, otherwise denote the set of final profiles in C_{new} as C_{final}.

Step 6: Use the mixed model approach to estimate the PA parameters from the profiles in C_{final}, to obtain the eblups, and to recompute $\hat{\boldsymbol{V}}_D$. Denote the estimated PA parameter as $\hat{\boldsymbol{\beta}}_{CPA}$. the eblups for the i^{th} PS curve by $\hat{\boldsymbol{b}}_{C,i}$ and variance-covariance matrix $\hat{\boldsymbol{V}}_D$ by $\hat{\boldsymbol{V}}_{CD}$. Here, the "$C$" in the subscript indicates that the estimates result from the cluster-based method.

Nearly identical results were obtained in all examples and simulations using either "ward" or "complete" linkage. Other linkage functions may work equally as well. Complete linkage was used in all results presented in subsequent sections. Additionally, the successive difference variance-covariance estimator is used to obtain $\hat{\boldsymbol{V}}$ because it has been shown by Sullivan and Woodall (1996) to work well in the presence of a sustained shift. If the user suspects the presence of out-of-control profiles from other than a sustained-shift change in the process, alternative robust estimators of the variance-

covariance matrix can be used instead of $\hat{\boldsymbol{V}}_D$. Then this estimator can be used to replace $\hat{\boldsymbol{V}}_D$ in steps 1 and 4 of the clustering algorithm. Some possible robust alternative variance-covariance estimators are presented below in the remainder of this section.

In a more general context, once the estimated profiles are obtained via some appropriate regression method, the collection of profiles can be represented by m appropriate $p \times 1$ vectors. In this case, the process of determining the possible presence of profiles from the out-of-control process is equivalent to detecting multivariate outliers among these $p \times 1$ vectors. A field rich in research history is available for this topic (see, for example, Rousseeuw and Leroy (2005)). Consequently, our proposed cluster-based profile monitoring method should also be compared with selected robust multivariate outlier detection methods.

The classical multivariate outlier detection method is the squared Mahalanobis distance computed as in (3.34) but with $T(x)$ and $C(x)$ replacing $\hat{\boldsymbol{\beta}}_{\boldsymbol{LMM}}$ and $\hat{\boldsymbol{V}}$, respectively. Here $T(x)$ is the arithmetic mean vector of the m vectors and C(x) is the classical moment covariance estimator. In this case, the squared Mahalanobis distance is equal to our T^2 statistic. However, since both $T(x)$ and $C(x)$ are notoriously non-robust to multivariate outliers other more robust estimators of center and covariance are recommended. The most commonly recommended choices are based on the MVE and minimum covariance determinant (MCD) estimators of center and covariance (see Rousseeuw and Leroy (2005) for a full description of these methods.), due to their high resistance to multiple multivariate outliers. Jensen et al. (2007) studied the performance of using robust versions of (3.34) based on both the MVE and MCD estimators to detect multivariate outliers in the presence of a sustained shift. They gave conditions in terms of m, n_i, shift size, and the proportion of multivariate outliers for when the MVE should be used instead of the MCD. Both Jensen et al. (2007) and Fan et al. (2013) also comment on the computational difficulties associated with the MVE and MCD procedures. Another robust method for multivariate control chart can be found in

Chenouri et al. (2009) with a robust covariance estimator called reweighted minimum covariance determinant (RMCD) estimator.

Within the profile monitoring context, several authors, including Vargas (2003) and Williams et al. (2007b), suggested using (3.34) with $\hat{\boldsymbol{\beta}}_{LMM}$ and $\hat{\boldsymbol{V}}$ replaced by the MVE estimators to robustify the profile monitoring process. It seems appropriate then that our cluster-based method be compared not only to the non-cluster based method of Jensen et al. (2008) but to a robust version of the (3.34) based on the MVE estimator. Following the guidelines of Jensen et al. (2007), the MVE is more appropriate than the MCD estimator for our selected values of m, n_i, shift size and the proportion profiles from the out-of-control process. We will also consider as a fourth estimator a robust version of our cluster-based methods computed replacing $\boldsymbol{V}_D$ with the MVE covariance estimator.

4.3 Detailed Simple Example

To aid in understanding the proposed algorithm, consider monitoring quadratic trend profiles whose in-control profiles were randomly generated from the model

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_2 + b_{1i})x_{ij} + (\beta_3 + b_{2i})x_{ij}^{\ 2} + \varepsilon_{ij}, i = 1, 2, \ldots, m_1, j = 1, 2, \ldots, n \tag{4.1}$$

and out-of-control profiles were generated as

$$\begin{aligned} &y_{ij} = (\beta_0' + b_{0i}) + (\beta_1' + b_{1i})x_{ij} + (\beta_2' + b_{2i})x_{ij}^{\ 2} + \varepsilon_{ij}, \\ &i = m_1 + 1, \ldots, m, \ j = 1, 2, \ldots, n \end{aligned} \tag{4.2}$$

where

$$\begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{bmatrix} \sim MN\left[\mathbf{0} \ , \begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{pmatrix} \right]$$

$$\varepsilon \sim N\left[0, \quad \sigma^2 I\right] \tag{4.3}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Here, chose, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2) = (12.5, -7, 2)$ for the in-control profiles and $\boldsymbol{\beta}'^T = (\beta_0', \beta_1', \beta_2') = (21.875, -14.5, 3.5)$ for the out-of-control profiles; $m_1 = 9, m = 12$ and $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 0.5$ and $\sigma^2 = 4$. $\sigma^2 = 4.$ Thus, profiles 1 through 9 represent profiles from the in-control process and profiles 10, 11, and 12 represent profiles from the out-of-control process. To simplify the illustration we assume that the covariates for these profiles are the same and equally spaced with

$$x_{ij} = j, \quad i = 1, 2, ..., m, \quad j = 1, 2, ..., n \tag{4.4}$$

Table 4.1: Dataset for the example

y/x	1	2	3	4	5	6	7	8
y_1	9.889	5.472	0.249	-2.385	-2.734	3.439	6.59	11.622
y_2	8.516	8.17	9.546	18.988	32.265	49.488	67.813	94.112
y_3	9.675	7.392	10.86	24.331	37.663	59.059	85.993	117.184
y_4	8.052	4.346	12.24	21.874	33.12	52.536	76.48	97.855
y_5	11.507	7.141	4.607	12.356	14.189	24.539	41.019	59.639
y_6	9.539	16.168	20.864	35.604	50.185	75.128	109.144	141.969
y_7	8.388	7.373	9.634	22.176	39.292	56.104	79.946	106.386
y_8	5.789	4.167	-2.028	-3.533	-2.625	-0.032	1.299	4.46
y_9	7.138	11.551	4.26	20.074	31.81	51.178	73.111	101.308
y_{10}	9.495	9.448	8.46	18.739	37.269	58.864	89.642	127.796
y_{11}	10.843	-0.678	-1.964	-0.057	6.97	16.974	31.161	51.937
y_{12}	7.832	3.833	1.719	7.512	16.532	34.221	53.323	82.262

The exploratory analysis is done by plotting the 12 observed profiles in Figure 4.1 where the blue curves represent the profiles from the in-control process while the red curves represent the profiles from the out-of-control process.

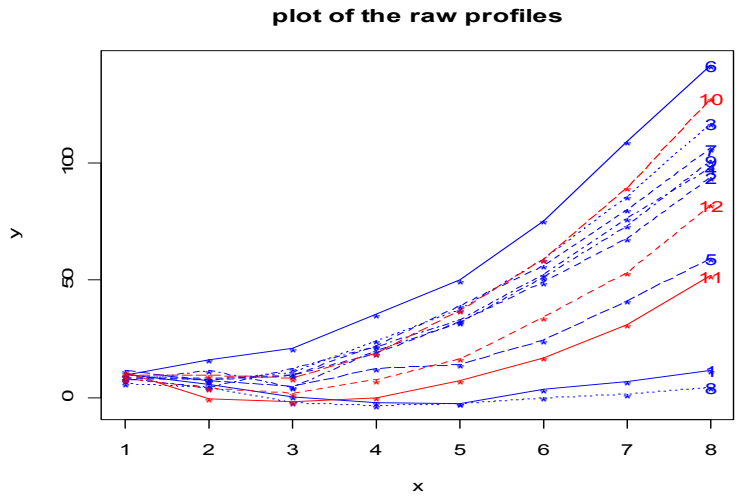

Figure 4.1: Plot of 12 observed profiles

Step 1

The Figure 4.1 shows the quadratic trend of the 12 profiles. It seems that profile 6 have larger observed values than other profiles but it is not trivial to conclude that profiles 10, 11 and 12 are "different" from the other nine profiles in the plot. The parameters for each profile are estimated individually using the fixed effects quadratic model in one regressor

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \beta_{2i} x_{ij}^{\ 2} + \varepsilon_{ij}, i = 1, 2, ..., m_1, j = 1, 2, ..., n.$$

using the method of least squares. The estimated parameters for each profile are listed in Table 4.2.

Table 4.2: 12×3 $\hat{B}$ matrix; the parameter estimates for 12 profiles

Index of profiles	$\hat{\beta}_{1i}$	$\hat{\beta}_{2i}$	$\hat{\beta}_{3i}$
1	18.393	-9.171	1.055
2	13.14	-7.072	2.149
3	15.41	-9.214	2.748
4	9.743	-5.554	2.1
5	20.558	-10.704	1.941
6	15.127	-6.44	2.791
7	11.069	-6.338	2.299
8	12.029	-6.316	0.68
9	14.907	-9.068	2.488
10	**21.645**	**-14.318**	**3.441**
11	**21.892**	**-14.832**	**2.324**
12	**20.081**	**-14.214**	**2.737**

Recall that the plot in Figure 4.1 does not show that profiles 10, 11, and 12 are apparently "different" from the other profiles, however, Table 4.2 shows that profiles 10, 11, 12 have a very "different" estimated parameter vectors. For example, all the estimated intercepts for profiles 10, 11, 12 have larger estimates compare to the other nine profiles. The corresponding successive difference estimate for the covariance matrix is calculated as

$$\hat{\boldsymbol{V}}_D = \begin{bmatrix} 12.987 & -7.291 & 0.181 \\ -7.291 & 4.677 & -0.280 \\ 0.181 & -0.279 & 0.508 \end{bmatrix}$$

Step 2

Using $\hat{V}$ in step 1 as the estimated variance-covariance matrix, obtain the similarity matrix S, presented in table 4.3.

Table 4.3: Similarity matrix using $s_{ij} = \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j\right)^T \hat{\boldsymbol{V}}_D^{-1} \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j\right)$

s_{ij}	Index of profiles i=1,2,…,12											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0	5.19	9	9.77	1.81	11.55	8.96	4.57	8.7	23.65	24.4	29.37
2	5.19	0	1.89	1.18	4.77	8.45	0.65	4.55	1.97	16.43	21.26	22.32
3	9	1.89	0	3.26	5.86	13.32	1.92	9.12	0.24	7.43	12.71	12.79
4	9.77	1.18	3.26	0	10.46	13.61	0.2	4.57	2.74	18.99	22.79	22.54
5	1.81	4.77	5.86	10.46	0	8.52	8.61	9.41	6.56	16.28	20.7	24.76
6	11.55	8.45	13.32	13.61	8.52	0	12.07	19.73	15.85	34.96	48.23	50.63
7	8.96	0.65	1.92	0.2	8.61	12.07	0	5.34	1.7	16.07	20.68	20.51
8	4.57	4.55	9.12	4.57	9.41	19.73	5.34	0	7.33	26.19	23.5	26.41
9	8.7	1.97	0.24	2.74	6.56	15.85	1.7	7.33	0	7.56	11.08	11.24
10	23.65	16.43	7.43	18.99	16.28	34.96	16.07	26.19	7.56	0	3.62	3.08
11	24.4	21.26	12.71	22.79	20.7	48.23	20.68	23.5	11.08	3.62	0	0.75
12	29.37	22.32	12.79	22.54	24.76	50.63	20.51	26.41	11.24	3.08	0.75	0

Step3

The cluster history for a complete-linkage clustering of S is displayed in Table 4.4, with Figure 4.2 providing a dendrogram representation of the clustering process. The main cluster requires at least more than half of the profiles which are at least 7 profiles in this example.

Table 4.4: Cluster history for example data

Step	1	2	3	4	5	6	7	8	9	10	11	12
1	1	2	3	4	5	6	7	8	9	10	11	12
2	1	2	3	4	5	6	4	7	8	9	10	11
3	1	2	3	4	5	6	4	7	3	8	9	10
4	1	2	3	4	5	6	4	7	3	8	9	9
5	1	2	3	2	4	5	2	6	3	7	8	8
6	1	2	3	2	1	4	2	5	3	6	7	7
7	1	2	2	2	1	3	2	4	2	5	6	6
8	1	2	2	2	1	3	2	4	2	5	5	5
9	1	2	2	2	1	3	2	2	2	4	4	4
10	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**3**	**3**	**3**
11	1	1	1	1	1	1	1	1	1	2	2	2
12	1	1	1	1	1	1	1	1	1	1	1	1

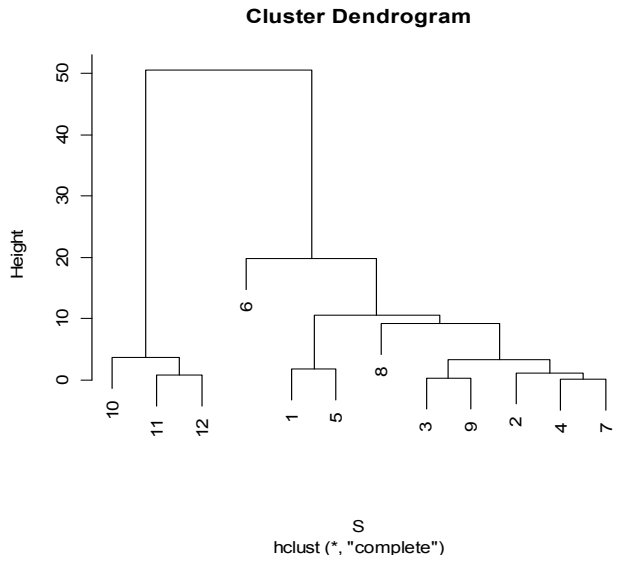

Figure 4.2: Dendrogram for clustering of example dataset.

After a total of ten steps, the initial main cluster contains profiles 1-4, and 7-9. Using the profile index to represent each profile, the initial main cluster is defined as $C_{main} = \{1, 2, 3, 4, 5, 8, 9\}$ and $C = C_{main} = \{1, 2, 3, 4, 5, 8, 9\}$. It is noted that there are six profiles are contained in the main cluster in cluster step 9 and that one more profile is added to this cluster in cluster step 10, resulting in seven profiles in the main cluster. Since this is the first cluster step in which at least more than half of the profiles are contained in the main cluster, the algorithm proceeds to step 4.

Step 4

The LMM is used to obtain the PA parameter estimate $\hat{\boldsymbol{\beta}}_{PA}$ based on the profiles in C as

$$\hat{\boldsymbol{\beta}}_{PA}{}^{T} = (14.406, \quad -7.930, \quad 1.932),$$

and the T_i^2 statistics for those profiles not contained in C are displayed below

Index of profiles not contained in C	6	10	11	12
T_i^2	10.695	14.381	17.446	19.049

The cutoff value of T_i^2 is

$$cutoff = \chi^2_{\left[1-\frac{\alpha}{m}\right],df=3} = 13.229$$

Since the 6th profile has the T^2 statistics less than the cutoff, this profile is added to C to obtain $C_{new} = \{1{:}9\}$.

Step 5

Since $C \neq C_{new}$, set $C = C_{new} = \{1{:}9\}$ and repeat step 4 using the LMM. The updated $\hat{\boldsymbol{\beta}}_{PA}$ and the T^2 statistics are obtained as

$$\hat{\boldsymbol{\beta}}_{PA}^T = (14.486, \quad -7.764, \quad 2.027)$$

Index of profiles not contained in C	10	11	12
T_i^2	15.611	19.811	21.502

Since the T^2 statistics above show that no profile can be added, the algorithm stops here with $C_{final} = \{1{:}9\}$.

Step 6

All profiles in the final set C_{final} are used with the LMM model to estimate the PA parameter vector $\hat{\boldsymbol{\beta}}_{CPA}$, eblups $\hat{\boldsymbol{b}}_{i,C}$ and variance-covariance matrix $\hat{\boldsymbol{V}}_{\boldsymbol{C}}$ as

$$\hat{\boldsymbol{\beta}}_{PA}^T = (14.486, \quad -7.764, \quad 2.027).$$

The successive difference estimate $\hat{\boldsymbol{V}}_{\boldsymbol{C},D}$ based on the ebulps is

$$\hat{\boldsymbol{V}}_{\boldsymbol{C},D} = \begin{bmatrix} 2.110 & -0.969 & -0.643 \\ -0.969 & 0.619 & 0.209 \\ -0.643 & 0.209 & 0.462 \end{bmatrix}.$$

Table 4.5: Eblups for the profiles in C_{final}

Index of profiles contained in C_{final}	$\widehat{b}_{0i}$	$\widehat{b}_{1i}$	$\widehat{b}_{2i}$
1	2.045	-0.735	-1.028
2	-0.524	0.271	0.164
3	-0.299	-0.354	0.586
4	-1.502	0.686	0.222
5	1.927	-0.933	-0.287
6	-1.27	0.499	0.358
7	0.474	-0.072	-1.19
8	-0.205	-0.44	0.347
9	-0.645	1.078	0.829

The example shows that the algorithm correctly identifies the three profiles from the out-of-control process. In the cluster phase, the algorithm gives the initial main cluster of profiles as $C_{main} = \{1{:}5,7{:}9\}$, and two corresponding minor clusters C_1 and C_2, with $C_1 = \{6\}$ and $C_2 = \{10,11,12\}$. In the profile clustering process, the profile in the minor cluster C_1 is added to the initial main cluster while the profiles in C_2 are not added. This, of course, is the desired result. After correctly identifying the profiles from the out-of-control process, the final PA profile and variance-covariance matrix were estimated by using the in-control profiles in C_{final}. The cluster phase shows that the 6th, 10th, 11th and 12th profiles in the two minor clusters do not behavior as similarly as other eight profiles in the initial main cluster.

The cluster-based method, using the T^2 statistics in terms of the estimated PA profile from all eight profiles that form the in-control process, correctly identified the 6th profile as a normal profile. The non-cluster-based method, on the other hand, using the T^2 statistics in terms of the estimated PA profile from all profiles in the HDS,

misclassified the 6^{th} profile as from the out-of-control process profile and the 10^{th}, 11^{th}, 12^{th} profiles as from the in-control process.

The MVE estimator was applied in this simulated example, replacing the successive difference estimator, for both the cluster-based method and the non-cluster based method, Using MVE estimator; the non-cluster based method classified the 6th, 11^{th} and 12^{th} profiles as from the out-of-control process. The cluster based method using the MVE in step 1 of the algorithm, on the other hand, misclassified the 6^{th} profile as from the out-of-control process. Neither method, when using the MVE in place of the successive difference estimator, correctly identified the in-control and out-of-control profiles. The performance of the MVE as a replacement for the successive difference estimator is further evaluated in the following automobile engine application in Section 4.4 and Monte Carlo study in Section 4.5.

4.4 Automobile Engine Application

In the automobile engine example there are 20 engines in the HDS and it is desired to study the relationship between engine speed (measured in revolutions per minute (RPM)) and engine torque. For each engine, the speed values are set equal to 1500, 2000, 2500, 2660, 2800, 2940, 3500, 4000, 4500, 5000, 5225, 5500, 5775, and 6000 RPM and the engine's corresponding torque values were measured. The profile for each engine is the relationship between torque produced by the engine and engine's speed in RPM. An engine with mechanical issues or other issues will yield a profile that is different from the other good engines. The raw data set (see Table 4.6), were individual data points for each engine are connected by straight-line segments, is shown in Figure 4.3. This data set has been analyzed using profile monitoring methods by Amiri et al. (2010) and Abdel-Salam et al. (2013).

Table 4.6: The Automotive Industry Data, Torque (T) vs. RPM

RPM	T_E1	T_E2	T_E3	T_E4	T_E5	T_E6	T_E7	T_E8	T_E9	T_E10
1500	98.53	96.35	96.7	96.75	97.61	100.06	94.55	96.48	96.83	100.07
2000	102.65	100.74	100.05	100.87	102.46	103.6	103.22	100.87	103.78	103.91
2500	113.82	110.67	111.17	110.14	112.18	112.74	112.99	110.81	114.3	112.52
2660	115.26	113.06	111.51	110.48	112.99	113.56	114.18	113.2	114.62	113.25
2800	116.24	114.58	112.01	110.94	114.54	112.85	116.48	114.73	117.19	114.1
2940	117.06	114.98	111.23	111.17	115	114.49	115.33	115.13	116.61	114.1
3500	109.89	108.55	105.64	105.78	108.99	108.95	109.59	108.69	110.43	109.21
4000	109.65	107.41	106.02	103.37	107.95	108.24	108.47	107.55	109.61	108.34
4500	105.72	103.9	103.11	102.23	103.65	105.56	105.27	104.03	106.32	104.87
5000	99.74	97.99	97.4	96.06	96.94	98.92	97.9	98.12	99.44	98.35
5225	95.97	94.27	93.88	92.39	92.78	95.41	94.67	94.39	95.62	94.76
5500	89.47	88.45	88.17	86.54	86.41	89.19	88.23	88.56	89.46	88.93
5775	81.96	81.44	81.18	79.31	78.6	81.85	80.86	81.54	82	82.19
6000	74.9	75	75.03	73.13	71.97	75.09	73.93	75.09	75.83	75.8
RPM	**T_E11**	**T_E12**	**T_E13**	**T_E14**	**T_E15**	**T_E16**	**T_E17**	**T_E18**	**T_E19**	**T_E20**
1500	97.98	97.29	93.13	93.11	95.38	98.28	96.79	96.45	91.53	98.37
2000	104.98	105.86	101.02	103.43	101.25	101.29	103.64	104.52	100.72	102.4
2500	114.9	115.25	111.25	112.02	111.53	112.2	112.73	113.78	110.71	112.67
2660	116.06	117.83	111.83	113.2	112.11	112.57	113.92	114.59	111.72	113.76
2800	116.65	117.97	113.27	113.77	112.6	113.06	113.35	115.4	112.29	115.41
2940	116.18	117.77	113.04	113.77	111.76	112.37	112.78	115.86	111.61	113.01
3500	109.65	111.31	105.6	109.15	108.12	107.03	108.2	110.78	105.21	110.08
4000	109.06	110.97	106.15	108.05	106.62	106.37	107.06	110.21	106.22	109.51
4500	105.01	107.37	104.12	103.46	102.92	104.1	105.27	106.75	101.73	106.09
5000	97.43	100.53	97.45	98.26	96.35	98.01	98.47	99.94	96.59	99.84
5225	94.04	97.17	94.68	94.26	93.14	94.21	95.67	96.94	93.78	96.46
5500	87.51	90.47	88.59	89.09	86.75	87.53	89.41	90.24	87.29	90.16
5775	79.36	83.51	81.08	81.06	80.27	80.08	82.57	82.65	78.97	82.74
6000	72.34	76.34	75.77	74.14	73.47	73.9	76.31	76.76	72.8	75.82

Figure 4.3 shows that it's reasonable to assume the quadratic relationship between the RPM and torque for each engine. The quadratic mixed model is applied in the cluster-based algorithm. For the quadratic mixed model, we assume that for the i^{th} engine, the torque produced by the j^{th} RPM is

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) x_{ij} + (\beta_2 + b_{2i}) x_{ij}^{\ 2} + \varepsilon_{ij}$$
$$i = 1, 2, ..., 20. \quad j = 1, 2, ..., 14.$$

Here the PA parameter vector is $\boldsymbol{\beta}_{PA} = (\beta_0, \quad \beta_1, \quad \beta_2)$ and the PS estimate for i^{th} engine is $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_0 + \hat{b}_{0i}, \quad \hat{\beta}_1 + \hat{b}_{1i}, \quad \hat{\beta}_2 + +\hat{b}_{2i})$. The first step of the proposed method is to represent each profile by using its estimated parameters. In our example the fixed quadratic model is used to obtain the estimated parameters for each profile and their estimates are listed in Table 4.7.

Table 4.7: The parameter estimates for 20 engines

Index of Engines	$\widehat{\beta}_{0i}$	$\widehat{\beta}_{1i}$	$\widehat{\beta}_{2i}$
1	59.35763	0.034016	-5.22E-06
2	58.42011	0.033058	-5.04E-06
3	62.73607	0.029578	-4.56E-06
4	64.27488	0.028636	-4.50E-06
5	59.11457	0.033636	-5.24E-06
6	65.22353	0.029948	-4.68E-06
7	54.68075	0.036109	-5.48E-06
8	58.49737	0.033101	-5.05E-06
9	57.80251	0.034873	-5.32E-06
10	66.45989	0.029254	-4.60E-06
11	60.04213	0.034305	-5.36E-06
12	58.36206	0.035396	-5.39E-06
13	57.00076	0.032874	-4.96E-06
14	54.81515	0.035156	-5.31E-06
15	58.9371	0.032373	-4.98E-06
16	63.04532	0.030415	-4.74E-06
17	62.93319	0.030579	-4.71E-06
18	57.70932	0.034596	-5.23E-06
19	53.82698	0.034500	-5.21E-06
20	60.25040	0.032634	-4.97E-06

The corresponding successive difference estimate for the covariance matrix is calculated as

$$\hat{\boldsymbol{V}}_{\mathrm{D}} = \begin{bmatrix} 1.225 & -7.319\mathrm{e}^{-3} & 9.834\mathrm{e}^{-7} \\ -7.319\mathrm{e}^{-3} & 5.361\mathrm{e}^{-6} & -7.398\mathrm{e}^{-10} \\ 9.834\mathrm{e}^{-7} & -7.398\mathrm{e}^{-10} & 1.035\mathrm{e}^{-13} \end{bmatrix}$$

Using the fixed estimates in Table 4.7 and their corresponding covariance estimate, we obtain the similarity matrix which is then used to cluster the engines.. The cluster history is listed in Table 4.8. One can see that the initial main cluster set contains 6 profiles in step 17 and that 5 more profiles are added to this initial main cluster set in cluster step 18, resulting in 11 profiles in the main cluster. Since this is the first step that the main cluster set contains greater than half of the profiles, the cluster step of the algorithm stops here. The cluster history (Table 4.8) shows that the proposed algorithm ended up with 11 engines in the initial main cluster set, consisting of engines 1,2,7, 8, 9, 12, 13, 14, 18, 19, and 20. The corresponding estimated PA parameter $\hat{\boldsymbol{\beta}}_{\text{PA}}$, which is obtained by fitting the quadratic mixed model to the data of these 11 engines, is

$$\hat{\boldsymbol{\beta}}_{\text{PA}} = \left(57.338, \quad 0.0342, \quad -5.199\text{e}^{-06}\right)$$

Using this estimated $\hat{\boldsymbol{\beta}}_{\text{PA}}$, the T^2 statistics for the engines not included in the initial main cluster set are calculated and listed below.

Index of Engine	3	4	5	6	10	11	15	16	17
T_i^2	2.4499	6.7032	7.1097	3.5364	5.2611	**12.2062**	1.3232	2.3276	1.2903

The cutoff value for the T^2 statistic here is $\chi^2_{1-\frac{\alpha}{m},q} = 11.93$, where $\alpha = 0.05$, m=20 and, q, the degree of freedom, equals 2 because the eblups for the quadratic terms are equivalent to 0 in this example. According to the observed T^2 statistics and the cutoff value, all engines in the minor set will be added to the initial main cluster except the 11$^{\text{th}}$ engine and the updated estimated PA parameter vector is

$$\hat{\boldsymbol{\beta}}_{\text{PA}} = \left(59.655, \quad 0.0327, \quad -5.010\text{e}^{-06}\right)$$

The T^2 statistic for the 11$^{\text{th}}$ engine was updated by using this updated estimated PA parameters but it is still greater than the cutoff value so we fail to add the 11$^{\text{th}}$ engine to the main cluster set and conclude that this engine probably has some mechanical issues or other issues. This agrees with the results found using a nonparametric mixed model profile method used by Abdel-Salam et al. (2013).

Table 4.8: Cluster history for 20 engines

Step	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	1	2	3	4	5	6	7	2	8	9	10	11	12	13	14	15	16	17	18	19
3	1	2	3	4	5	6	7	2	8	6	9	10	11	12	13	14	15	16	17	18
4	1	2	3	4	5	6	7	2	1	6	8	9	10	11	12	13	14	15	16	17
5	1	2	3	4	5	6	7	2	1	6	8	9	10	11	12	13	14	15	16	15
6	1	2	3	4	5	6	7	2	1	6	8	9	10	2	11	12	13	14	15	14
7	1	2	3	4	5	6	7	2	1	6	8	1	9	2	10	11	12	13	14	13
8	1	2	3	4	5	6	7	2	1	6	8	1	9	2	10	11	3	12	13	12
9	1	2	3	4	5	6	7	2	1	6	8	1	9	2	10	4	3	11	12	11
10	1	2	3	4	5	6	7	2	1	6	5	1	8	2	9	4	3	10	11	10
11	1	2	3	4	5	6	2	2	1	6	5	1	7	2	8	4	3	9	10	9
12	1	2	3	4	5	6	2	2	1	6	5	1	7	2	8	4	3	9	7	9
13	1	2	3	4	5	6	2	2	1	6	5	1	7	2	8	4	3	1	7	1
14	1	2	3	4	5	6	2	2	1	6	5	1	7	2	4	4	3	1	7	1
15	1	2	3	4	5	6	2	2	1	6	5	1	2	2	4	4	3	1	2	1
16	1	2	3	4	5	3	2	2	1	3	5	1	2	2	4	4	3	1	2	1
17	1	2	3	3	4	3	2	2	1	3	4	1	2	2	3	3	3	1	2	1
18	**1**	**1**	**2**	**2**	**3**	**2**	**1**	**1**	**1**	**2**	**3**	**1**	**1**	**1**	**2**	**2**	**2**	**1**	**1**	**1**
19	1	1	2	2	2	2	1	1	1	2	2	1	1	1	2	2	2	1	1	1
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

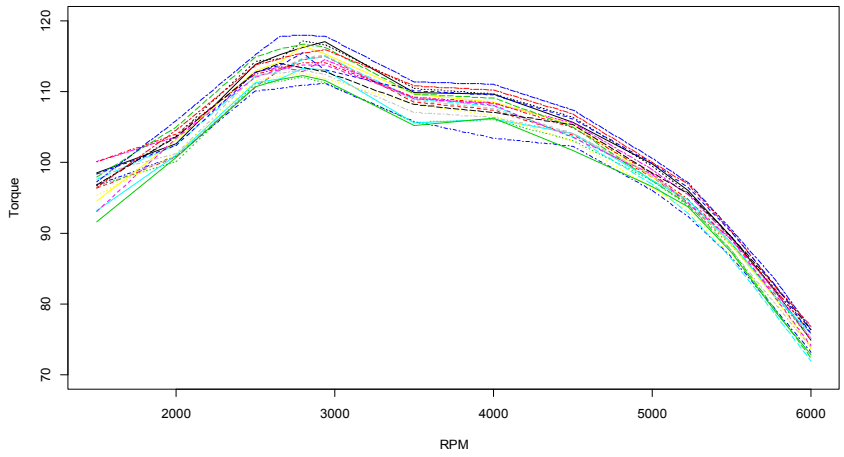

Figure 4.3: The raw data set for 20 automobile engines

The cluster dendrogram in Figure 4.4 shows that engine 5 and engine 11 are clustered in the same minor set. After the sequentially addition of the remaining engines to the initial main cluster set, the cluster-based method identified engine 11 as from the out-of-control

process and engine 5 as from the in-control process. The non-cluster-based method did not detect any engine from an out-of-control process. .

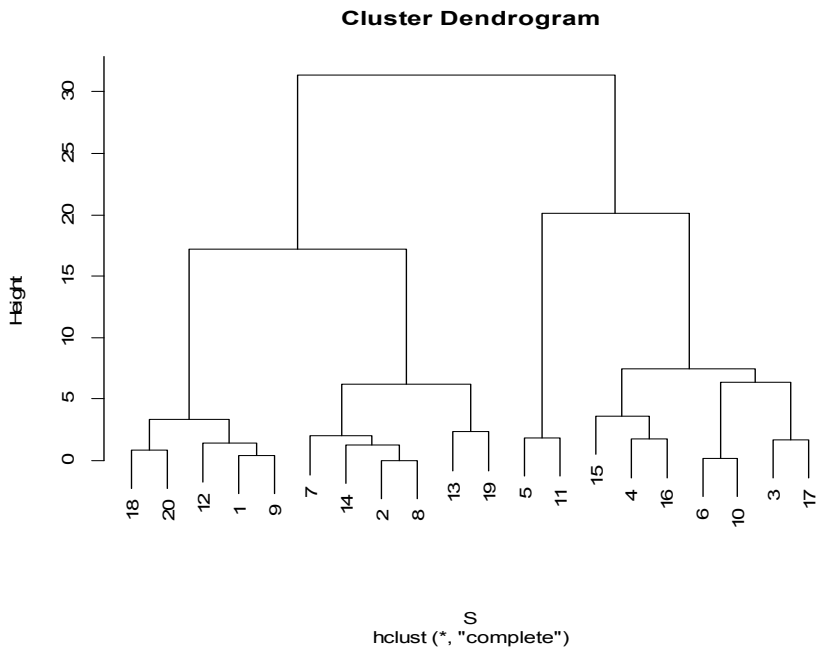

Figure 4.4: Dendrogram for clustering of 20 engines

4.5 A Monte Carlo Study

In the previous chapter, we introduced the non-cluster-based method proposed by Jensen, et al. (2008) for Phase I profile monitoring in LMM. A Monte-Carlo study is performed in order to understand and compare the cluster-based method proposed in this chapter to the non-cluster-based method.

Recall that the main purpose of Phase I profile monitoring is to correctly separate the in-control profiles from the out-of-control profiles. Both the cluster-based method and the non-cluster-based method identify those profiles in-control and those profiles that are not in-control. Various statistics can be computed to compare these methods.

One performance characteristic of Phase I analysis is the probability of signal (POS). The POS only represents the ability of the method to determine the presence of the profiles from the out-of-control process in the HDS. The POS does not give any information about whether the classification of profiles into the two categories of in-control and out-of-control is correctly specified. Each method's ability to make both correct classifications and incorrect classifications can be evaluated by computing the following performance characteristics: fraction correctly classified (FCC), sensitivity, specificity, false positive rate (FPR) and false negative rate (FNR). Fraker et al. (2008) pointed out that similar metrics are used in biosurveillance for applications in which outbreak time periods are to be distinguished from non-outbreak time periods. The definitions of these terms will be given below.

In each HDS, there are two sets of profiles. One set is from the in-control process, and the other is from the out-of-control process. After completing the Phase I analysis, the following classification table (Table 4.9) can be constructed.

Table 4.9: Classification table for Phase I analysis

Classified set / Actual set	In-control process	Out-of-control process
In-control process	A	B
Out-of-control process	C	D

In Table 4.9, "A" represents the number of profiles from the in-control process that are correctly identified as from the in-control process and "D" represents the number of profiles from the out-of-control process that are correctly identified as from the out-of-control process, respectively, after the Phase I analysis. "B" represents the number of profiles which are from the in-control process but mistakenly classified as from the out-of-control process while "C" represents the number of profiles which are from the out-of-control process but classified as from the in-control process. With this table, the FCC can be defined as

$$FCC = \frac{A + D}{A + B + C + D} \tag{4.5}$$

The sensitivity measures the ability of the classification method to identify those profiles from the in-control process correctly as profiles from the in-control process and it can be calculated as

$$Sensitivity = \frac{A}{A + B} \tag{4.6}$$

The specificity, on the other hand, represents the ability to identify those profiles from the out-of-control process correctly as profiles from the out-of-control process and it can be obtained as

$$Specificity = \frac{D}{C + D} \tag{4.7}$$

FPR is the fraction of those profiles classified as from the in-control process that are actually from the out-of-control process. FNR is the fraction of those profiles classified as from the out-of-control process that are actually from the in-control process. FPR and FNR are computed as

$$FPR = \frac{C}{A + C} \tag{4.8}$$

and

$$FNR = \frac{B}{B + D} \tag{4.9}$$

It is easy to show that all these metrics are bounded by 0 and 1, and that a method will perform well in Phase I analysis by achieving large values for FCC, sensitivity and specificity and small values for FPR and FNR.

A Monte-Carlo study is used to compare the non-cluster-based method and the cluster-based method using the performance measures of POS, FCC, sensitivity and specificity, FPR, and FNR discussed above. This Monte-Carlo study assumes the in-control profiles are randomly generated from the model

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \beta_{2i} x_{ij}^{\ 2} + \varepsilon_{ij}, i = 1, 2, \ldots, m_1, \; j = 1, 2, \ldots, n. \tag{4.10}$$

where

$$\beta_{0i} = \beta_2 \bar{x}^2 + b_{0i},$$

$$\beta_{1i} = \beta_1 - 2\beta_2 \bar{x} + b_{1i},$$

$$\beta_{2i} = \beta_2 + b_{2i}.$$

Here, $\boldsymbol{\beta}^{\mathrm{T}} = \left(\beta_1 \; \beta_2 \; \beta_3\right)$ represents the fixed parameters and $\boldsymbol{b}_i^T = \left(b_{1i} \; b_{2i} \; b_{3i}\right)$ represents the random effects. Note, the corresponding PA parameter vector can be written as $\boldsymbol{\beta}_{PA}^T = \left(\beta_2 \bar{x}^2, \;\; \beta_1 - 2\beta_2 \bar{x}, \;\; \beta_2\right)$, where β_1 and β_2 are fixed parameters and $\bar{x} = \dfrac{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}}{mn}$. Consequently, the PA profile can be written as

$$y_{PA,j} = \beta_2 \bar{x}^2 + (\beta_1 - 2\beta_2 \bar{x}) x_j + \beta_2 x_j^{\ 2} \; j = 1, 2, \ldots, n. \tag{4.11}$$

It is easy to show that (4.11) can be simplified as

$$y_{PA,j} = \beta_1 x_j + \beta_2 \left(x_j - \bar{x}\right)^2, \quad j = 1, 2, \ldots, n. \tag{4.12}$$

The out-of-control profiles are also generated from equation (4.10), but with

$$\beta_{0i} = (\beta_2 + shift)\bar{x}^2 + b_{0i},$$
$$\beta_{1i} = \beta_1 - 2(\beta_2 + shift)\bar{x} + b_{1i},$$
$$\beta_{2i} = (\beta_2 + shift) + b_{2i},$$

and its corresponding PA profile is

$$y_{PA,j} = (\beta_2 + shift)\bar{x}^2 + \left[\beta_1 - 2(\beta_2 + shift)\bar{x}\right]x_j + \left[\beta_2 + shift\right]x_j{}^2,$$
$$j = 1, 2, ..., n \tag{4.13}$$

Also, (4.13) can be simplified as

$$y_{PA,j} = \beta_1 x_j + (\beta_2 + shift)\left(x_j - \bar{x}\right)^2, \quad j = 1, 2, ..., n. \tag{4.14}$$

Note, with the simplification forms of PA profiles (4.12) and (4.14), the difference between the in-control PA profile and the out-of-control PA profile is based on the value of the shift. For example, when the shift =0, (4.12) and (4.14) are equivalent, which means all profiles are from the stable process. When the shift does not equal 0, the PA profiles from (4.12) and (4.14) are different and imply that the stable process has changed so not all the profiles are from the stable process. The performance of the two methods is evaluated based on the different values of the shift. In above equations, it is assumed that

$$\begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{bmatrix} \sim MN\left[\mathbf{0}, \begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{pmatrix}\right],$$

$$\varepsilon \sim N\left[\mathbf{0}, \quad \sigma^2 I\right],$$

and (4.15)

$$x_{ij} = j, \; i = 1,2,...,m, \; j = 1,2,...,n.$$

Here, $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 0.5, \quad \sigma^2 = 1$ and $\beta_1 = 3, \beta_2 = 2$ .It is also assumed that $m_1 = 20, \quad m = 30$ and $n = 10$. Thus, there are 20 profiles from the in-control process and 10 from the out-of-control process. The PA parameter vector for the in-control process is set at $\boldsymbol{\beta}_{PA}{}^T = \left(\beta_2 \bar{x}^2, \quad \beta_1 - 2\beta_2 \bar{x}, \quad \beta_2\right) = \left(60.5, \quad -19, \quad 2\right)$.

Before comparing the POS of the non-cluster-based method and the cluster-based method, both methods need to be calibrated to insure they have the same the same nominal POS, α_0 , when the shift=0. To achieve this, the empirical nominal POS value, α_0 , is calculated for the cluster-based method by setting shift=0 and repeating the cluster-based method using (4.3) MC times, where MC represents the number of Monte Carlo repetitions. Here, MC=10,000. The total number Monte Carlo replications with at least one declared profile from the out-of-control process (that is, at least one "signal") is determined and denoted by s_{total} . The empirical nominal value α_0 is then calculated as

$$\alpha_0 = \frac{s_{total}}{MC} \tag{4.16}$$

The next step in the calibration is to find the critical value for the non-cluster-based method so that the non-cluster-based method will have the same α_0 as the cluster-based method when the shift=0. To do this, the maximum T_i^2 statistic is calculated among the m T_i^2 for each Monte Carlo replication. That is, compute

$$T\_Stat_k = \max(T_i^2), \quad k = 1,2,...,MC \tag{4.17}$$

with shift=0 and obtain the empirical critical value for the non-cluster-based method as the $(1-\alpha_0)^{th}$ quantile of $\{T\_Stat_k,\ k=1,2,...,MC\}$.

In this Monte Carlo study, the cluster-based method has empirical $\alpha_0 = 0.0454$ and the corresponding critical value for the non-cluster-based method is 15.2497. Note that the empirical critical value for the non-cluster-based method is consistent with m=30, $\chi^2_{1-\frac{0.0454}{m},df=3} = 15.3880$ suggested for use by the non-cluster-based method. With both methods properly calibrated, the Monte Carlo study continues with MC=5,000 and with shift = (0.05, 0.075, 0.1, 0.125, 0.175, 0.2, 0.225, 0.25, 0.275, 0.3). For each value of the shift factor, the performance measures FCC, sensitivity, specificity, FNR, FPR and POS are averaged over the MC replications. The results are presented in Table 4.10.

Table 4.10: Average of performances based on a Monte Carlo study
(Within each cell the metrics are listed in the following order: cluster and non-cluster results with successive difference estimator and cluster and non-cluster results with the MVE estimator. The bold cells represent the better value)

Shift	FCC	Sensitivity	Specificity	FPR	FNR	POS
0.05	**0.6674**	**0.9981**	**0.0059**	**0.3324**	0.3922	0.0864
	0.6670	0.9978	0.0055	0.3326	0.4429	**0.0904**
	0.6663	0.9982	0.0024	0.9976	**0.0018**	0.0480
	0.6663	0.9979	0.003	0.997	0.0021	0.0476
0.075	**0.6704**	**0.9978**	**0.0156**	**0.3303**	**0.2173**	0.1578
	0.6693	0.9974	0.0132	0.3310	0.2814	**0.1594**
	0.6663	0.9983	0.0023	0.9977	0.0017	0.0468
	0.6662	0.9979	0.0029	0.9971	0.0021	0.0454
0.1	**0.6782**	**0.9978**	**0.0391**	**0.3250**	0.1016	**0.2876**
	0.6731	0.9955	0.0282	0.3280	0.2409	0.2812
	0.6663	0.998	0.0028	0.9972	**0.0020**	0.0422
	0.6661	0.9976	0.0032	0.9968	0.0024	0.0486
0.125	**0.6948**	**0.9983**	**0.0879**	**0.3136**	**0.0381**	**0.4478**
	0.6805	0.9944	0.0528	0.3226	0.1749	0.4314
	0.6667	0.9983	0.0035	0.9965	0.0017	0.0497
	0.6665	0.998	0.0036	0.9964	0.002	0.0492
0.15	**0.7268**	**0.9986**	**0.1832**	**0.2903**	0.0154	**0.6396**

	0.6913	0.9920	0.0899	0.3145	0.1518	0.5854
	0.6669	0.9983	0.0042	0.9958	**0.0017**	0.0498
	0.6667	0.998	0.004	0.996	0.0020	0.0498
0.175	**0.7697**	**0.9992**	**0.3106**	**0.2565**	**0.0050**	**0.7812**
	0.7060	0.9902	0.1378	0.3033	0.1249	0.7236
	0.6680	0.9985	0.0071	0.9929	0.0015	0.0508
	0.6673	0.9980	0.0061	0.9939	0.0020	0.0586
0.2	**0.8234**	**0.9993**	**0.4716**	**0.2091**	0.0030	**0.8790**
	0.7227	0.9871	0.1940	0.2899	0.1176	0.8230
	0.6702	0.9987	0.0132	0.9868	**0.0013**	0.0756
	0.6686	0.9981	0.0096	0.9904	0.0019	0.0752
0.225	**0.8766**	**0.9995**	**0.6309**	**0.1559**	**0.0016**	**0.9438**
	0.7432	0.9854	0.2588	0.2733	0.1012	0.8968
	0.6760	0.9988	0.0304	0.9696	0.0012	0.1026
	0.6711	0.998	0.0172	0.9828	0.0020	0.1106
0.25	**0.9219**	**0.9994**	**0.7670**	**0.1044**	**0.0016**	**0.9750**
	0.7627	0.9821	0.3241	0.2560	0.0996	0.9336
	0.6873	0.999	0.064	0.936	0.001	0.1688
	0.6761	0.998	0.0322	0.9678	0.002	0.1612
0.275	**0.9548**	**0.9996**	**0.8654**	**0.0631**	**0.0010**	**0.9896**
	0.7855	0.9806	0.3953	0.2357	0.0896	0.9698
	0.7068	0.9991	0.1222	0.8778	0.0009	0.2640
	0.6844	0.9979	0.0574	0.9426	0.0021	0.2606
0.3	**0.9749**	**0.9995**	**0.9256**	**0.0359**	0.0011	**0.9956**
	0.8052	0.9775	0.4604	0.2163	0.0890	0.9806
	0.7342	0.999	0.2045	0.7955	**0.0010**	0.3560
	0.6955	0.9974	0.0918	0.9082	0.0026	0.3550

Table 4.10 shows that when the shift is very small (shift less than or equal to 0.075), the non-cluster-based method has a slightly larger POS than the cluster-based method, but the cluster-based method has superior performance based on the other criteria. For example, when the shift is 0.05, the cluster-based method has FNR=0.3922 while the non-cluster-based method has FNR= 0.4429. Also, the cluster-based method has larger values for FCC, specificity and sensitivity with smaller FPR when the shift is 0.05. When the shift is greater than 0.075, the cluster-based method gives uniformly superior results compared to the non-cluster-based method based on all performance criteria. For example, when the shift is equal to 0.2, the cluster-based method has the FCC and FNR equal 0.8234 and 0.0030, respectively, while the non-cluster-based method has the FCC and FNR are equal to 0.7277 and 0.1176, respectively. Also, the

POS of the cluster-based method is 0.8790 while the POS for the non-cluster-based method is 0.8230. Clearly the cluster-based method is superior to the non-cluster-based method when one third of the profiles from the out-of-control process are due to a relatively large shift in the process.

When the MVE is used the cluster-based method is still superior to the non-cluster-based method although the advantage is not nearly as great as when the successive difference estimator is used. It is clearly seen that the results for the cluster-based method using the successive difference estimator are superior to those for this method when using the MVE. In fact, it is also illustrated that the results when using the MVE are generally very poor, disappointedly so, when compared to those using the successive difference estimator.

The average estimated PA parameters are also calculated for each shift factor. Table 4.11 lists the result for both cluster-based method and non-cluster-based method. Table 4.11 shows that both estimators have bias in parameter estimation compared to the true in-control PA parameters $\boldsymbol{\beta}_{PA}^{T} = (60.5, \quad -19, \quad 2)$ when one third of the profiles are from the out-of-control process. However, the estimated PA parameters from the cluster-based method have smaller bias than those from the non-cluster-based method. When the shift is small, both methods provide estimators with small bias. However, for the non-cluster-based method, the bias increases monotonically as the shift increases. For example, when the shift is 0.05, the non-cluster-based method has estimated PA parameters of $\hat{\boldsymbol{\beta}}_{PA}^{T} = (61.0026, \quad -19.1802, \quad 2.0190)$, while when the shift equals 0.3, the non-cluster-based method has estimated PA parameters of $\hat{\boldsymbol{\beta}}_{PA}^{T} = (63.5235, \quad -20.0981, \quad 2.1023)$. The bias of the estimate from the cluster-based method, on the other hand, is increasing as the shift increases and then decreases when the shift is larger than 0.15. For example, in Table 6, the cluster-based method provides the estimate with the smallest bias when the shift equals 0.3 and the second smallest bias when the shift is equal to 0.05. Table 4.11 also shows that the parameter estimates resulting from the cluster-based method

have smaller standard errors than the parameter estimated based on the non-cluster-based method.

The results in Table 4.10 illustrate the value of a procedure that can correctly distinguish those profiles from the in-control-process from those from the out-of-control process. Table 4.11 illustrates that when profiles from the out-of-control process are eliminated from the HDS, improved estimates of the PA coefficients can be obtained.

Table 4.11: Average of PA parameter estimates based on a Monte Carlo study
(Within each cell the metrics are listed in the following order: cluster and non-cluster results with successive difference estimator and cluster and non-cluster results with the MVE estimator. The bold cells represent estimates closer to the true parameter values of $\boldsymbol{\beta}_{PA}^{T} = (60.5, \quad -19, \quad 2)$ *and smaller standard error)*

Shift	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$se(\hat{\beta}_0)$	$se(\hat{\beta}_1)$	$se(\hat{\beta}_2)$
0.05	**60.9942**	**-19.1802**	**2.0149**	**0.0088**	**0.0090**	**0.0094**
	61.0026	-19.1814	2.0190	0.0098	0.0097	0.0100
	61.0059	-19.1858	2.0167	0.0093	0.0090	0.0094
	61.0068	-19.1860	2.0167	0.0109	0.0098	0.0100
0.075	**61.2410**	**-19.2674**	**2.0227**	**0.0103**	**0.0093**	**0.0096**
	61.2574	-19.2736	2.0234	0.0117	0.0099	0.0100
	61.2581	-19.2777	2.0249	0.0114	0.0092	0.0092
	61.2589	-19.2777	2.0250	0.0145	0.0103	0.0100
0.1	**61.4596**	**-19.3482**	**2.0308**	**0.0018**	**0.0096**	**0.0011**
	61.5068	-19.3648	2.0357	0.0136	0.0102	0.0100
	61.5095	-19.3691	2.0333	0.0137	0.0096	0.0011
	61.511	-19.3694	2.0333	0.019	0.0112	0.01
0.125	**61.6299**	**-19.4085**	**2.036**	**0.0129**	**0.0099**	**0.0097**
	61.7615	-19.4569	2.0401	0.0154	0.0106	0.0100
	61.7604	-19.4602	2.0413	0.0159	0.01	0.0092
	61.7631	-19.461	2.0417	0.0239	0.0124	0.0100
0.15	**61.6991**	**-19.4370**	**2.0384**	**0.0139**	**0.0101**	**0.0098**
	62.0110	-19.5481	2.0523	0.0117	0.0099	0.0100
	62.0107	-19.5511	2.0497	0.0180	0.0104	0.0098
	62.0152	-19.5527	2.0500	0.0290	0.0137	0.0100
0.175	**61.6867**	**-19.4296**	**2.0372**	**0.0145**	**0.0102**	**0.0097**
	62.2657	-19.6402	2.0568	0.0182	0.0112	0.0100
	62.2562	-19.6403	2.0576	0.0197	0.0109	0.0093
	62.2673	-19.6444	2.0583	0.0342	0.0152	0.0101

0.2	**61.0702**	**-19.2083**	**2.0176**	**0.0142**	**0.0102**	**0.0098**
	63.0193	-19.9148	2.0857	0.0218	0.0120	0.0100
	62.4933	-19.7266	2.0655	0.0211	0.0113	0.0099
	62.5193	-19.736	2.0667	0.0394	0.0168	0.0101
0.225	**61.3216**	**-19.2955**	**2.0262**	**0.0147**	**0.0103**	**0.0097**
	62.7699	-19.8236	2.0734	0.0206	0.0117	0.0100
	62.7021	-19.8024	2.0726	0.0223	0.0117	0.0095
	62.7714	-19.8277	2.0750	0.0447	0.0185	0.0101
0.25	**61.0702**	**-19.2083**	**2.0176**	**0.0142**	**0.0102**	**0.0098**
	63.0193	-19.9148	2.0857	0.0218	0.0120	0.0100
	62.8608	-19.8598	2.078248	0.023015	0.011915	0.0098
	63.0235	-19.9194	2.0833	0.0499	0.0201	0.0101
0.275	**60.8742**	**-19.1325**	**2.0108**	**0.0133**	**0.0100**	**0.0098**
	63.2740	-20.0069	2.0901	0.0229	0.0123	0.0100
	62.9347	-19.8868	2.0805	0.0235	0.0121	0.0096
	63.2756	-20.011	2.0917	0.0551	0.0219	0.0102
0.3	**60.7290**	**-19.0814**	**2.0081**	**0.0122**	**0.0099**	**0.0098**
	63.5235	-20.0981	2.1023	0.0240	0.0125	0.0100
	62.9074	-19.8762	2.0790	0.0236	0.0122	0.0098
	63.5277	-20.1027	2.1000	0.0603	0.0236	0.0102

The above Monte Carlo study compares the cluster-based method and non-cluster-based method when 1/3 of profiles are from the out-of-control process, other ratios of out-of-control profiles have been tried as well and the conclusions obtained are very similar to those for the 1/3 ratio. For example, when the proportion of profiles from the out-of-control process is 1/10, the FPR values for the cluster-based method are 0.1283, 0.0137 and 0.0072 with shift values 0.1, 0.2 and 0.3 respectively. The corresponding FPR values for the non-cluster-based method are 0.1489, 0.2466 and 0.1191. When 4/10 of profiles are from the out-of-control process and for shift values equal 0.1, 0.2 and 0.3, the FCC values for the cluster-based method are 0.6081, 0.7560 and 0.9623, respectively. The corresponding FCC values for the non-cluster-based method are 0.6041, 0.6365 and 0.6956. Our simulation results show that the cluster-based method gives superior results when compared to the non-cluster-based method as long as the shift size is moderate or large for ratio values up to 1/2. At the ratio value of 1/2 both methods are equally distorted.

4.6 Further Analysis based on the Monte Carlo Study

The Monte Carlo study showed that when the shift is very small, the non-cluster-based method works as well as the cluster-based method. However, when shift is not very small, the cluster-based method works uniformly better than the non-cluster-based method. To gain insight into those situations where the cluster-based method outperforms the non-clusted based method, both methods are compared using three simulated data sets based on (4.13) under the condition of either a small, moderate or large shift. The three corresoponding shift values are chosen as shift=0.05, 0.175 and 0.3.

The first Monte Carlo simulation is given when the shift=0.05. The corresponding plots for 30 true PS curves and estimated PS curves are given in Figure 4.5 . The PS curves in blue are from the in-control process and those in red are from the not-in-control process.

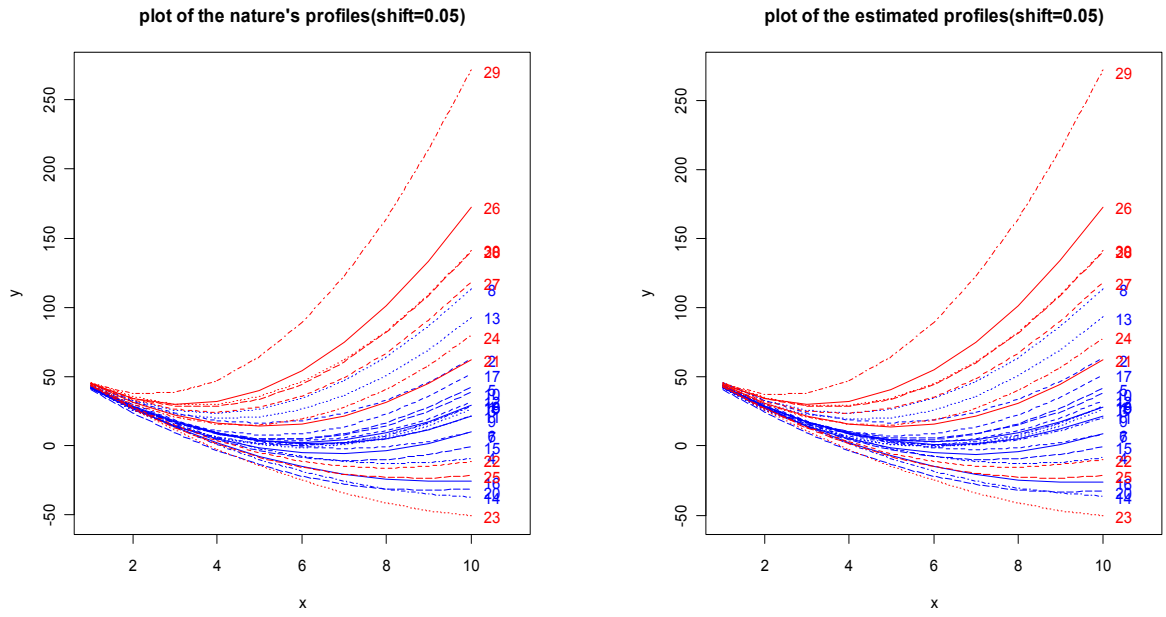

Figure 4.5: Plot of true profiles with shift=0.05

(Red curves represent the profiles from the out-of-control process).

In Figure 4.5, it seems that the 29th profile has an apparently "different" trend compare to other profiles. Also, the 26th and the 28th profiles also show a "different"

trend compare to other profiles. Both methods determined the 29th profile as from the out-of-control process and classified the other 29 profiles as from the in-control process. Thus, nine profiles which from the out-of-control process were misclassified as from the in-control process. The corresponding classification tables are given below.

Table 4.12: Classification table for non-cluster-based method (shift=0.05)

Classified set / Actual set	In-control process	Out-of-control process
In-control process	20	0
Out-of-control process	9	1

Table 4.13: Classification table for cluster-based method (shift=0.05)

Classified set / Actual set	In-control process	Out-of-control process
In-control process	20	0
Out-of-control process	9	1

With these two classification tables, it is easy to show that both methods give the same FCC, Sensitivity, Specificity, FNR and FPR. This one simulation result is consistent with the result obtained in the Monte Carlo study above that the two methods are nearly equivalent when the shift is very small.

Another Monte Carlo simulation is given when shift=0.175 and the corresponding plot for the 30 true and estimated profiles are shown in Figure 4.6. As in the first example, the PS curves in blue are from the in-control process and those in red are from the not-in-control process.

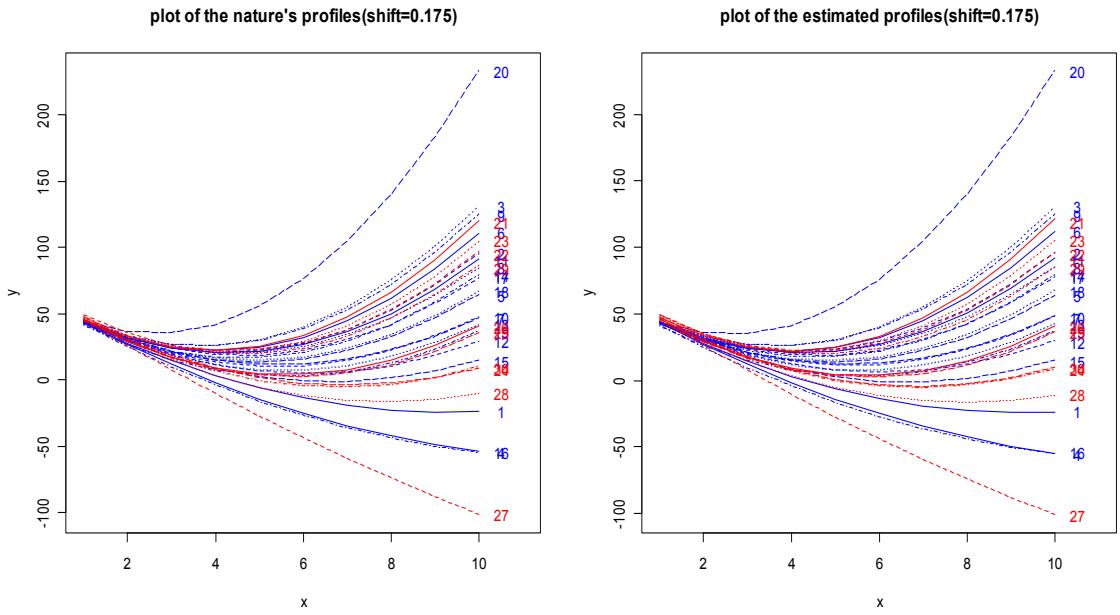

Figure 4.6: Plot of true profiles with shift=0.175

(*Red curves represent the profiles from the out-of-control process*).

In Figure 4.6, it is not easy to conclude that the profiles from the out-of-control process have a "different" trend from the other profiles. One may only conclude that the 20th and the 27th profiles look "different" compare to other profiles. The non-cluster-based method specified the 12th, 19th, 22th, 23th, 25th, 26th, and 30th profiles (which are all located in the middle of the profiles in Figure 4.6) as from the out-of-control process. Thus, this method correctly identified five out of ten profiles that are from the out-of-control process and misclassified two profiles that are from the in-control process as from the out-of-control process. The corresponding classification table is given in Table 4.14.

Table 4.14: Classification table for cluster-based method (shift=0.175)

Classified set / Actual set	In-control process	Out-of-control process
In-control process	18	2
Out-of-control process	5	5

The cluster-based method, on the other hand, correctly identified eight of ten profiles from the out-of-control process, misclassifying the 24^{th} and 28^{th} profiles. Recall in Figure 4.6, one may conclude that both 20^{th} and 27^{th} profiles are from the out-of-control process based on the plot. However, the cluster-based method correctly identified the 27^{th} profile as from the out-of-control process and specified the 20^{th} profile as from the in-control process. Table 4.15 is the classification table for the cluster-based method.

Table 4.15: Classification table for cluster-based method (shift=0.175)

Classified set / Actual set	In-control process	Out-of-control process
In-control process	20	0
Out-of-control process	2	8

Once again, this example gives tabled results very similar to those seen in Table 4.10. It can be showed that when the shift=0.175, both methods cannot correctly identify all the profiles from the out-of-control process. However, the profiles specified from the out-of-control process by the cluster-based method are actually from the out-of-control process, while the profiles specified from the out-of-control process by the non-cluster-based method are not. For example, in this one simulation case, the 12^{th} and 19^{th} profiles are actual from the in-control process but are specified as the out-of-control process by the non-cluster method.

To further see the difference of the two methods, one more simulation is repeated with shift=0.3. The plots of the 30 true and estimated PS curves are given in Figure 4.7. Also, the PS curves in blue are from the in-control process and those in red are from the not-in-control process.

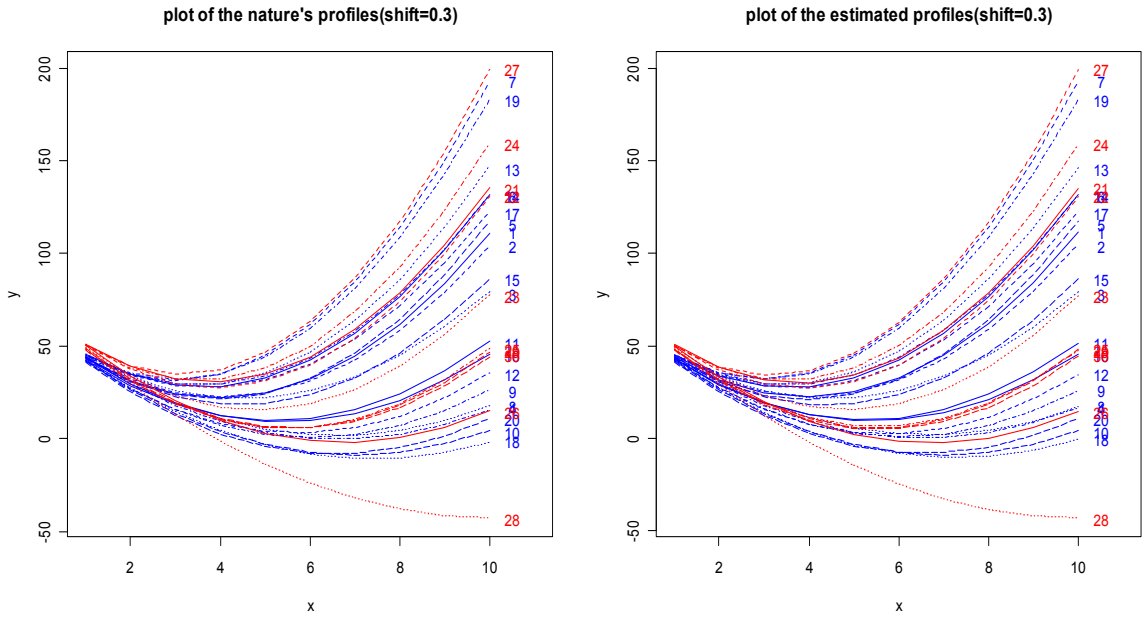

Figure 4.7: Plot of true profiles with shift=0.3

(*Red curves represent the profiles from the out-of-control process*).

Table 4.16: The parameter estimates for 30 profiles (shift=0.3)

Index of profiles	$\hat{\beta}_{1i}$	$\hat{\beta}_{2i}$	$\hat{\beta}_{3i}$
1	61.342	-19.616	2.463
2	59.091	-18.201	2.274
3	58.584	-16.469	1.848
4	58.582	-17.767	1.355
5	62.525	-21.027	2.653
6	61.763	-18.778	2.581
7	61.186	-19.656	3.285
8	61.697	-18.082	1.363
9	58.344	-18.911	1.564
10	60.553	-19.953	1.433
11	60.501	-19.356	1.848
12	60.83	-20.423	1.781
13	60.511	-18.417	2.7
14	58.83	-17.031	2.428
15	62.278	-19.889	2.23
16	59.15	-18.364	1.7
17	61.943	-18.442	2.451
18	59.833	-19.308	1.329
19	59.612	-18.444	3.077

20	60.879	-21.045	1.607
21	**68.363**	**-20.227**	**2.688**
22	**69.019**	**-21.484**	**2.767**
23	**67.704**	**-21.689**	**2.259**
24	**68.502**	**-21.214**	**3.026**
25	**68.072**	**-22.996**	**2.101**
26	**69.203**	**-21.206**	**1.573**
27	**69.394**	**-22.368**	**3.536**
28	**70.453**	**-22.654**	**1.131**
29	**68.889**	**-22.747**	**2.065**
30	**70.411**	**-23.264**	**2.069**

Similar to the previous plots, Figure 4.7 does not reveal the profiles from the out-of-control process as "different" from those profiles from the in-control process except the 28th profile seems to have the obvious "different" trend. However, the estimated coefficients for each profile based on the fixed effects model, listed in Table 4.16, show that the profiles from the out-of-control process do have the "different" estimated coefficients compare to the other profiles. For example, the profiles from the out-of-control process have larger estimates for the intercept term and smaller estimates for the linear term.

When shift=0.3, the non-cluster-based method correctly specified 5 out of the 10 profiles that from the out-of-control process, which are from the 26th to 30th profiles and it misclassified the 3rd profile as from the out-of-control process. The cluster-based method, on the other hand, correctly separated all the profiles from the in-control and out-of-control process. The classification tables for the non-cluster-based method and cluster-based method are Tables 4.17 and 4.18 respectively.

Table 4.17: Classification table for cluster-based method (shift=0.3)

Classified set / Actual set	In-control process	Out-of-control process
In-control process	19	1
Out-of-control process	5	5

Table 4.18: Classification table for cluster-based method (shift=0.3)

Classified set / Actual set	In-control process	Out-of-control process
In-control process	20	0
Out-of-control process	0	10

One might be curious that why the cluster-based method can correclty identify all profiles that from the out-out-control process while the non-cluster-based method can only correctly identify some of them. To gain insight into the reason, the detailed explantation with two 3D plots are listed below.

Recall that both cluster-based method and non-cluster-based method can use the parametric approach to obtain the T^2 statistic. In equation (3.34), one can compute the T^2 statistic for each PS curve by using the standardized distance of its estimated parameter vector to the estimated PA parameter vector. With this property , one can treat each $p \times 1$ parameter vector as a data point in the p dimensional space and the T^2 statisic for the i^{th} PS cruve then can be explained as the standardized distance between the i^{th} data point to the data point that represents the PA parameter vector in this p dimensional space. In the above example, each profile can be represented as a quadratic function and thus p=3. Figures 4.8 and 4.9 give the the 3D plots for the 30 estimated PS parameter vectors and PA parameter vector for both methods, when the shift =0.3.

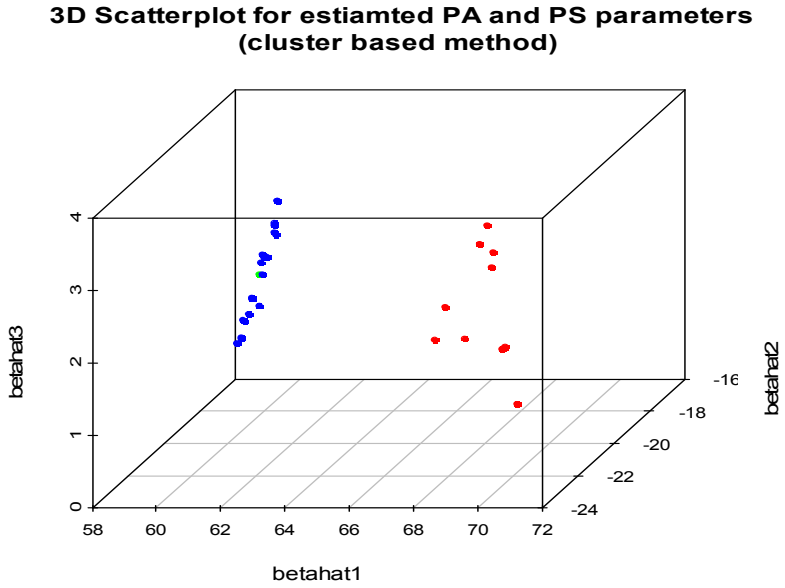

Figure 4.8: 3D Plot of estimated PA and PS parameter vectors when the shift=0.3 (*Blue dots and red dots represent the estimated parameter vectors for profiles from in-control and out-of-control process respectively; the green dot represents the PA parameter vector*)

Recall that the cluster-based method uses the profiles in the final set C_{final} to estimate the PA parameter vector. In this particular example, the final set C_{final} includes all 20 profiles from the in-control process and no profiles from the out-of-control process. Then, the PA parameter vector is estiamted by using all 20 profiles that from the in-control process, which is represented by the green dot in Figure 4.8. In Figure 4.8, the green dot is in the middle of the blue dots (which represent the estimated PS parameter vectors for the profiles from the in-control process). It is seen that all red dots (which represent the estimated PS parameter vectors for the profiles from the out-of-control process) have larger distances to the green dot than that of the blue dots. Thus , it is resonable that no outlying profile can be added to the main cluster by using the cluster-based method. Additionally,Figure 4.8 also shows that the the blue dots are clustered together and the red dots are clustered together, provided a true representation of the fact that all profiles from the in-control process are from one population and all profiless

from out-of-control process are from another population. The non-cluster-based method, on the other hand, estimates the PA profile first by using all 30 profiles and its estimated PA and PS parameters are plotted in Figure 4.9.

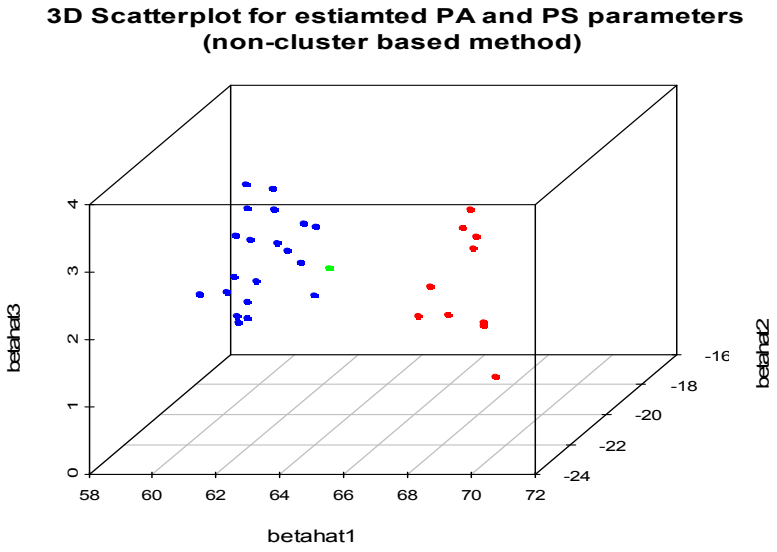

Figure 4.9: 3D Plot of estimated PA and PS parameters
(*Blue dots and red dots represent the profiles from in-control and out-of-control process respectively; the green dot represents the PA parameters)*

Unlike Figure 4.8, Figure 4.9 shows that the green dot (which represents the estimated PA parameter vector) is "pulled" to the direction of the red dots (which represent the estimated outlying parameter vectors). It can be noted that some of the red dots have smaller standardized distances to the green dot than some blue dots (which represent the estimated PS parameter vector for the profiles from the in-control process). As a result, the non-cluster-based method misclassifies some of the profiles from the out-of-control process as from the in-control process and some of the profiles from the in-control process as from the out-of-control process. Using the above classification tables, the performance of each simulation with three different shifts is listed in Table 4.18.

Table 4.19: Performance of one simulation study with different shift
(*The above cells represent the result from the cluster-based method and the bolded cells represent the optimal*)

Shift	FCC	Sensitivity	Specificity	FPR	FNR
0.05	0.700	1.00	0.100	0.310	0.000
	0.700	1.00	0.100	0.310	0.000
0.175	**0.933**	**1.00**	**0.800**	**0.091**	**0.000**
	0.767	0.783	0.500	0.217	0.286
0.3	**1.000**	**1.00**	**1.000**	**0.000**	**0.000**
	0.800	0.95	0.500	0.208	0.167

The numbers in Table 4.19 do not match exactly the corresponding numbers in Table 4.5. After all, the numbers in Table 4.10 are the average of 5,000 repetitions, each repetition of the form of the three examples illustrated here. However, the results are consistent with the conclusion that the cluster-based method works better than non-cluster-based method when we have a moderate or large shift. Further, the simulation examples in this section give a more intuitive explanation that the cluster-based method is superior to the non-cluster-based method even when they have similar POS. For example, in the shift=0.175 example, both methods resulted in a signal but the non-cluster-based method signaled due to correctly detecting five profiles from the out-of-control process and incorrectly detecting two profiles from the in-control process as from the out-of-control process. On the other hand, the cluster-based method signaled due to correctly detecting eight profiles as from the out-of-control process. Similarly, when shift=0.3, both methods signaled but the non-cluster-based method only detected 5 of 10 profiles that from the out-of-control process and misclassified one profile from the in-control process as from the out-of-control process while the cluster-based method detect all 10 profiles that from the out-of-control process and did not misclassify any profiles.

The Monte Carlo study with these three particular simulations illustrates that the cluster-based method provides improved ability to classify the profiles into the proper categories of in-control and out-of-control than the non-cluster-based method for the

parametric case. The Monte Carlo study gives the average performance of the two methods, while these three particular simulations provide a more detailed illustration and some intuitive justification for the superiority of the cluster-based method.

4.7 Chapter Summary

In this chapter, an innovative profile monitoring methodology, which is referred to as the cluster-based profile monitoring method, is introduced for Phase I analysis. The proposed method incorporates a cluster analysis phase to aid in determining if nonconforming profiles are present in the HDS. To cluster the profiles, the proposed method first replaces the data for each profile with an estimated profile curve, using some appropriate regression method, and clusters the profiles based on their estimated parameter vectors. This cluster phase then yields a main cluster which contains more than half of the profiles. The initial estimated PA parameters are obtained by fitting a linear mixed model to those profiles in the main cluster. In-control profiles, determined using the Hotelling's T^2 statistic, that are not contained in the initial main cluster are iteratively added to the main cluster and the mixed model is used to update the estimated PA parameters. A simulated example, a Monte Carlo study and an application to a real data set demonstrate the performance advantage of this proposed method over a current non-cluster-based method with respect to more accurate estimates of the PA parameters and better classification performance in determining those profiles from an in-control process from those from an out-of-control process.

Also, in this chapter, it showed that when the profiles can be represented by m appropriate $p \times 1$ vectors, the profile monitoring process is equivalent to the detection of multivariate outliers. For this reason, we also compare our proposed method to a popular method used to identify outliers when dealing with a multivariate response. More specifically, the successive difference and the MVE methods for estimating the variance-covariance matrix for the estimated profile model parameters are also used in computing both the cluster-based and non-cluster-based procedures. The successive difference

estimator has been recommended for use when the out-of–control process is due to a sustained shift in the profile parameters. The MVE method is commonly suggested for use in detecting multivariate outliers. Our study demonstrates that when the out-of-control process due to a sustained shift, the cluster-based method using the successive difference estimator is clearly the superior method, among those methods we considered, based on all performance criteria.

Chapter 5. Phase II Control Charts based on Phase I Analysis

One use of the estimates from Phase I is to obtain the proper control limits for the Phase II control chart. However, if the HDS in Phase I contains the profiles from the out-of-control process, the estimates from Phase I will be biased, as a result, the control limits obtained in Phase II will be distorted and the corresponding control chart cannot detect the shift quickly. This chapter will show how the Phase I estimates affect the performance in Phase II analysis.

5.1 Profile Monitoring in Phase II

In Phase II analysis, the performance of a control chart is measured using the average run length (ARL). The ARL is the average number of samples that are obtained before a chart signals. This signal can be either a true signal of a change (or shift) in the process or a false alarm. The false alarm means the chart signaled when, in fact, there was no shift. Two corresponding ARLs are utilized during the Phase II process the out-of-control ARL, denote as ARL_1, and the in-control ARL, denote as ARL_0. The out-of-control ARL_1 represents the expected number of samples to signal when a true shift has occurred. While the in-control ARL, ARL_0, denotes the expected number of samples to signal when the process remains in control. In this case, no shift has occurred thus resulting in a false alarm. The goal of Phase II analysis is to detect the small shifts quickly which is equivalent to minimizing the out-of-control ARL_1 over a range of process shifts while requiring a prespecified value of the in-control ARL_0.

Two control charts are comparable if they have the same in-control ARL_0. To compare the performance of two comparable control charts one compares the out-of-control ARL_1 over a range of process shifts. For example, if two control charts have the same in-control ARL_0 and one has smaller out-of-control ARL_1 for all shifts than the

other, that is, one chart signals more quickly than the other control chart, then we can conclude that the control chart with smaller out-of-control ARL_1 works better than the other one.

During the profile monitoring process, the T^2 control chart is commonly used to detect a change in the process. For example, Williams et al. (2007a) and Jensen et al. (2007), used the T^2 control chart to detect the presence of profiles from the out-of-control process during Phase I analysis. For the Phase II analysis, the T^2 control chart is based on the estimates that result from the Phase I analysis. For example, in the previous chapter, the estimated PA parameter $\hat{\boldsymbol{\beta}}_{PA}$ and its variance-covariance matrix $\hat{\boldsymbol{V}}$ are obtained. These estimates would then be used to set the control limit and to compute the T^2 statistic used during the Phase II analysis.

One proposed method to obtain the T^2 control chart in Phase II profile monitoring is to calculate the T^2 statistics for the i^{th} new profile as

$$T_i^2 = \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{PA}\right)^T \hat{\boldsymbol{V}}^{-1} \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{PA}\right), \tag{5.1}$$

where $\hat{\boldsymbol{\beta}}_{PA}$ and $\hat{\boldsymbol{V}}$ are the estimates for the PA parameter and variance-covariance matrix from Phase I HDS and $\hat{\boldsymbol{\beta}}_i$ is the least square estimates for the i^{th} new profile in Phase II. The control limit for the Phase II T^2 control chart is established using $\hat{\boldsymbol{\beta}}_{PA}$ and $\hat{\boldsymbol{V}}$ so that the prespecified value for ARL_0 is achieved. For example, if ARL_0=200, the control limit is set so that when the shift=0, it takes, on average, 200 samples to signal. One only needs to find the upper control limit (UCL), so that, on average, it takes 200 samples for the observed T^2 statistic to exceed the UCL. The UCL is obtained by educated trial and error so that based on 10,000 simulations, the observed $ARL_0 \approx 200$.

Four estimated PA parameters are provided in Chapter 4, two obtained by the non-cluster-based method (Jensen et. al (2007)) (one based on the successive difference estimator of the covariance matrix and the other based on the MVE covariance estimator) and the other two obtained by the proposed cluster-based method (one based on the successive difference estimator of the covariance matrix and the other based on the MVE covariance estimator). In this chapter, two control charts are defined based on the estimated PA parameters determined in Chapter 4. One is the non-cluster-based T^2 control chart. That is, the $\hat{\boldsymbol{\beta}}_{PA}$ and $\hat{\boldsymbol{V}}$, using the successive difference estimator, in equation (5.1) are obtained by the non-cluster-based method provided by Jensen et al. (2007). The other is the cluster-based T^2 control chart where the $\hat{\boldsymbol{\beta}}_{PA}$ and $\hat{\boldsymbol{V}}$, again using the successive difference estimator, in equation (5.1) are obtained by the proposed cluster-based method. Since the results in Chapter 4 demonstrated a poor performance when using the MVE estimator, it will not be considered in this chapter. The evaluation of the cluster-based T^2 control chart and the non-cluster-based T^2 control chart are presented below.

5.2 Detailed Simple Example

Recall the example in Chapter 4 in which we generated first nine profiles from the in-control process as

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_2 + b_{1i})x_{ij} + (\beta_3 + b_{2i})x_{ij}^{\ 2} + \varepsilon_{ij}, i = 1, 2, \ldots, m_1, j = 1, 2, \ldots, n \tag{4.1}$$

and generated the last three profiles as

$$y_{ij} = (\beta_0' + b_{0i}) + (\beta_1' + b_{1i})x_{ij} + (\beta_2' + b_{2i})x_{ij}^{\ 2} + \varepsilon_{ij}, \quad i = m_1 + 1, \ldots, m, \quad j = 1, 2, \ldots, n \tag{4.2}$$

where $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2) = (12.5, -7, 2)$ for the in-control process and $\boldsymbol{\beta}'^T = (\beta_0', \beta_1', \beta_2') = (21.875, -14.5, 3.5)$ for the out-of-control process with n=10, $m_1 = 9, m = 12$, $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 0.5$, and $\sigma^2 = 4$.

The non-cluster-based method detected the 6th profile as from the out-of-control process, and then calculated the estimates for the PA parameter using

$$\hat{\boldsymbol{\beta}}_{PA}^T = (16.2608, \quad -9.7092, \quad 2.1782)$$

and the estimated variance-covariance matrix to

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 3.660 & -2.470 & -0.405 \\ -2.470 & 2.035 & -0.132 \\ -0.405 & -0.132 & 0.493 \end{bmatrix}.$$

The cluster-based method, on the other hand, correctly detected the 10th, 11th and 12th profiles as from the out-of-control process and obtained the estimates for the PA parameter and variance-covariance matrix to be

$$\hat{\boldsymbol{\beta}}_{PA}^T = (12.5493, \text{-}7.2394, 1.7764),$$

and

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 0.203 & 0.120 & 0.270 \\ 0.120 & 0.236 & 0.090 \\ 0.270 & 0.090 & 0.389 \end{bmatrix}$$

respectively.

The control limits are obtained by simulating the in-control process with the Phase I estimates and ARL_0 =200. After calibrating the control limits with ARL_0=200 for the non-cluster-based T^2 control chart and the cluster-based T^2 control chart, the performance of each chart can be evaluated by comparing the ARL_s for different shifts. Recall that the PA profile in this example is

$$\boldsymbol{y}_{PA} = \beta_0 + \beta_1 \boldsymbol{x} + \beta_2 \boldsymbol{x}^2, \tag{5.2}$$

where $\boldsymbol{\beta}^T = (12.5, -7, 2)$. Equation (5.2) can be rewritten as

$$\boldsymbol{y}_{PA} = \beta_{C1} \boldsymbol{x} + \beta_{C2} (\boldsymbol{x} - 2.5)^2, \tag{5.3}$$

where $\boldsymbol{\beta}_C^{\ T} = (\beta_{C1}, \beta_{C2}) = (3, 2)$. Assuming that the out-of-control process has the PA profile

$$\boldsymbol{y}_{PA} = \beta_{C1} \boldsymbol{x} + (\beta_{C2} + shift)(\boldsymbol{x} - 2.5)^2 \tag{5.4}$$

with shift=(0, 0.25, 0.5, 0.75, 1, 1.25, 1.5), the ARL_0 and ARL_1 based on the cluster-based T^2 control chart and non-cluster-based T^2 control chart are given to compare the performance of these two control charts.

When the shift equals 0, the ARL_0 from the cluster-based T^2 control chart and the non-cluster-based T^2 control chart are 66.8264 and 15.6412, respectively. Compared to the ARL_0 =200 from the simulated process, one can see that both control charts have many more false alarms than expected. This is reasonable because our estimates from both methods are not very close to the true parameters when the HDS only contains 9 of profiles from the in-control process. However, the cluster-based T^2 control chart works

much better than the non-cluster T^2 control chart, which is consistent to our results in Chapter 4 that the cluster-based method obtained estimates closer to the true parameters.

The ARL_1 for two control charts are also obtained according to different shift values. Table 5.1 compares ARL_1 from these two control charts. In Table 5.1, the ARL_CB represents the ARL for the cluster-based method and ARL_NCB represents the ARL for the non-cluster-based method.

Table 5.1: ARL_CB and ARL_NCB with $ARL_0 \approx 200$

(*Bolded cells represent the better values*)

Phase II Shift	ARL_CB	ARL_NCB	Phase II Shift	ARL_CB	ARL_NCB
0.25	**97.8645**	105.0901	1	**6.3235**	29.5877
0.5	**31.1519**	932.751	1.25	**2.1247**	5.7479
0.75	**26.9371**	277.8033	1.5	**1.0271**	2.0891

Table 5.1 shows that when the shift is greater than 0 the ARL based on the cluster-based T^2 control chart is uniformly smaller than the ARL based on the non-cluster-based T^2 control chart, especially when the Phase II shift is small. For example, when shift equals 0.5, the ARL based on the non-cluster-based T^2 control chart is 932.751 while the ARL based on the cluster-based T^2 control chart is 36.1519. When the Phase II shift equals 1, the cluster-based T^2 control chart gives ARL=6.3235, which means it can detect the change of process quickly. The non-cluster-based T^2 control chart, on the other hand, on average, takes about 30 samples to detect the change of process.

In addition, Table 5.1 shows that the ARL based on the non-cluster-based T^2 control chart are greater than 200 when the shift is less than 1.This result indicates that the non-cluster-based T^2 control chart takes much longer than expected to detect the change of process at this point. This seemingly illogical event occurs because the non-cluster-based method in Phase I misclassified the profiles from in-control and out-of-

control process, and its estimates are severely biased. Recall that the profiles from the out-of-control process in HDS were generated with PA profile

$$\boldsymbol{y}_{PA} = 21.875 - 14.5\boldsymbol{x} + 3.5\boldsymbol{x}^2,$$

which can be written as

$$\boldsymbol{y}_{PA} = 3\boldsymbol{x} + (2+1.5)\left(\boldsymbol{x} - 2.5\right)^2. \tag{5.5}$$

Note, (5.5) is equivalent to (5.4) with a shift=1.5. The non-cluster-based estimator involves 3 profiles from the out-of-control process with shift=1.5 and as a result, its estimated PA profile is pulled to the direction of the PA profile with shift=1.5. The corresponding Phase II control chart is also distorted, resulting in many more runs, on the average, than expected to detect the change of process when the shift is less than 0.1.

5.3 ARL based on Monte Carlo Study

In the previous chapter, a Monte Carlo study was performed and the average estimated PA parameters and the corresponding estimated variance-covariance matrix were obtained. The Monte Carlo study concluded that the estimated PA parameters from the cluster-based method have smaller bias than the bias from the non-cluster-based method. In this section, a further study is performed to evaluate the performance of two Phase II T^2 control charts based on these estimates in Phase I.

Recall that in the Monte Carlo study in Chapter 4, the comparison of the cluster-based estimates and non-cluster-based estimates are obtained according to different shifts in the process. The Phase II T^2 control charts in this section are also obtained according to same shifts used in Chapter 4. For each value of the Phase I shift, the ARL was obtained to achieve ARL_0=200 based on the both the cluster-based process and the non-cluster-based process. Also, the true PA parameters and the shift values were the

same as those used in the above example. The number of profiles, m, is 30, the number of profiles from the out-of-control process, m-m_1, is 10, the number of observations for each profiles, n, is 10, values taken uniformly on the interval from 0 to 10 for each profile. The number of Monte Carlo repetitions is 10,000.

Table 5.2 lists the average ARLs of the Phase II T^2 control charts, listed for appropriate values of the shift in Phase II, obtained by using the estimates from both methods when the out-of-control process has a shift=0.05 in Phase I. When the Phase I shift=0.05, the simulation results from Chapter 4 showed that the average estimated PA parameter based on the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (61.002, -19.182, 2.016)$ and the estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 0.387 & 0.007 & -0.003 \\ 0.007 & 0.407 & 0.002 \\ -0.003 & 0.002 & 0.441 \end{bmatrix}.$$

With ARL_0=200, the UCL for the cluster-based T^2 control chart is 50.8. The average estimated PA parameter based on the non-cluster-based method from chapter 4 was $\hat{\boldsymbol{\beta}}^T = (61.005, -19.183, 2.016)$ and corresponding estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 0.476 & -0.001 & -0.002 \\ -0.001 & 0.467 & 0.003 \\ -0.002 & 0.003 & 0.501 \end{bmatrix}.$$

With ARL_0=200, the UCL for the non-cluster-based T^2 control chart is 41.4.The ARLs for different Phase II shifts are presented in Table 5.2.

Table 5.2: ARL_CB and ARL_NCB with Phase I shift=0.05, $ARL_0 \approx 200$

(*Bolded cells represent the better values*)

Phase II Shift	ARL_CB	ARL_NCB	Phase II Shift	ARL_CB	ARL_NCB
0	**186.1936**	186.0836	0.175	**1.4145**	1.4175
0.05	**77.6124**	78.2726	0.2	**1.1624**	1.1636
0.075	**22.0527**	22.3944	0.225	**1.0568**	1.0573
0.1	**8.1647**	8.2065	0.25	**1.0144**	1.0152
0.125	**3.6739**	3.7025	0.275	**1.0044**	1.0045
0.15	**2.0643**	2.0747	0.3	**1.0004**	**1.0004**

In Table 5.2, it is easy to see that when the Phase II shift equals 0, both methods have ARL_0 less than 200, that indicates both methods have slightly more false alarms than desired. Also, the ARL_0 from the cluster-based method is slightly largerthan that from the non-cluster-based method which implies that the cluster-based method has fewer false alarms than the non-cluster-based method.

While the Phase II shift is greater than 0 the ARL_1 is less than 200, which shows both control charts can fairly quickly detect a change in the process. Also, all the smaller ARL_1 values are all from the cluster-based T^2 control chart. However, the difference between the ARL_1 values from the cluster-based T^2 control chart and the ones from the non-cluster-based T^2 control chart are very small. For example, when the Phase II shift=0.15, the ARL_CB=2.0643 while the ARL_NCB is a little larger at ARL_NCB=2.0747. The result in Table 5.2 is consistent to the conclusion in the previous chapter that the cluster-based method and non-cluster-based method have similar performance when the process experiences a very small Phase I shift.

Table 5.3 gives the ARL from the cluster-based T^2 control chart and the non-cluster-based T^2 control chart with the process has a Phase I shift=0.075. When Phase I shift=0.075, the simulation results from Chapter 4 showed that the average estimated PA parameter based on the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (61.240, -19.269, 2.024)$ and the estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 0.533 & -0.038 & 0.001 \\ -0.038 & 0.433 & 0.001 \\ 0.001 & 0.001 & 0.457 \end{bmatrix}.$$

With ARL_0=200, the UCL for the cluster-based T^2 control chart is 40.20. The average estimated PA parameter based on the non-cluster-based method from Chapter 4 was $\hat{\boldsymbol{\beta}}^T = (61.257, -19.257, 2.024)$ and the corresponding estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 0.679 & -0.073 & 0.004 \\ -0.073 & 0.491 & 0.001 \\ 0.004 & 0.001 & 0.502 \end{bmatrix}$$

With ARL_0=200, the UCL for the non-cluster-based T^2 control chart is 33.20. The ARLs for different Phase II shifts are presented in Table 5.3.

Table 5.3: ARL_CB and ARL_NCB with Phase I shift=0.075, ARL_0 ≈200
(*Bolded cells represent the better values*)

Phase II Shift	ARL_CB	ARL_NCB	Phase II Shift	ARL_CB	ARL_NCB
0	**124.7831**	122.6432	0.175	**1.5882**	1.5981
0.05	**124.5165**	128.8425	0.2	**1.2248**	1.2342
0.075	**30.7486**	32.3163	0.225	**1.0801**	1.0823
0.1	**11.5982**	12.0644	0.25	**1.0235**	1.0241
0.125	**4.64440**	4.8667	0.275	**1.0056**	1.0073
0.15	**2.51520**	2.5607	0.3	**1.0011**	1.0014

Similar to the results in Table 5.3, Table 5.4, representing a Phase I shift of 0.1, indicates that both control charts have more false alarms than expected when the Phase II shift is 0. With a Phase II shift greater than 0, the ARL values from Table 5.3 are all less than 200. However, the non-cluster-based T^2 control chart has a ARL greater than

200 in Table 5.4. That is, when Phase II shift equals 0.05, the ARL from the non-cluster-based method is 283.445. This result indicates that the non-cluster-based T^2 control chart takes much longer than expected to detect the change of process at this point. This seemingly illogical event occurs because the non-cluster-based method in Phase I misclassified the profiles from the in-control and out-of-control process, and its estimates are severely biased.

Table 5.4: ARL_CB and ARL_NCB with Phase I shift=0.1, $ARL_0 \approx 200$
(*Bolded cells represent the better values*)

Phase II Shift	ARL_CB	ARL_NCB	Phase II Shift	ARL_CB	ARL_NCB
0	**112.556**	105.0964	0.175	**1.7074**	2.0191
0.05	**197.2859**	283.445	0.2	**1.2909**	1.4103
0.075	**52.1539**	69.882	0.225	**1.1283**	1.1553
0.1	**18.8977**	19.6289	0.25	**1.0442**	1.056
0.125	**6.1996**	7.2638	0.275	**1.0123**	1.0145
0.15	**2.9147**	3.4116	0.3	**1.0029**	1.004

When Phase I shift=0.125, the simulation results from chapter 4 showed that the average estimated PA parameter based on the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (61.627, -19.410, 2.037)$ and the estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 0.837 & -0.144 & 0.012 \\ -0.144 & 0.488 & -0.003 \\ 0.013 & -0.003 & 0.472 \end{bmatrix}.$$

With ARL_0=200, the UCL for the cluster-based T^2 control chart is 30.3. The average estimated PA parameter based on the non-cluster-based method from Chapter 4 was $\hat{\boldsymbol{\beta}}^T = (61.509, -19.367, 2.032)$ and corresponding estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 0.924 & -0.163 & 0.012 \\ -0.163 & 0.524 & -0.001 \\ 0.012 & -0.001 & 0.502 \end{bmatrix}$$

With ARL_0=200, the UCL for the non-cluster-based T^2 control chart is 20.3. The ARLs for different Phase II shifts are presented in Table 5.5.

Table 5.5: ARL_CB and ARL_NCB with Phase I shift=0.125, $ARL_0 \approx 200$
(*Bolded cells represent the better values*)

Phase II Shift	ARL_CB	ARL_NCB	Phase II Shift	ARL_CB	ARL_NCB
0	**107.9900**	99.4453	0.175	**2.3053**	2.8679
0.05	**445.0099**	634.667	0.2	**1.5371**	1.7763
0.075	**105.7215**	173.0917	0.225	**1.2057**	1.3101
0.1	**27.6380**	43.489	0.25	**1.0575**	1.1138
0.125	**9.4015**	21.5579	0.275	**1.0219**	1.0403
0.15	**4.1336**	7.8762	0.3	**1.0053**	1.0090

As the Phase I shift increases to 0.125, the non-cluster-based T^2 control chart requires more runs to signal than the cluster-based chart even when the Phase II shift is relatively small. For example, Table 5.5 shows that the non-cluster-based T^2 control chart requires 42% to 60% more runs to signal than the cluster-based chart for Phase II shifts less than or equal to 0.1.

When the Phase I shift=0.15, the simulation results from Chapter 4 showed that the average estimated PA parameter based on the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (61.711, -19.441, 2.039)$ and the estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 0.966 & -0.190 & 0.014 \\ -0.190 & 0.507 & -0.004 \\ 0.014 & -0.004 & 0.476 \end{bmatrix}.$$

With ARL_0=200, the UCL for the cluster-based T^2 control chart is 27.39. The average estimated PA parameter based on the non-cluster-based method from Chapter 4 was $\hat{\boldsymbol{\beta}}^T = (62.014, -19.550, 2.049)$ and corresponding estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 1.425 & -0.347 & 0.029 \\ -0.347 & 0.591 & -0.008 \\ 0.029 & -0.008 & 0.503 \end{bmatrix}$$

With ARL_0=200, the UCL for the non-cluster-based T^2 control chart is 21.58. The ARLs for different Phase II shifts are presented in Table 5.6.

Table 5.6: ARL_C and ARL_NCB with Phase I shift=0.15, $ARL_0 \approx 200$
(*Bolded cells represent the better values*)

Phase II Shift	ARL_CB	ARL_NCB	Phase II Shift	ARL_CB	ARL_NCB
0	**104.4567**	97.5154	0.175	**2.6131**	4.3159
0.05	**590.4931**	927.2438	0.2	**1.6734**	2.3991
0.075	**146.7290**	392.7659	0.225	**1.2665**	1.5832
0.1	**36.5752**	98.9797	0.25	**1.0961**	1.3512
0.125	**11.8519**	27.6586	0.275	**1.0315**	1.0824
0.15	**4.95580**	9.7622	0.3	**1.0067**	1.0257

Table 5.6, when the Phase I shift is 0.15, shows that the non-cluster-based T^2 control chart requires 57% up to 97% more runs to signal that the cluster-based chart for Phase II shifts less than or equal to 0.15. As can be seen, the highly biased and more variable non-cluster-based estimates from Phase I have caused the Phase II T^2 control chart for the non-cluster method to be very inefficient at detecting small shifts in the process.

Note, from Table 5.2 to Table 5.6, representing Phase I shifts from 0.05 to 0.15, the ARL from both the cluster-based and the non-cluster-based T^2 control charts are increasing when Phase II shift is greater than 0 and decreasing when Phase II shift equals 0. For example, in Table 5.2 with the Phase I shift=0.05, the ARL_CB=186.1956 when

the Phase II shift=0, while in Table 5.3 with Phase I shift=0.075, ARL_CB=124.7831 when the Phase II shift =0. Table 5.4 and Table 5.5, with Phase I shift=0.1 and 0.125, give ARL_CB=112.5560 and ARL_CB=107.9900 for Phase II shifts of 0, respectively. Table 5.6 with Phase I shift=0.15, on the other hand, gives ARL_C=104.4567 when the Phase II shift=0. While the Phase II shift equals 0.1, the ARL from the cluster-based method increases from 8.1647 to 36.5752 and the ARL from the non-cluster-based method increases from 8.2065 to 98.9797. This pattern changes with Table 5.7.

When the Phase I shift=0.175, the simulation results from Chapter 4 showed that the average estimated PA parameter based on the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (61.6867, -19.4296, 2.0372)$ and the estimated variance-covariance matrix as

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 1.047 & -0.222 & 0.020 \\ -0.222 & 0.519 & -0.005 \\ 0.020 & -0.005 & 0.474 \end{bmatrix}.$$

With ARL_0=200, the UCL for the cluster-based T^2 control chart is 25.09. The average $\hat{\boldsymbol{\beta}}^T = (62.2657, -19.6402, 2.0568)$ and the corresponding estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 1.662 & -0.434 & 0.037 \\ -0.434 & 0.623 & -0.010 \\ 0.037 & -0.010 & 0.503 \end{bmatrix}.$$

With ARL_0=200, the UCL for the non-cluster-based T^2 control chart is 20.69. The ARLs for different Phase II shifts are presented in Table 5.7.

Table 5.7: ARL_CB and ARL_NCB with Phase I shift=0.175, $ARL_0 \approx 200$
(*Bolded cells represent the better values*)

Phase II	ARL_C	ARL_NCB	Phase II	ARL_C	ARL_NCB

Shift			Shift		
0	**106.6622**	89.1635	0.175	**2.4277**	7.1194
0.05	**420.5775**	955.2781	0.2	**1.5893**	3.4238
0.075	**112.1817**	746.6541	0.225	**1.2299**	2.0361
0.1	**29.6726**	218.4644	0.25	**1.0830**	1.428
0.125	**10.1933**	58.7566	0.275	**1.0257**	1.1647
0.15	**4.4187**	18.3763	0.3	**1.0063**	1.0596

Comparing Table 5.7 with Phase I shift =0.175, to Table 5.6, one can see that, when Phase II shift equals 0, the ARL from the cluster-based T^2 control chart is increasing while the ARL from the non-cluster-based T^2 control chart continues decreasing. That means the expected number of false alarms from the cluster-based T^2 control chart are decreasing while the expected number of false alarms from the non-cluster-based T^2 control chart continues increasing. When the Phase II shift is greater than 0, the ARL from the cluster-based T^2 control chart is decreasing while the ARL_1 from the non-cluster-based T^2 control chart continues increasing. This pattern holds for all Phase II shifts greater than or equal to 0.05 and for all Phase I shift greater than or equal to 0.175. On the other hand, the ARL_0 values for the non-cluster-based method continue to decrease, and the ARL_1 values continue to increase for most values of the Phase II shifts across the Phase I shifts from 0.175 to 0.3.

When Phase I shift =0.2, the simulation results from Chapter 4 showed that the average estimated PA parameter based on the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (61.5422, -19.3761, 2.0338)$ and the estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 1.097 & -0.243 & 0.022 \\ -0.243 & 0.527 & -0.006 \\ 0.022 & -0.006 & 0.474 \end{bmatrix}.$$

With ARL_0=200, the UCL for the cluster-based T^2 control chart is 23.00. The average estimated PA parameter based on the non-cluster-based method from Chapter 4 was

$\hat{\boldsymbol{\beta}}^T = (62.5151, -19.7314, 2.0690)$ and the corresponding estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 1.895 & -0.520 & 0.044 \\ -0.520 & 0.654 & -0.013 \\ 0.044 & -0.013 & 0.503 \end{bmatrix}.$$

With ARL_0=200, the UCL for the non-cluster-based T^2 control chart is 20.19. The ARLs for different Phase II shifts are presented in Table 5.8.

Table 5.8: ARL_CB and ARL_NCB with Phase I shift=0.2, $ARL_0 \approx 200$
(*Bolded cells represent the better values*)

Phase II Shift	ARL_CB	ARL_NCB	Phase II Shift	ARL_CB	ARL_NCB
0	**131.6531**	78.8067	0.175	**2.3022**	12.4719
0.05	**373.5815**	815.1500	0.2	**1.5347**	5.3136
0.075	**94.7297**	1000.0640	0.225	**1.2069**	2.7970
0.1	**26.0579**	426.6021	0.25	**1.0754**	1.7611
0.125	**9.1649**	17.3800	0.275	**1.0223**	1.3601
0.15	**4.0700**	35.6289	0.3	**1.0052**	1.1121

When Phase I shift =0.225, the simulation results from Chapter 4 showed that the average estimated PA parameter based on the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (61.3216, -19.2955, 2.0262)$ and the estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 1.079 & -0.235 & 0.019 \\ -0.235 & 0.526 & -0.005 \\ 0.019 & -0.005 & 0.474 \end{bmatrix}.$$

With ARL_0=200, the UCL for the cluster-based T^2 control chart is 21.53. The average estimated PA parameter based on the non-cluster-based method from chapter 4 was

$\hat{\boldsymbol{\beta}}^T = (62.7699, -19.8236, 2.0734)$ and the corresponding estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 2.129 & -0.605 & 0.052 \\ -0.605 & 0.685 & -0.016 \\ 0.052 & -0.016 & 0.504 \end{bmatrix}.$$

With ARL_0=200, the UCL for the non-cluster-based T^2 control chart is 19.80. The ARLs for different Phase II shifts are presented in Table 5.9.

Table 5.9: ARL_CB and ARL_NCB with Phase I shift=0.225, $ARL_0 \approx 200$

(*Bolded cells represent the better values*)

Phase II Shift	ARL_C	ARL_NCB	Phase II Shift	ARL_C	ARL_NCB
0	**147.4249**	69.558	0.175	**1.8311**	22.6995
0.05	**180.6139**	659.9176	0.2	**1.3322**	8.6265
0.075	**46.9285**	1031.412	0.225	**1.1226**	3.9904
0.1	**18.3672**	2464.647	0.25	**1.0499**	2.3064
0.125	**6.9726**	777.4915	0.275	**1.0108**	1.5469
0.15	**3.3172**	202.6	0.3	**1.0029**	1.2152

In Table 5.8 and Table 5.9, one can see the cluster-based T^2 control chart can detect the Phase II shifts in the range from 0.05 to 0.2 with far fewer runs on average than the non-cluster-based method. For example, when the Phase I shift is 0.225 and the Phase II shift is 0.1, the non-cluster-based method requires 134 times more runs than the cluster-based method.

More comparisons of the ARL based on the cluster-based T^2 control chart and the non-cluster-based T^2 control chart are given in Table 5.10, Table 5.11 and Table 5.12 with Phase I shift=0.25, 0.275, and 0.3 respectively.

When Phase I shift =0.25, the simulation results from Chapter 4 showed that the average estimated PA parameter based on the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (61.0702, -19.2083, 2.0176)$ and the estimated variance-covariance matrix as

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 1.009 & -0.205 & 0.015 \\ -0.205 & 0.518 & -0.005 \\ 0.015 & -0.005 & 0.477 \end{bmatrix}.$$

With ARL_0=200, the UCL for the cluster-based T^2 control chart is 21.00. The average estimated PA parameter based on the non-cluster-based method from Chapter 4 was $\hat{\boldsymbol{\beta}}^T = (63.0193, -19.9148, 2.0857)$ and the corresponding estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 2.368 & -0.693 & 0.060 \\ -0.693 & 0.717 & -0.019 \\ 0.060 & -0.019 & 0.504 \end{bmatrix}.$$

With ARL_0=200, the UCL for the non-cluster-based T^2 control chart is 19.60. The ARLs for different Phase II shifts are presented in Table 5.10.

Table 5.10: ARL_CB and ARL_NCB with Phase I shift=0.25, $ARL_0 \approx 200$

(*Bolded cells represent the better values*)

Phase II Shift	ARL_CB	ARL_NCB	Phase II Shift	ARL_CB	ARL_NCB
0	**141.1499**	62.3618	0.175	**1.4785**	43.2735

0.05	**77.9210**	543.7347	0.2	**1.1837**	15.1529
0.075	**22.6256**	946.8634	0.225	**1.0669**	6.2917
0.1	**8.1912**	849.1215	0.25	**1.0192**	3.18
0.125	**3.7372**	402.229	0.275	**1.0045**	1.9456
0.15	**2.1641**	135.1522	0.3	**1.0007**	1.3942

When the Phase I shift =0.275, the simulation results from Chapter 4 showed that the average estimated PA parameter based on the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (60.8742, -19.1315, 2.0108)$ and the estimated variance-covariance matrix as

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 0.886 & -0.158 & 0.011 \\ -0.158 & 0.504 & -0.003 \\ 0.011 & -0.003 & 0.479 \end{bmatrix}.$$

With ARL_0=200, the UCL for the cluster-based T^2 control chart is 21.95. The average estimated PA parameter based on the non-cluster-based method from chapter 4 was $\hat{\boldsymbol{\beta}}^T = (63.2740, -20.0069, 2.0901)$ and corresponding estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 2.615 & -0.783 & 0.068 \\ -0.783 & 0.750 & -0.022 \\ 0.068 & -0.022 & 0.504 \end{bmatrix}.$$

With the ARL_0=200, the UCL for the non-cluster-based T^2 control chart is 19.45. The ARLs for different Phase II shifts are presented in Table 5.11.

Table 5.11: ARL_CB and ARL_NCB with Phase I shift=0.275, $ARL_0 \approx 200$

(*Bolded cells represent the better values*)

Phase II Shift	ARL_CB	ARL_NCB	Phase II Shift	ARL_C	ARL_NCB
0	**150.7668**	54.5125	0.175	**1.3088**	79.1809

0.05	**41.7068**	435.2054	0.2	**1.1170**	26.8399
0.075	**14.1886**	802.5751	0.225	**1.0378**	10.2806
0.1	**5.7258**	881.0241	0.25	**1.0088**	4.6416
0.125	**2.8796**	571.9755	0.275	**1.0026**	2.5887
0.15	**1.7900**	231.1461	0.3	**1.0003**	1.6871

When Phase I shift =0.3, the simulation results from chapter 4 showed that the average estimated PA parameter based on the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (61.7290, -19.0814, 2.0081)$ and the estimated variance-covariance matrix as

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 0.748 & -0.108 & 0.006 \\ -0.108 & 0.489 & -0.002 \\ 0.006 & -0.002 & 0.481 \end{bmatrix}$$

With ARL_0=200, the UCL for the cluster-based T^2 control chart is 24.55. The average estimated PA parameter based on the non-cluster-based method from Chapter 4 was $\hat{\boldsymbol{\beta}}^T = (63.5235, -20.0981, 2.1032)$ and corresponding estimated variance-covariance matrix was

$$\hat{\boldsymbol{V}} = \begin{bmatrix} 2.874 & -0.878 & 0.076 \\ -0.878 & 0.785 & -0.025 \\ 0.076 & -0.025 & 0.504 \end{bmatrix}.$$

With ARL_0=200, the UCL for the non-cluster-based T^2 control chart is 19.27. The ARLs for different Phase II shifts are presented in Table 5.12.

Table 5.12: ARL_CB and ARL_NCB with Phase I shift=0.3, $ARL_0 \approx 200$

(*Bolded cells represent the better values*)

Phase II Shift	ARL_CB	ARL_NCB	Phase II Shift	ARL_CB	ARL_NCB
0	**189.3782**	50.3346	0.175	**1.2450**	146.0853
0.05	**33.8738**	371.4248	0.2	**1.0871**	51.5057

0.075	**11.1115**	706.0651	0.225	**1.0272**	18.4944
0.1	**4.7847**	891.7677	0.25	**1.0063**	7.5449
0.125	**2.5182**	718.3847	0.275	**1.0014**	3.6936
0.15	**1.6267**	375.3079	0.3	**1.0003**	2.1833

All three tables show that the cluster-based T^2 control chart can now detect a change in the process at all different Phase II shifts with far fewer than the default ARL_0 of 200. The non-cluster-based T^2 control chart continues to require a very large number of runs to detect a process shift, especially for Phase II shifts in the 0.05 to 0.2 range. These three tables show that the cluster-based T^2 control chart can detect the change of process very quickly even the shift value is very small. For example, Table 5.12 shows that, on average, it only takes less than 34 samples to signal on the average when the shift=0.05, and it signals almost immediately when shift is greater than or equal to 0.15. Also, these three tables show that the non-cluster-based T^2 control chart has many more false alarms than that for the cluster-based T^2 control chart when the Phase II shift equals 0.

The eleven tables in this chapter show how the Phase I shift in HDS affects the Phase II ARL for both the cluster-based T^2 control chart and the non-cluster-based T^2 control chart. One can conclude that the cluster-based T^2 control chart works uniformly better than the non-cluster-based T^2 control chart when the Phase I shift is small or moderate. When the Phase I shift is large, say 0.175 or greater, the performance of the cluster-based T^2 chart is clearly much better than the performance of the non-cluster-based chart especially so for small Phase II shift values. For example, the great improvement in the cluster-based T^2 control chart begins to be seen for a Phase I shift of 0.175 and greater as seen in Tables 5.6 -5.12. The reason is that the cluster-based method in Phase I, by clustering the profiles before estimating the PA parameters, is more likely to cluster the profiles from out-of-control process (one third of profiles are from the out-of-control process in the example) in the HDS as a minor cluster when Phase I shift is large. Thus, the PA parameters, estimated by only using the in-control profiles in the main cluster, are more accurate and precise. On the other hand, the estimated PA

parameters from the non-cluster-based method, using all profiles in HDS, will be severely biased when the shift is moderate or large. Consequently, the better the Phase I estimates are, the better the Phase II results will be as indicated by smaller ARL_1 values for smaller Phase II shifts. The poorer the Phase I estimates are, the poorer the Phase II results will be as indicated by very large ARL_1 values will be, even for moderate Phase II shifts.

5.4 Chapter Summary

In this chapter, the ARLs obtained by using the cluster-based T^2 control chart were compared to those obtained by using the non-cluster-based T^2 control chart in Phase II analysis. Both the simple example and the Monte Carlo study showed that the estimates from the Phase I analysis can dramatically affect the performance in Phase II analysis. The cluster-based T^2 control chart can be far more efficient in detecting Phase II shifts than the non-cluster-based T^2 control chart.

Chapter 6. Cluster-Based Nonparametric Profile Monitoring

The proposed cluster-based profile monitoring method in previous chapters is based on the estimated parametric profiles fit to data in the HDS. However, in many cases, the profiles cannot be well represented by a parametric function. This chapter will present the cluster-based method for nonparametric profile monitoring in Phase I analysis.

6.1 Cluster-Based Nonparametric Profile Monitoring

In Chapter 4, the cluster-based profile monitoring for parametric profiles has been presented. In parametric profile monitoring, the profiles are represented by m appropriate $p \times 1$ estimated parameter vectors. The cluster-based method clusters the profiles based on their estimated parameter vectors. After the clustering phase, an initial main cluster set with at least half of the profiles in the HDS is obtained and an initial estimated PA parameter vector is calculated. The profiles in the minor sets may be sequentially updated to the initial main cluster set to form a final main cluster set. Finally, the PA parameter vector is estimated based on the profiles in the final main cluster set and the control limits for Phase II can be set by using the estimated PA parameters.

In Chapter 4, the cluster-based method was demonstrated to more correctly identify those parametric profiles from the in-control process and out-of-control process than a non-cluster-based method. However, in some situations, the quality of the product or process is best represented by a nonparametric relationship between a response variable and some explanatory variables. In this case, the profiles can be represented by equation (3.6) in Chapter 3

$$y_{ij} = f\left(x_{ij}\right) + \xi_i\left(x_{ij}\right) + \varepsilon_{ij}\,, i = 1,2,3,...m \;\; j = 1,2,3,...n.$$

Here $f\left(x_{ij}\right)$ and $\xi_i\left(x_{ij}\right)$ are some nonparametric functions. Complete details describing the monitoring of nonparametric profiles via a non-cluster-based method can be found in Chapter 3. The goal of this chapter is to apply the cluster-based method to monitor the nonparametric profiles.

Recall that the first step when using the cluster-based method for the parametric profile monitoring is to represent each profile by a $p \times 1$ estimated parameter vector and then cluster the profiles based on these estimated parameter vectors. Nonparametric profile monitoring proceeds in exactly the same way as parametric profile monitoring. That is, each nonparametric profile is represented by a $p \times 1$ estimated parameter vector obtained by some appropriate nonparametric regression method. P-spline is the method used in this dissertation but other methods could also be considered. The cluster-based method is then applied to cluster the profiles based on these m estimated vectors.

In Chapter 3, it was shown that the nonparametric model in equation (3.6), can be written as

$$y_{ij} = f_i\left(x_{ij}\right) + \varepsilon_{ij}, i = 1,2,3,...m \;\; j = 1,2,3,...n.$$

where $f_i\left(x_{ij}\right) = f\left(x_{ij}\right) + \xi_i\left(x_{ij}\right)$ is some nonparametric function. A first-order p-spline regression model using the truncated power basis of degree one can be used to represent $f_i\left(x_{ij}\right)$ as

$$f_i\left(x_{ij}\right) \approx \beta_{0i} + \beta_{1i} + \sum_{k=1}^{K} \mu_{ki}\left(x_{ij} - \kappa_k\right)_+, i = 1,2,3,...m, \;\; j = 1,2,3,...n., \quad (6.1)$$

where $k = 1,2,3,...,K$, with K as the number of knots. Define the $(K+2) \times 1$ vector ϕ_i as $\phi_i = \left(\beta_{0i} \quad \beta_{1i} \quad \mu_{1i} \quad ... \quad \mu_{Ki}\right)$ and then the estimated fitted curve for the i^{th} profile can be represented by $\hat{\phi}_i = \left(\hat{\beta}_{0i} \quad \hat{\beta}_{1i} \quad \hat{\mu}_{1i} \quad ... \quad \hat{\mu}_{Ki}\right)$.

After obtaining the $p \times 1$ vector $\hat{\phi}_i$ for each profile with $p = K + 2$, the next step is to use these $\hat{\phi}_i$ to cluster the profiles in order to obtain an initial main cluster set. Using the initial main cluster set, an initial estimated PA profile, based on the nonparametric mixed model, will be used to calculate the T^2 statistics for the profiles in the minor sets. The profiles with in-control T^2 statistics will be added to the initial main cluster set to form a new main cluster set and the estimated PA will be updated based on this new main cluster set. The above procedure is repeating by updating the profiles in the minor sets until either the smallest T^2 statistic is beyond the control limits or all profiles have been added to the main cluster set. Here the in-control profiles are the profiles with the T^2 statistic less than or equal to $\chi^2_{(df,1-\alpha/m)}$, where α represents the significant level, m is the number of profiles in the HDS and *df* represents the degrees of freedom, which is equal to $K+1$.

The T^2 statistics for the profiles in the minor set will be calculated as

$$T_i^2 = \left(\hat{\phi}_i - \bar{\phi}\right)^T \hat{V}^{-1} \left(\hat{\phi}_i - \bar{\phi}\right) \tag{6.1}$$

where $$\hat{V} = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \left(\hat{\phi}_{i+1} - \hat{\phi}_i\right)\left(\hat{\phi}_{i+1} - \hat{\phi}_i\right)^T$$

and $\bar{\phi}$ is based on the estimated PA profile in the main cluster set, and it will be updated each time after new profiles added to the main cluster. Complete details of the algorithm are illustrated in the automobile engine application example in Section 6.2.

6.2 An Automobile Engine Application

Recall that in the automobile engine example in Chapter 4, the quality of the engine is represented by the relationship between the torque produced by the engine and the engine's speed in RPM. In Chapter 4, a parametric model is used to represent the relationship between torque and speed in RPM for each engine. However, Abdel-Salam et al. (2013) showed that the relationship can be better represented by a nonparametric

model. In this section, we will apply the cluster-based method when it is assumed, that the relationship between the torques and speed in RPM is nonparametric.

The first step of the proposed algorithm is to use p-spline regression to fit to the data for each engine data with the model

$$f_i\left(x_{ij}\right)=\beta_{0i}+\beta_{1i}+\sum_{k=1}^{K}\mu_{ki}\left(x_{ij}-\kappa_k\right)_+ +\varepsilon_{ij},\ i=1,2,3,....20.\ j=1,2,3,...,14.\ \kappa=1,2,3,...,\mathrm{K}.$$

Define $\phi_i=\left(\beta_{0i}\quad \beta_{1i}\quad \mu_{1i}\quad ...\quad \mu_{Ki}\right)$ and the corresponding $\hat{\phi}_i$ for the i^{th} engine can be represented as $\hat{\phi}_i=\left(\hat{\beta}_{0i}\quad \hat{\beta}_{1i}\quad \hat{\mu}_{1i}\quad ...\quad \hat{\mu}_{Ki}\right)$. In this case, since there are 14 observations for each engine, choosing K=4 equally spaced knots seems reasonable. Table 6.1 lists $\hat{\phi}_i,\quad \hat{\phi}_i$, for each engine.

Using the $\hat{\phi}_i,\ i=1,2,...,14$, in Table 6.1 and their corresponding estimated covariance matrix, we obtain the similarity matrix which is then used to cluster the engines. The cluster history is listed in Table 6.2. One can see that the initial main cluster set contains 9 profiles at step 17 and that 6 more profiles are added to this initial main cluster set in cluster step 18, resulting in 15 profiles in the main cluster. Since this is the first step that the main cluster set contains greater than half of the profiles, the cluster step of the algorithm stops here. The cluster history (Table 6.2) shows that the proposed algorithm ended up with 15 engines in the initial main cluster set, consisting of engines 1-10, and 13-17. The corresponding estimated PA profile is then obtained by fitting the p-spline mixed model to the data for the 15 engines in the main cluster. The estimated PA profile is

$$y_{PA,j}=f(x_j)\approx\hat{\beta}_0+\hat{\beta}_1+\sum_{k=1}^{K=4}\hat{\mu}_k\left(x_j-\kappa_k\right)_+$$

Define $\overline{\phi}=\left(\hat{\beta}_0\quad \hat{\beta}_1\quad \hat{\mu}_1\quad \hat{\mu}_2\quad \hat{\mu}_3\quad \hat{\mu}_4\right)$,and the $\overline{\phi}$ based on 15 engines is

$$\overline{\phi}=\left(71.831\quad 0.0160\quad -0.0176\quad -0.0040\quad -0.0071\quad -0.0151\right).$$

Using $\bar{\phi}$ and $\hat{\phi}_i$ in equation (6.1), the T^2 statistics for the engines not included in the initial main cluster set are calculated and listed below.

Index of Engines	11	12	18	19	20
T_i^2	18.450	14.564	8.762	17.544	20.387

The cutoff value for the T^2 statistic here is $\chi^2_{1-\frac{\alpha}{m},df} = 18.38$, where $\alpha = 0.05$ and $df = K + p = 5$. According to the observed T^2 statistics and the cutoff value, all engines in the minor sets will be added to the initial main cluster set except the 11th engine and 20th engine. A new main cluster set with engines 1-10, and engines 12-19 is obtained after added the in-control engines, and the estimated PA profile is then updated by fitting p-spline mixed model using this new main cluster set. The updated $\bar{\phi}$ for the updated estimated PA profile is

$$\bar{\phi} = \begin{pmatrix} 70.999 & 0.0165 & -0.0185 & -0.0035 & -0.0072 & -0.0153 \end{pmatrix}$$

The T^2 statistics for the 11th and 20th engines were then updated by using this updated $\bar{\phi}$ in equation (6.1) and the updated T^2 statistics is listed in the following table.

Index of Engines	11	20
T_i^2	19.644	18.559

The updated T^2 statistics for 11th and 20th engines are still greater than the cutoff value 18.33, so we fail to add them to the main cluster set and conclude that these two engines probably have some mechanical issues or other issues. This agrees with the results found using a nonparametric and semiparametric mixed model profile methods used by Abdel-Salam et al. (2013).

Table 6.1: Estimated $\hat{\phi}_i$, $i = 1, 2..14$ for each engine

Index of Engines	$\hat{\beta}_{0i}$	$\hat{\beta}_{1i}$	$\hat{\mu}_{1i}$	$\hat{\mu}_{2i}$	$\hat{\mu}_{3i}$	$\hat{\mu}_{4i}$
1	73.3139	0.016	-0.0141	-0.0085	-0.0065	-0.0157
2	71.7374	0.0157	-0.0119	-0.0107	-0.0056	-0.0139
3	73.6999	0.0146	-0.0173	-0.0019	-0.0074	-0.0139
4	75.8218	0.0134	-0.0153	-0.0037	-0.0063	-0.0152
5	74.4416	0.0149	-0.0127	-0.009	-0.0078	-0.0143
6	79.9753	0.0128	-0.0133	-0.0044	-0.0084	-0.0147
7	66.3589	0.0187	-0.0193	-0.0057	-0.0074	-0.0147
8	71.8467	0.0157	-0.0119	-0.0107	-0.0056	-0.0139
9	70.2356	0.0174	-0.0169	-0.0069	-0.0075	-0.0136
10	80.1105	0.0128	-0.0126	-0.0057	-0.0079	-0.0125
11	71.6214	0.0172	-0.0207	-0.0031	-0.008	-0.0154
12	68.9162	0.0188	-0.0214	-0.0034	-0.0075	-0.0150
13	66.0206	0.0179	-0.0211	-0.002	-0.0068	-0.0143
14	65.7138	0.0185	-0.0203	-0.0042	-0.0063	-0.0166
15	70.4448	0.0162	-0.0184	-0.0031	-0.0083	-0.0132
16	75.8862	0.0141	-0.0166	-0.0023	-0.0081	-0.0151
17	71.938	0.0163	-0.0227	0.0024	-0.0084	-0.0143
18	70.6005	0.0172	-0.018	-0.0043	-0.0085	-0.0143
19	62.7847	0.0191	-0.0243	-0.0001	-0.0062	-0.0181
20	74.8780	0.0149	-0.0154	-0.004	-0.0088	-0.015

The cluster dendrogram in Figure 6.1 shows that engine 11 and 12, engine 18, 19 and 20 are clustered in the same minor sets respectively. After the sequentially addition of the remaining engines to the initial main cluster set, the cluster-based method identified engine 11 and engine 20 as from the out-of-control process and engine 12, 18 and engine 19 as from the in-control process

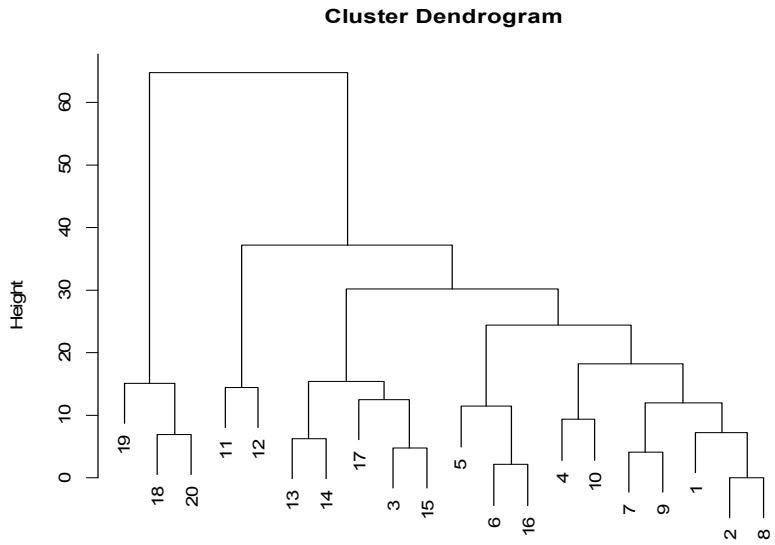

Figure 6.1: Dendrogram for clustering of 20 engines by nonparametric approach

Table 6.2: Cluster history based on eblups for 20 engines

Step	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	1	2	3	4	5	6	7	2	8	9	10	11	12	13	14	15	16	17	18	19
3	1	2	3	4	5	6	7	2	8	9	10	11	12	13	14	6	15	16	17	18
4	1	2	3	4	5	6	7	2	7	8	9	10	11	12	13	6	14	15	16	17
5	1	2	3	4	5	6	7	2	7	8	9	10	11	12	3	6	13	14	15	16
6	1	2	3	4	5	6	7	2	7	8	9	10	11	11	3	6	12	13	14	15
7	1	2	3	4	5	6	7	2	7	8	9	10	11	11	3	6	12	13	14	13
8	1	1	2	3	4	5	6	1	6	7	8	9	10	10	2	5	11	12	13	12
9	1	1	2	3	4	5	6	1	6	3	7	8	9	9	2	5	10	11	12	11
10	1	1	2	3	4	4	5	1	5	3	6	7	8	8	2	4	9	10	11	10
11	1	1	2	3	4	4	1	1	1	3	5	6	7	7	2	4	8	9	10	9
12	1	1	2	3	4	4	1	1	1	3	5	6	7	7	2	4	2	8	9	8
13	1	1	2	3	4	4	1	1	1	3	5	5	6	6	2	4	2	7	8	7
14	1	1	2	3	4	4	1	1	1	3	5	5	6	6	2	4	2	7	7	7
15	1	1	2	3	4	4	1	1	1	3	5	5	2	2	2	4	2	6	6	6
16	1	1	2	1	3	3	1	1	1	1	4	4	2	2	2	3	2	5	5	5
17	1	1	2	1	1	1	1	1	1	1	3	3	2	2	2	1	2	4	4	4
18	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**1**	**1**	**1**	**1**	**1**	**3**	**3**	**3**
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

6.3 A Monte Carlo Study

In Chapters 4 and 5, a Monte Carlo study clearly illustrated the advantages of the cluster-based method over a non-cluster-based method in Phase I and Phase II applications when the profiles were correctly represented by a parametric model. A Monte Carlo study is used in this chapter to evaluate the average performance of the cluster-based method in monitoring profiles fit not with a parametric model but a nonparametric model using some appropriate nonparametric method. In this Monte Carlo study, the in-control PA profile is a combination of the parametric PA profile in Chapter 4 and an additive component $\gamma\left(10\left(Sin\left(\frac{\pi(x_j-1)}{2.25}\right)\right)\right)$. Thus the in-control profiles are generated from the model

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + \beta_{2i}x_{ij}^{\ 2} + \gamma\left(10\left(Sin\left(\frac{\pi(x_j-1)}{2.25}\right)\right)\right) + \varepsilon_{ij}, i = 1,2,\ldots,m_1,\ j = 1,2,\ldots,n, \tag{6.2}$$

where

$$\beta_{0i} = \beta_2\bar{x}^2 + b_{0i},$$

$$\beta_{1i} = \beta_1 - 2\beta_2\bar{x} + b_{1i},$$

$$\beta_{2i} = \beta_2 + b_{2i}.$$

And the out-of-control profiles are also generated from equation (6.2) but with

$$\beta_{0i} = (\beta_2 + shift)\bar{x}^2 + b_{0i},$$

$$\beta_{1i} = \beta_1 - 2(\beta_2 + shift)\bar{x} + b_{1i},$$

$$\beta_{2i} = (\beta_2 + shift) + b_{2i}.$$

One can show that the PA profile for the in-control process and out-of-control process are

$$y_{PA,j} = \beta_1 x_j + \beta_2\left(x_j - \bar{x}\right)^2 + \gamma\left(10\left(Sin\left(\frac{\pi(x_j-1)}{2.25}\right)\right)\right), j = 1,2,\ldots,n, \tag{6.3}$$

and

$$y_{PA,j} = \beta_1 x_j + (\beta_2 + shift)(x_j - \bar{x})^2 + \gamma\left(10\left(Sin\left(\frac{\pi(x_j - 1)}{2.25}\right)\right)\right), \quad j = 1,2,...,n, \qquad (6.4)$$

respectively. The parameter γ in above equations is called the misspecification parameter and varies from 0 to 4. When γ is 0, the PA curve is exactly the quadratic function and when γ is 4, the PA curve departs considerably from the quadratic model. Values of γ between 0 and 4 represent a continuous departure from the quadratic model. A plot of the PA profiles using different values for $\gamma = 0$, $\gamma = 1$, $\gamma = 2$, $\gamma = 3$ and $\gamma = 4$ is given in Figure 6.2.

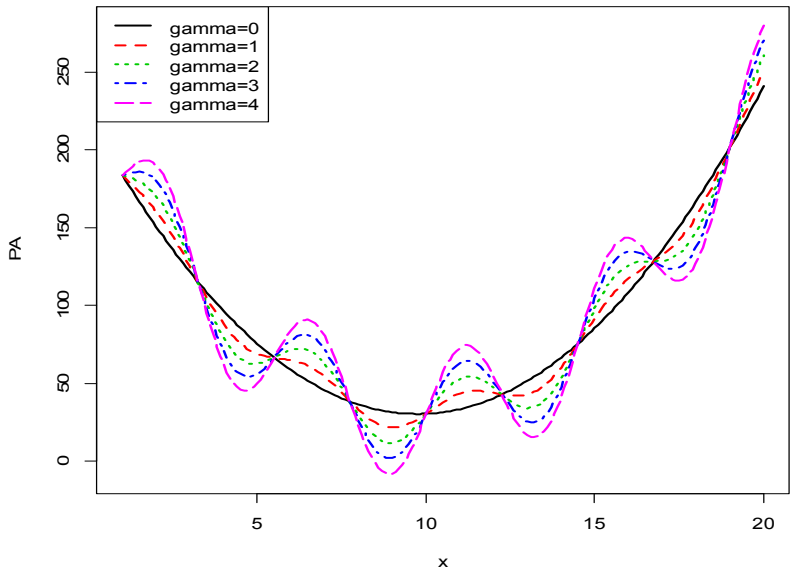

Figure 6.2: Plot of PA profile with different γ values.

All other parameters are similar to the Monte Carlo study in Chapter 4, where

$$\begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{bmatrix} \sim MN\left[\mathbf{0}, \begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{pmatrix}\right],$$

$$\varepsilon \sim N\left[\mathbf{0}, \quad \sigma^2 I\right],$$

$$x_{ij} = j, \; i = 1,2,...,m, \; j = 1,2,...,n.$$

Here, $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 0.5$, $\sigma^2 = 1$ and $\beta_1 = 3, \beta_2 = 2$. It is also assumed that we have $m_1 = 20$ in-control profiles, $m = 30$ profiles in the HDS and $n = 20$.

It is assumed throughout this Monte Carlo study that the user, not knowing the true model given by (6.2), observes a quadratic trend in each profile and assumes that the proper model is a quadratic (model (6.2) with $\gamma = 0$) for each profile. Of course, as γ increases from 0 to 4, the user's model becomes more and more misspecified. This misspecification will greatly hinder proper Phase I analysis, with increasingly poor performance as gamma increases toward 4. On the other hand, a proper nonparametric method due to its ability to adapt to the data should provide a better analysis in Phase I that should remain relatively constant across increasing values of γ. Comparing the Phase I results for the misspecified parametric model to the nonparametric methods should illustrated the difficulty in making proper Phase I decisions with a misspecified model and the advantages of using a nonparametric procedure when the true profile model is not known to the user. The nonparametric method used here is the p-spline. While this information is valuable, these results will completely support similar results appearing in Abdel-Salam et al. (2013). Rather, our emphasis will be on comparing the cluster-based method to the non-cluster-based method used by Abdel-Salam et al. (2013) to demonstrate the advantages provided by the clustering technique.

Consequently, in this Monte Carlo study, a parametric model and a nonparametric model with the cluster-based and non-cluster-based methods will be used to fit the data. Thus, there are four scenarios, a parametric model to fit the data with the cluster-based and non-cluster-based method as well as a nonparametric model with the cluster-based and non-cluster-based method. The shift values in this Monte Carlo study are set to 0.05, 0.1, 0.15, 0.2, 0.25 and 0.3, and the misspecification parameter, γ is set to 0, 2 and 4. The average performance of the four methods for different shift values with different γ is compared by using performance metrics provided in Chapter 4.

Figure 6.3 to Figure 6.5 display the plots of the FCC based on the four methods for different shift values based on the four methods with $\gamma = 0$, $\gamma = 2$ and $\gamma = 4$ respectively. Recall that the FCC measures the proportion of correctly identified in-control and out-of-control profiles and larger values of the FCC are better than smaller values. As expected, all plots show that FCC is increasing as shift value increases.

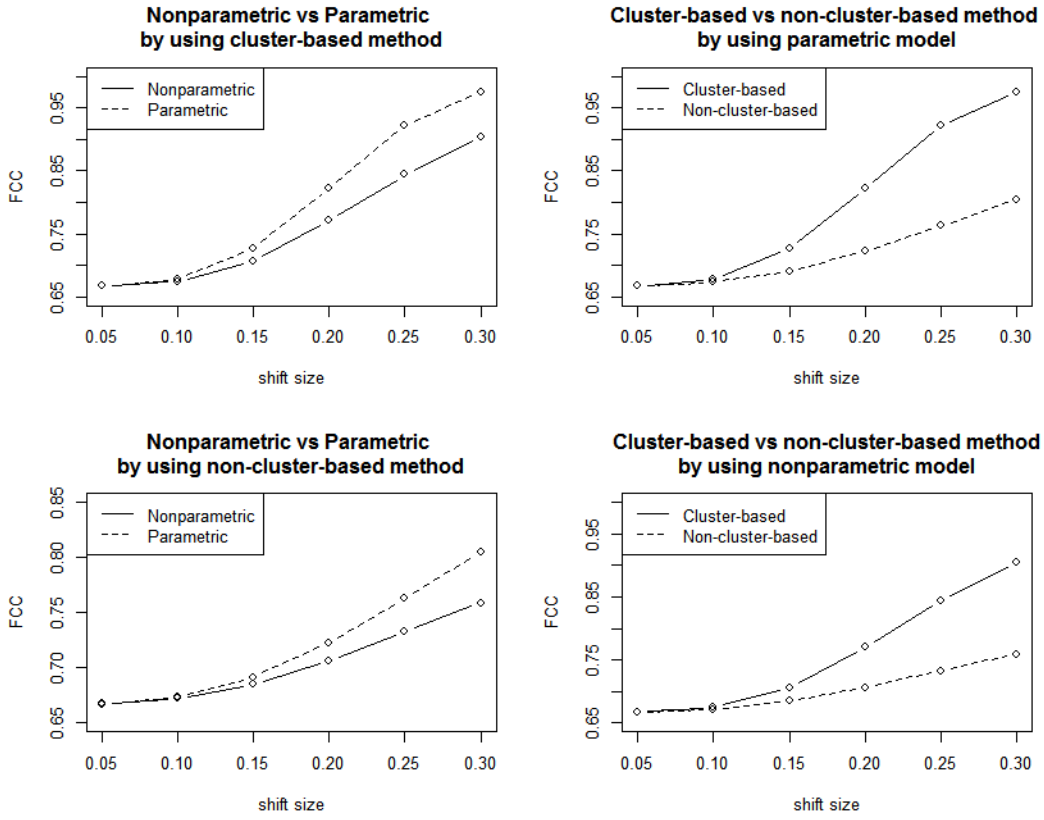

Figure 6.3: FCC for different shift values with γ=0

Figure 6.3 displays the plots of the FCC based on the four methods for different shift values with γ=0. The plots show that when γ equals 0, the FCC based on the parametric methods are superior to the nonparametric methods. From equation (6.2), one can see that when γ equals 0, the true model is a parametric model and as expected, the parametric model will fit the data better than the nonparametric model. As a result, the FCC based on the parametric methods is larger than the FCC based on the nonparametric

methods. Also, Figure 6.3 tells us that the cluster-based method works better than the non-cluster-based method regardless of using which model is used to fit the data.

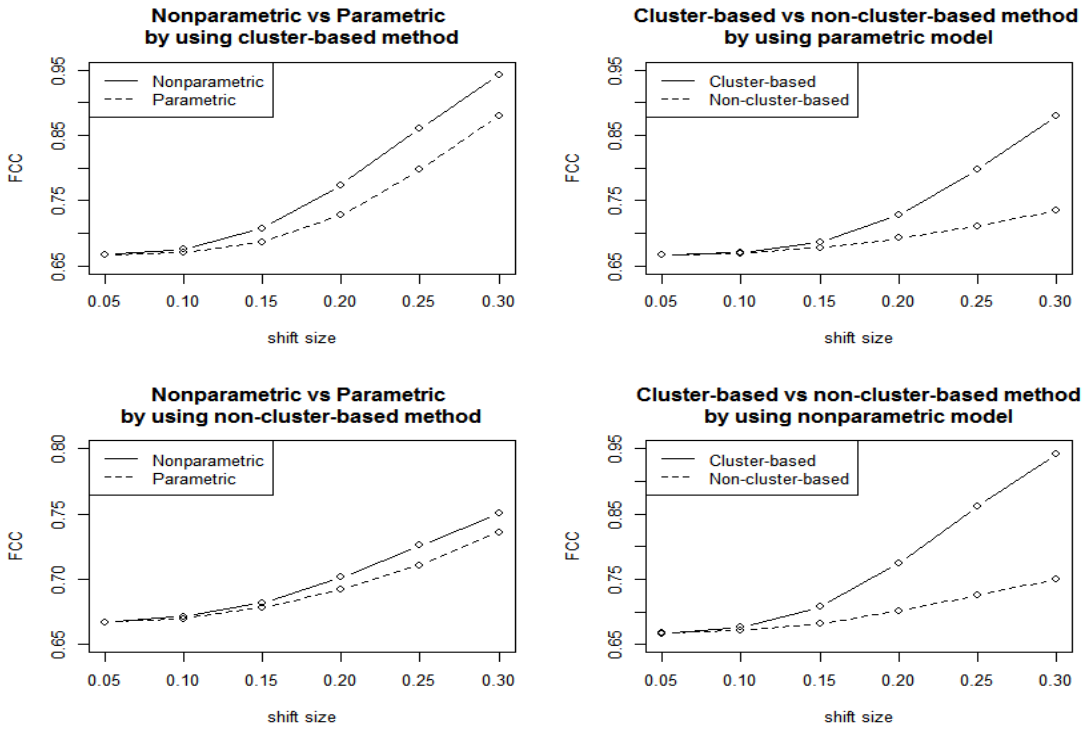

Figure 6.4: FCC for different shift values with γ=2

Figure 6.4 lists the plots of the FCC based on the four methods for different shift values when γ equals 2. In this case, the FCC from the nonparametric model is uniformly larger than the FCC from the parametric model for both cluster-based and non-cluster-based methods. The reason for this result is that when γ equals 2, the user's parametric model is slightly misspecified and is unable to provide as good a fit to each profile as in the $\gamma = 0$ case. The nonparametric model, on the other hand, is robust to the model misspecification and thus has a better performance. Also, the plots imply that the cluster-based method works uniformly better than the non-cluster-based method as the results from Figure 6.3. For example, when shift value equals 0.3, the nonparametric cluster-based method has FCC= 0.9418 while the nonparametric non-cluster-based method has FCC=0.7503. Also, the parametric cluster-based method has FCC=0.8795,

while the parametric non-cluster-based method has FCC= 0.7357. (See Appendix, Table E-H for details).

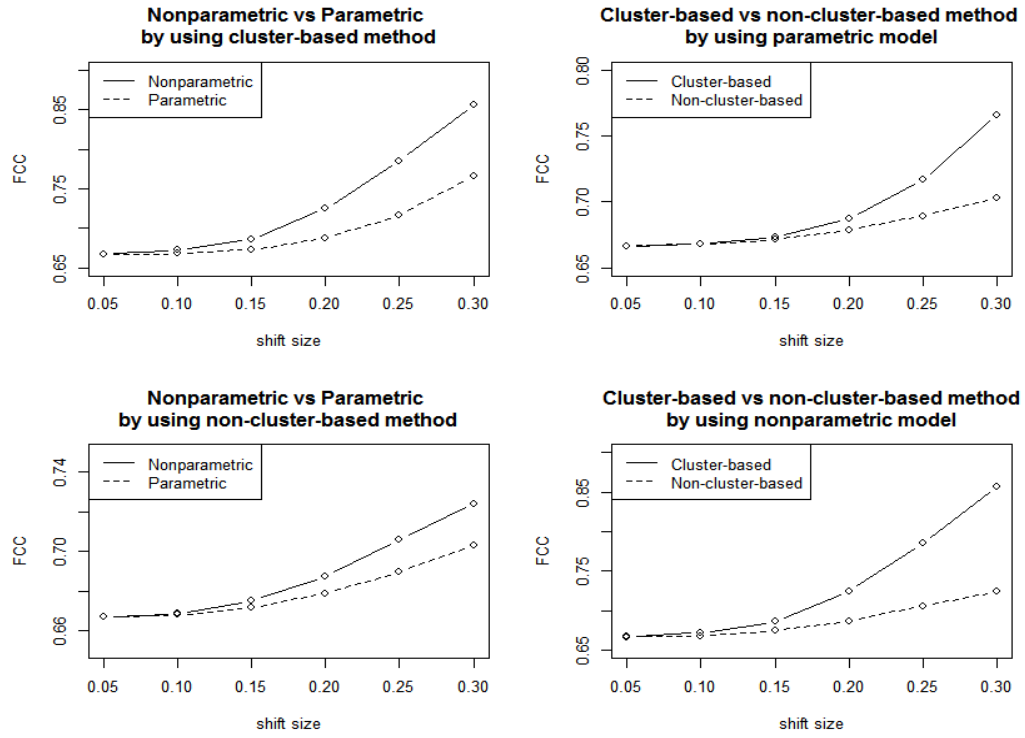

Figure 6.5: FCC for different shift values with γ=4

When γ equals 4, the results for the FCC are similar to the results when γ equals 2. The cluster-based methods are superior to the non-cluster-based methods and the nonparametric models work better than the parametric models. However, the FCC from the fours methods is smaller. For example, when $\gamma = 2$ and shift=0.3, the FCC for the four methods (parametric cluster-based and non-cluster-based methods; nonparametric cluster-based and non-cluster-based methods) are 0.8795, 0.7357, 0.9418 and 0.7503 respectively, while γ equals 4 with shift equals 0.3, the FCC values for the four methods are 0.7661, 0.7031, 0.8566 and 0.7242, respectively. The reason for these results is that as γ increases, the proportion of the nonparametric part increases in the model, but the shift values only affect the parametric part so that the relative shift value in this case is smaller.

Other performance metrics are also obtained in this Monte Carlo study, for example, Figure 6.6 to Figure 6.8 display the plots FPR for the four methods with different shift and γ values.

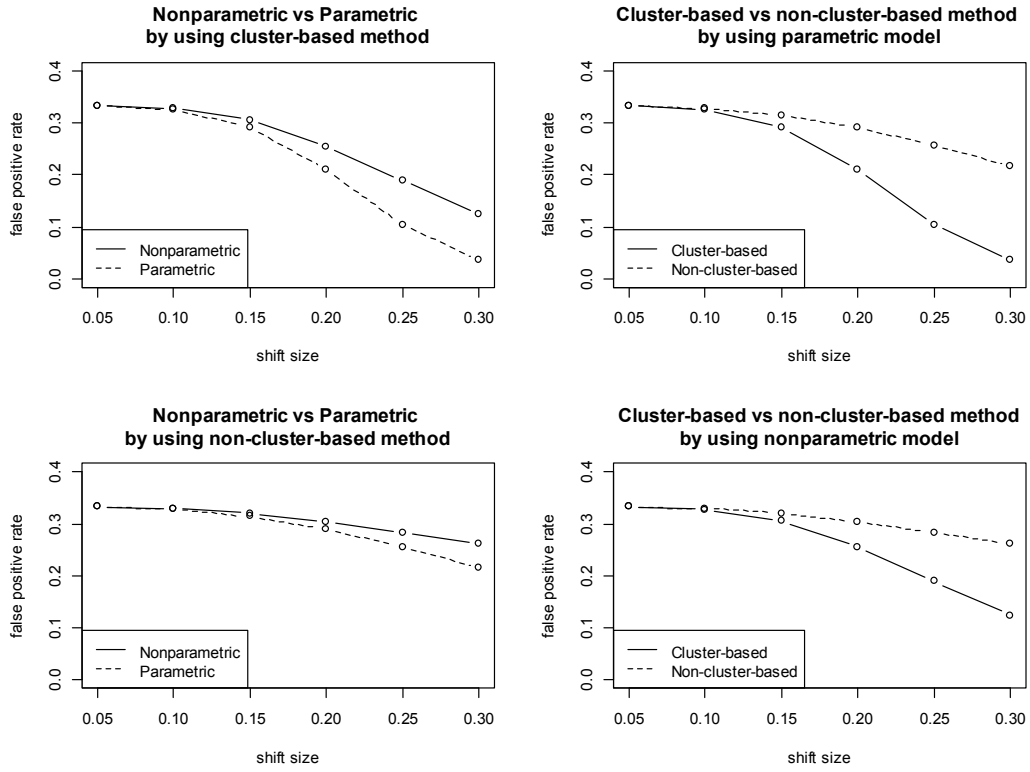

Figure 6.6: FPR for different shift values with γ=0

Figure 6.6 contains the plots of the FPR for different shift values with γ=0 . Not surprisingly, the plots indicate that the FPR is decreasing as shift value increases. Also, similar to the conclusion in Figure 6.3, when the true model is a parametric model, using the parametric model has uniformly better performance (smaller FPR) compare to the nonparametric model. Also, as mentioned in the previous plots, the cluster-based method performs uniformly better than the non-cluster-based method regardless of using the parametric or nonparametric model to fit the data.

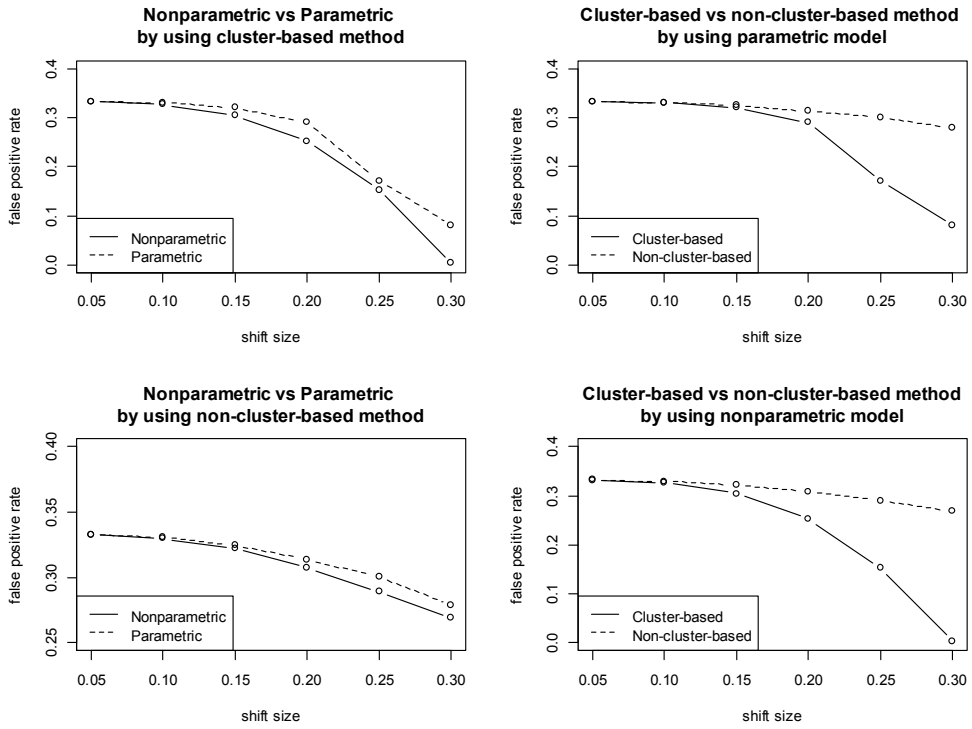

Figure 6.7: FPR for different shift values with γ=2

Figure 6.7 is the plots of the FPR for different shift values when γ=2 . As expected, the parametric models have an inferior performance when compared to the nonparametric models. Also the cluster-based methods have superior performance than the non-cluster-based methods. Further, the plots in Figure 6.6 imply that the difference of the cluster-based and non-cluster-based methods is much bigger than the difference of the nonparametric and parametric methods. For instance, when shift value equals 0.3, the FPR from the four methods (parametric model with cluster-based and non-cluster-based methods; nonparametric model with cluster-based and non-cluster-based methods) are 0.0796, 0.2792, 0.0034 and 0.2691 respectively. One can see that the difference between the parametric and nonparametric approach is trivial, which is 0.0796 versus 0.0034. However, there is a significant difference between the cluster-based methods and the non-cluster-based methods; the cluster-based methods have the FPR close to 0 while the non-cluster-based methods have the FPR greater than 0.2.

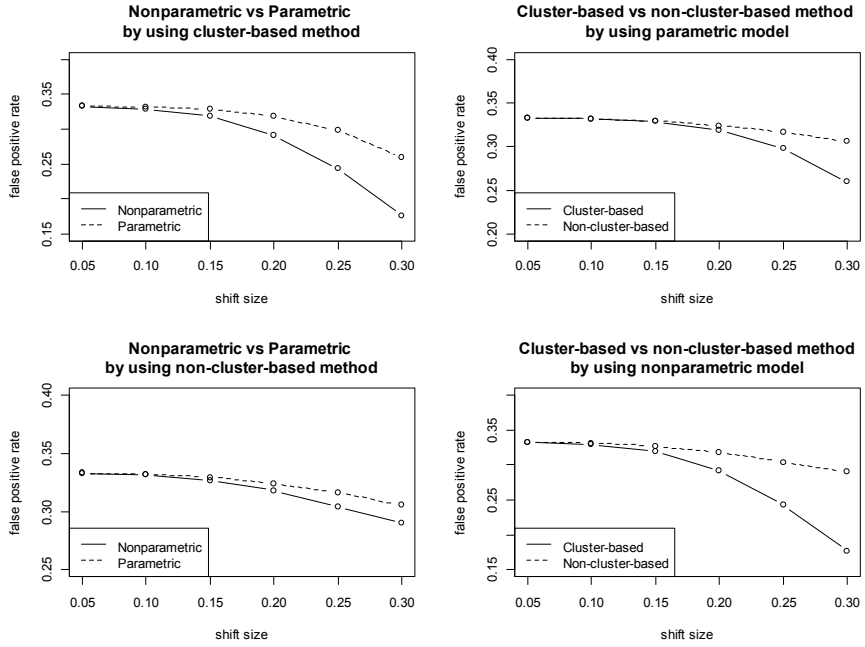

Figure 6.8: FPR for different shift values with γ=4

Figure 6.8 shows the plots of FPR with different values when γ=4 . The conclusion is consistent with the conclusion in Figure 6.4 that as γ increases, the relative shift values are decreasing thus the FPR is increasing. The misspecified parametric models have an inferior performance when compared to the nonparametric models. Also the cluster-based methods have superior performance than the non-cluster-based methods.

The other performance metrics are also obtained by this Monte Carlo study and all the performance metrics are listed in Appendix (Table A to Table L).

6.4 Conclusion

In this Chapter, we applied the cluster-based method to the nonparametric profile monitoring in Phase I analysis. By clustering the profiles, we first fit each profiles based on the p-spline regression model and then cluster the profiles based on the estimated parameters. After clustering the profiles, we updated profiles to the main cluster set by

using the similar manner as the cluster-based method in monitoring the parametric profiles.

The automobile engine example showed the details of the algorithm and the Monte Carlo study provided the average performance of using cluster-based method in nonparametric profile monitoring. The results of the Monte Carlo study indicate that the cluster-based method works uniformly better than the non-cluster-based method. Also, two models are used to fit the profiles in the Monte Carlo study, a parametric model and nonparametric model. The results show that when the true model is the parametric model (that is, when $\gamma = 0$), the parametric method to fit the data is superior to the nonparametric method. However, when the user's parametric model is misspecified, the nonparametric method has superior performance, which agrees with the conclusions from Abdel-Salam (2013). In summary, the greatest gains in Phase I performance are obtained using the cluster-based method as opposed to the non-cluster-based method. And, clearly, if the user's model cannot be correctly specified, a nonparametric model should be considered and an appropriate nonparametric method (such as p-splines) used over a possible misspecified parametric model and incorrect parameter estimates.

Chapter 7. Conclusions and Outlook for Future Work

This chapter summaries the cluster-based method used in robust regression and in monitoring the parametric and nonparametric profiles. Also, an outlook for future study is proposed.

7.1 Conclusions

Profile monitoring is a very popular approach in SPC which assumes that the quality of a product or of a process can be represented by a relationship between a response variable and one or more explanatory variables. In this work, an innovative cluster-based profile monitoring method is proposed in monitoring either parametric or nonparametric profiles.

Before introducing the cluster-based method in profile monitoring, the advantage of using cluster-based methods in robust regression, referred to as *cluster-based bounded influence regression or CBI*, was introduced. It is known that the majority of the previous regression methods either can be easily affect by high influence points (such as OLS) or can be very inefficient (such as LTS). CBI, by using the hierarchical cluster method to cluster the observations, has been demonstrated to be a high breakdown and efficient estimator. CBI regression represents a data space via a special set of anchor points and then obtains a similarity measure of the observations by using a collection of single-point-added ordinary least squares regression estimators. A hierarchical cluster method then yields a main cluster set containing at least half of the total observations and one or more minor sets. An initial regression estimator arises from the main cluster, with a group-additive DFFITS argument used to carefully activate the minor clusters through a bounded influence regression framework. CBI achieves a 50% breakdown

point, is regression equivariant, scale and affine equivariant and asymptotically normally distributed. Both the case studies (PH data and HKB data) and the Monte Carlo study show that this regression methodology is competitive with methods such as LTS (Ruppert and Carroll (1980)), S1S (Coakley and Hettmansperger (1993)) and REWLS (Gervini and Yohai (2002)) when the data is highly contaminated but also be able to compete with the efficient M and BI regression methods (Huber and Ronchetti (2009)) when the data has few or no problematic observations. Specifically, the first case study (PH data) shows that the CBI outperformed the other high breakdown procedures under the low contamination situation. The Monte Carlo study, on the other hand, shows that the CBI is one of the two procedures (S1S and CBI) that provide unbiased regression coefficients. Between the unbiased procedures, the CBI has the smaller standard errors of the regression coefficients and has more stable of the coefficient estimates.

The cluster-based profile monitoring in Phase I incorporated a cluster analysis phase to determine if out-of-control profiles are present in the HDS. The proposed method first replaces the data for each profile with an estimated parameter vector, using some appropriate regression method, and then clusters the profiles based on these estimated parameter vectors. This cluster phase then yields a main cluster which contains at least half of the profiles. The initial estimated PA parameters are obtained by fitting a linear mixed model to those profiles in the main cluster. The in-control profiles, determined using the Hotelling's T^2 statistic, that are not contained in the initial main cluster are iteratively added to the main cluster and the mixed model is used to update the estimated PA parameters. A simulated example, a Monte Carlo study, and the application to the automobile engine data set demonstrates the performance advantage of this proposed method over a current non-cluster-based method with respect to more accurate estimates of the PA parameters and better classification performance in determining those profiles from an in-control process from those from an out-of-control process for both parametric and nonparametric methods for estimating the profiles.

Also, this work showed that when the profiles can be represented by m appropriate $p \times 1$ vectors, the profile monitoring process is equivalent to the detection of multivariate outliers. For this reason, we also compare our proposed method for the parametric modelling of profiles to a popular method used to identify outliers when dealing with a multivariate response. More specifically, the successive difference and the MVE methods for estimating the variance-covariance matrix for the estimated profile model parameters are also used in computing both the cluster-based and non-cluster-based procedures. The successive difference estimator has been recommended for use when the out-of–control process is due to a sustained shift in the profile parameters. The MVE method is commonly suggested for use in detecting multivariate outliers. Our study demonstrated that when the out-of-control process is due to a sustained shift, the cluster-based method using the successive difference estimator is clearly the superior method, among those methods we considered, based on all performance criteria.

Besides, the Phase II ARLs obtained by using the cluster-based T^2 control chart to those obtained by using the non-cluster-based T^2 control chart in Phase II analysis. Both the simple example and the Monte Carlo study showed that the estimates from the Phase I analysis can dramatically affect the performance in Phase II analysis. The cluster-based T^2 control chart can be far more efficient in detecting Phase II shifts than the non-cluster-based T^2 control chart.

In Summary, the cluster-based method has been demonstrated to be superior to the current non-cluster-based method in monitoring parametric or nonparametric profiles. The cluster-based method is more likely to correctly identify the in-control and out-of-control profiles thus will give us more accurate estimates and more efficient Phase II control charts.

7.2 Outlook for Future Work

Our research assumed that the response variable is continuous and from the normal distribution. However, these assumptions will not always be true. For example, the response variable could be counts, a binary variable or a continuous variable from a distribution other than the normal. In these cases, the profile or relationship between the response variable and explanatory variables can be represent by using the generalized linear model, such as a Poisson, logistic or any other appropriate model. For future work, the cluster-based methods could be applied in those situations where the response variable comes from the exponential family and the relationship between the response variable and explanatory variables can be represented by using the generalized linear model.

References

Abdel-salam, A.S. G. (2009), "Profile Monitoring with Fixed and Random Effects Using Nonparametric an Semiparametric Methods". Dissertation. Virginia Polytechnic Institute and State University, 2009.

Abdel-Salam, A. S. G., Birch, J. B., and Jensen, W. A. (2013), "A Semiparametric Mixed Model Approach to Phase I Profile Monitoring," *Quality and Reliability Engineering International*, 29, 555-569.

Alfaro, J. L., and Ortega, J. F. (2008), "A Robust Alternative to Hotelling's T-2 Control Chlart Using Trimmed Estimators," *Quality and Reliability Engineering International*, 24, 601-611.

Amiri, A., Jensen, W. A., and Kazemzadeh, R. B. (2010), "A Case Study on Monitoring Polynomial Profiles in the Automotive Industry," *Quality and Reliability Engineering International*, 26, 509-520.

Berk, R. A. (2008), "Springer Series in Statistics," *Statistical Learning from a Regression Perspective*.

Birch, J. B. (1992), "Estimation and Inference in Multiple Regression Using Robust Weights: A Unified Approach."

Birch, J. B. (2010), "Exploratory and Robust Data Analysis," *Pre-publication Course Packet, Virginia Tech, 2010*.

Birch, J. B., and Agard, D. B. (1993), "Robust Inference in Regression - a Comparative-Study," *Communications in Statistics-Simulation and Computation*, 22, 217-244.

Cali Manning, D., and Adams, B. M. (2005), "Robust Monitoring of Contaminated Data," *Journal of Quality Technology*, 37, 163.

Chenouri, S. e., Steiner, S. H., and Variyath, A. M. (2009), "A Multivariate Robust Control Chart for Individual Observations," *Journal of Quality Technology*, 41, 259-271.

Chicken, E., Pignatiello, J. J. J. R., and Simpson, J. R. (2009), "Statistical Process Monitoring of Nonlinear Profiles Using Wavelets," *Journal of Quality Technology*, 41, 198.

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009), "Asymptotic Properties of Penalized Spline Estimators," *Biometrika*, 96, 529-544.

Coakley, C. W., and Hettmansperger, T. P. (1993), "A Bounded Influence, High Breakdown, Efficient Regression Estimator," *Journal of the American Statistical Association*, 88, 872-880.

Cook, R. D., Hawkins, D. M., and Weisberg, S. (1993), "Exact Iterative Computation of the Robust Multivariate Minimum Volume Ellipsoid Estimator," *Statistics & Probability Letters*, 16, 213-218.

Croarkin, C. (1982), "Measurement Assurance for Dimensional Measurements on Integrated-Circuit Photomasks," *NBS Technical Note 1164, U.S. Department of Commerce, Washington, D.C., USA.*

de Mast, J., and Roes , K. (2004), "Robust Individuals Control Chart for Exploratory Analysis," *Quality Engineering*, 16, 407.

Demidenko, E. (2004), *Mixed Models : Theory and Applications*, Hoboken, N.J.: Wiley-Interscience.

Ding, Y., Zeng, L., and Zhou, S. (2006), "Phase I Analysis for Monitoring Nonlinear Profiles in Manufacturing Processes," *Journal of Quality Technology*, 38, 199-216.

Doruska, P. F. (1998), "Methods for Quantitatively Describing Tree Crown Profiles of Loblolly Pine (Pinus Taeda L.)." Dissertation Virginia Polytechnic Institute and State University, 1998.

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing with B-Splines and Penalties," *Statistical Science*, 11, 89-102.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: M. Dekker.

Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing*, New York: Marcel Dekker.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Boca Raton, Fla.: Chapman & Hall/CRC.

Fan, S. K., Huang, H. K., and Chang, Y. J. (2013), "Robust Multivariate Control Charts for Outlier Detection Using Hierarchical Cluster Tree in Sw2," *Quality and Reliability in Engineering International* 29, 975-981.

Fraker, S. E., Woodall, W. H., and Mousavi, S. (2008), "Performance Metrics for Surveillance Schemes," *Quality Engineering*, 20, 451-464.

Gervini, D., and Yohai, V. J. (2002), "A Class of Robust and Fully Efficient Regression Estimators," *Annals of Statistics*, 30, 583-616.

Gray, R. J. (1994), "Spline-Based Tests in Survival Analysis," *Biometrics*, 50, 640-652.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models : A Roughness Penalty Approach*, London; New York: Chapman & Hall.

Gupta, S., Montgomery, D. C., and Woodall, W. H. (2006), "Performance Evaluation of Two Methods for Online Monitoring of Linear Calibration Profiles," *International Journal of Production Research*, 44, 1927-1942.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics : The Approach Based on Influence Functions*, New York: Wiley.

Härdle, W. K. (1992), *Applied Nonparametric Regression*, Cambridge University Press.

Hawkins, D. M. (1993), "A Feasible Solution Algorithm for the Minimum Volume Ellipsoid Estimator in Multivariate Data," *Computational Statistics 8, pp. 95-107*.

Hawkins, D. M. (1994), "The Feasible Solution Algorithm for Least Trimmed Squares Regression," *Computational Statistics & Data Analysis*, 17, 185-196.

Hawkins, D. M., Bradu, D., and Gordon, V. K. (1984), "Location of Several Outliers in Multiple-Regression Data Using Elemental Sets," *Technometrics*, 26, 197-208.

Hawkins, D. M., and Olive, D. J. (1999), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational statistics & data analysis*, 30, 1-11.

Holmes, D. S., and Mergen, A. E. (1993), "Improving the Performance of the T2 Control Chart," *Quality Engineering*, 5, 619-625.

Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley & Sons.

Huber, P. J., and Ronchetti, E. (2009), *Robust Statistics*, Hoboken, N.J.: Wiley.

Jensen, W. A., and Birch, J. B. (2009), "Profile Monitoring Via Nonlinear Mixed Models," *Journal of Quality Technology*, 41, 18-34.

Jensen, W. A., Birch, J. B., and Woodall, W. H. (2007), "High Breakdown Estimation Methods for Phase I Multivariate Control Charts," *Quality and Reliability Engineering International*, 23, 615-629.

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.

Jeong, M. K. (2006), "Wavelet-Based Spc Procedure for Complicated Functional Data," *International Journal of Production Research*, 44, 729-744.

Jin, J. H., and Shi, J. J. (2001), "Automatic Feature Extraction of Waveform Signals for in-Process Diagnostic Performance Improvement," *Journal of Intelligent Manufacturing*, 12, 257-268.

Jobe, J. M., and Pokojovy, M. (2009), "A Multistep, Cluster-Based Multivariate Chart for Retrospective Monitoring of Individuals," *Journal of Quality Technology*, 41, 323-339.

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

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

Lada, E. K., Jye-Chyi, L., and Wilson, J. R. (2002), "A Wavelet-Based Procedure for Process Fault Detection," *Semiconductor Manufacturing, IEEE Transactions on*, 15, 79-90.

Lawrence D. (2003), "Cluster-Based Bounded Influence Regression," Ph.D. dissertation, Department of Statistics, Virginia Tech.

Lawrence, D., Birch, J. B. and Chen, Y. (2013), "Cluster-Based Bounded Influence Regression," Quality Engineering and Reliability International, Early View.

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

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006), *Robust Statistics : Theory and Methods*, Chichester, England: J. Wiley.

Mason, R. L., and Young, J. C. (2002), *Multivariate Statistical Process Control with Industrial Applications*, Philadelphia, PA; Alexandria, VA: Society for Industrial and Applied Mathematics ; ASA.

Montgomery, D. C. (2009), *Introduction to Statistical Quality Control*, Hoboken, N.J.: Wiley.

Noorossana, R., Saghaei, A., and Amiri, A. (2012), *Statistical Analysis of Profile Monitoring* (Vol. 865), John Wiley & Sons.

O'Sullivan, F., Yandell, B. S., and Raynor, W. J., Jr. (1986), "Automatic Smoothing of Regression Functions in Generalized Linear Models," *Journal of the American Statistical Association*, 81, 96-103.

Pendleton, O. J., and Hocking, R. R. (1981), "Diagnostic Techniques in Multiple Linear Regression Using Proc Matrix," *SUGI*, 195-201.

Pinheiro, J. C., and Bates, D. M. (2000), *Mixed-Effects Models in S and S-Plus*, New York: Springer.

Qiu, P. (2010), "Nonparametric Profile Monitoring by Mixed Effects Modeling," *Technometrics*, 52, 265-277.

Ramsay, J. O., and Silverman, B. W. (2002), *Applied Functional Data Analysis : Methods and Case Studies*, New York: Springer.

Reis, M. S., and Saraiva, P. M. (2006), "Multiscale Statistical Process Control of Paper Surface Profiles," *Quality Technology & Quantitative Management*, 3, 263-282.

Rocke, D. M. (1989), "Robust Control Charts," *Technometrics*, 31, 173-184.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871-880.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York John Wiley & Sons.

Rousseeuw, P. J., and Leroy, A. M. (2003), *Robust Regression and Outlier Detection*, Hoboken, NJ: Wiley-Interscience.

Rousseeuw, P. J., and Leroy, A. M. (2005), "References," in *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc., pp. 292-310.

Rousseeuw, P. J., and Van Driessen, K. (2006), "Computing LTS Regression for Large Data Sets," *Data Mining and Knowledge Discovery*, 12, 29-45.

Ruppert, D., and Carroll, R. J. (1980), "Trimmed Least-Squares Estimation in the Linear-Model," *Journal of the American Statistical Association*, 75, 828-838.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge; New York: Cambridge University Press.

Schabenberger, O., and Pierce, F. J. (2002), *Contemporary Statistical Models for the Plant and Soil Sciences*, Boca Raton: CRC Press.

Shiau, J. J. H., and Sun, J. H. (2010), "A New Strategy for Phase I Analysis in Spc," *Quality and Reliability Engineering International*, 26, 475-486.

Simpson, D. G., Ruppert, D., and Carroll, R. J. (1992), "On One-Step Gm Estimates and Stability of Inferences in Linear-Regression," *Journal of the American Statistical Association*, 87, 439-450.

Staudte, R. G. and Sheather. S. J. (1990), *Robust Estimation and Testing*, New York: John Wiley & Sons.

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

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

Tatum, L. G. (1997), "Robust Estimation of the Process Standard Deviation for Control Charts," *Technometrics*, 39, 127-141.

Vargas, N. J. A. (2003), "Robust Estimation in Multivariate Control Charts for Individual Observations," *Journal of Quality Technology*, 35, 367-376.

Verbeke, G., and Lesaffre, E. (1996), "A Linear Mixed-Effects Model with Heterogeneity in the Random-Effects Population," *Journal of the American Statistical Association*, 91, 217-221.

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

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

Wand, M. P. (2003), "Smoothing and Mixed Models," *Computational Statistics*, 18, 223-249.

Wang, K. B., and Tsung, F. (2005), "Using Profile Monitoring Techniques for a Data-Rich Environment with Huge Sample Size," *Quality and Reliability Engineering International*, 21, 677-688.

Wang, Y. (2011), *Smoothing Splines : Methods and Applications*, Boca Raton, FL: CRC Press.

Wegener, M., and Kauermann, G. (2008), "Examining Heterogeneity in Implied Equity Risk Premium Using Penalized Splines," *Asta-Advances in Statistical Analysis*, 92, 35-56.

Williams, J. D., Birch, J. B., Woodall, W. H., and Ferry, N. M. (2007a), "Statistical Monitoring of Heteroscedastic Dose-Response Profiles from High-Throughput Screening," *Journal of Agricultural Biological and Environmental Statistics*, 12, 216-235.

Williams, J. D., Woodall, W. H., and Birch, J. B. (2007b), "Statistical Monitoring of Nonlinear Product and Process Quality Profiles," *Quality and Reliability Engineering International*, 23, 925-941.

Williams, J. D., Woodall, W. H., Birch, J. B., and Sullivan, J. H. (2006), "Distribution of Hotelling's T-2 Statistic Based on the Successive Differences Estimator," *Journal of Quality Technology*, 38, 217-229.

Woodall, W. H. (2007), "Current Research on Profile Monitoring," *Produção*, 17, 420-425.

Woodall, W. H., Spitzner, D. J., Montgomery, D. C., and Gupta, S. (2004), "Using Control Charts to Monitor Process and Product Quality Profiles," *Journal of Quality Technology*, 36, 309-320.

Wu, H., and Zhang, J.-T. (2006), *Nonparametric Regression Methods for Longitudinal Data Analysis : [Mixed-Effects Modeling Approaches]*, Hoboken, N.J.: Wiley-Interscience.

Yanez, S., Gonzalez, N., and Alberto Vargas, J. (2010), "Hotelling's T(2) Control Charts Based on Robust Estimators," *Dyna-Colombia*, 77, 239-247.

Zhang, J., Li, Z., and Wang, Z. (2009), "Control Chart Based on Likelihood Ratio for Monitoring Linear Profiles," *Computational statistics & data analysis*, 53, 1440-1448.

Appendix

Table A: Average performance metrics based on parametric cluster-based method with $\gamma = 0$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6674	0.3324	0.3922	0.9981	0.0059	0.0864
0.1	0.6782	0.325	0.1016	0.9978	0.0391	0.2876
0.15	0.7268	0.2903	0.0154	0.9986	0.1832	0.6396
0.2	0.8234	0.2091	0.003	0.9993	0.4716	0.8790
0.25	0.9219	0.1044	0.0016	0.9994	0.7670	0.9750
0.3	0.9749	0.0359	0.0011	0.9995	0.9256	0.9956

Table B: Average performance metrics based on parametric non-cluster-based method with $\gamma = 0$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6670	0.3326	0.4429	0.9978	0.0055	0.0904
0.1	0.6731	0.328	0.2409	0.9955	0.0282	0.2812
0.15	0.6913	0.3145	0.1518	0.992	0.0899	0.5854
0.2	0.7227	0.2899	0.1176	0.9871	0.1940	0.8230
0.25	0.7627	0.256	0.0996	0.9821	0.3241	0.9336
0.3	0.8052	0.2163	0.089	0.9775	0.4604	0.9806

Table C: Average performance metrics based on nonparametric cluster-based method with $\gamma = 0$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6673	0.3326	0.3731	0.9985	0.0049	0.0722
0.1	0.6749	0.3274	0.1168	0.9981	0.0284	0.1876
0.15	0.7061	0.3056	0.0258	0.9984	0.1214	0.4258
0.2	0.7712	0.2548	0.0115	0.9982	0.3173	0.6506
0.25	0.8440	0.1891	0.004	0.9989	0.5342	0.8202
0.3	0.9048	0.1242	0.0036	0.9987	0.7169	0.9164

Table D: Average performance metrics based on nonparametric non-cluster-based method with $\gamma = 0$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6673	0.3327	0.3787	0.9988	0.0043	0.0693
0.1	0.6719	0.3292	0.2087	0.9972	0.0212	0.192
0.15	0.6848	0.3199	0.1306	0.9952	0.0644	0.4052
0.2	0.7066	0.3039	0.0885	0.9936	0.1326	0.6368
0.25	0.7327	0.2834	0.0759	0.9911	0.2159	0.7846
0.3	0.7589	0.2619	0.0672	0.9893	0.2981	0.8856

Table E: Average performance metrics based on parametric cluster-based method with $\gamma = 2$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6668	0.333	0.4743	0.9987	0.0029	0.0712
0.1	0.6703	0.3305	0.1933	0.9983	0.0144	0.1622
0.15	0.6861	0.3198	0.0439	0.9986	0.061	0.3460
0.2	0.7275	0.2899	0.008	0.9993	0.1841	0.6180
0.25	0.801	0.2298	0.0017	0.9997	0.4036	0.8296
0.3	0.8795	0.0796	0.0015	0.9995	0.6396	0.9348

Table F: Average performance metrics based on parametric non-cluster-based method with $\gamma = 2$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6667	0.3328	0.5383	0.998	0.0035	0.0765
0.1	0.6697	0.3307	0.2642	0.9975	0.0142	0.1469
0.15	0.6782	0.3244	0.1792	0.9952	0.0442	0.3360
0.2	0.6925	0.3136	0.1466	0.992	0.0935	0.5522
0.25	0.7092	0.3004	0.1303	0.9889	0.1482	0.7071
0.3	0.7357	0.2792	0.1123	0.985	0.2371	0.8213

Table G: Average performance metrics based on nonparametric cluster-based method with $\gamma = 2$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6675	0.3324	0.3531	0.9985	0.0057	0.0653
0.1	0.6759	0.3268	0.0733	0.9988	0.0303	0.1745
0.15	0.7077	0.3045	0.0188	0.9988	0.1256	0.4030
0.2	0.7743	0.2525	0.0072	0.9988	0.3252	0.6571
0.25	0.8612	0.1527	0.0038	0.9989	0.5859	0.8375
0.3	0.9418	0.0034	0.0024	0.999	0.8273	0.9454

Table H: Average performance metrics based on nonparametric non-cluster-based method with $\gamma = 2$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6671	0.3327	0.3939	0.9987	0.0040	0.0774
0.1	0.6714	0.3296	0.1853	0.9979	0.0186	0.1672
0.15	0.6818	0.3221	0.1318	0.9959	0.0536	0.3882
0.2	0.7017	0.3075	0.0957	0.9938	0.1175	0.5840
0.25	0.7259	0.2889	0.0784	0.9917	0.1943	0.7446
0.3	0.7503	0.2691	0.0685	0.9900	0.2710	0.835

Table I: Average performance metrics based on parametric cluster-based method with $\gamma = 4$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6666	0.3332	0.5195	0.9988	0.0022	0.0642
0.1	0.6680	0.332	0.3150	0.9985	0.0067	0.1076
0.15	0.673	0.3287	0.1269	0.9984	0.0223	0.2266
0.2	0.6875	0.3189	0.0388	0.9987	0.065	0.3916
0.25	0.7166	0.298	0.0135	0.999	0.1519	0.6008
0.3	0.7661	0.2595	0.0043	0.9994	0.2996	0.7560

Table J: Average performance metrics based on parametric non-cluster-based method with $\gamma = 4$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6666	0.3331	0.5517	0.9987	0.0021	0.0678
0.1	0.6675	0.3321	0.4194	0.9975	0.0071	0.1161
0.15	0.6716	0.3294	0.2096	0.9973	0.0201	0.1858
0.2	0.6789	0.324	0.1579	0.9958	0.0453	0.3603
0.25	0.6895	0.3162	0.1349	0.9937	0.0812	0.5343
0.3	0.7031	0.3057	0.1214	0.9912	0.127	0.6255

Table K: Average performance metrics based on nonparametric cluster-based method with $\gamma = 4$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6673	0.3326	0.3668	0.9987	0.0044	0.0628
0.1	0.6723	0.3291	0.1625	0.998	0.0209	0.1571
0.15	0.6864	0.3194	0.0614	0.9979	0.0633	0.3538
0.2	0.725	0.2915	0.0181	0.9984	0.1784	0.5692

0.25	0.7855	0.243	0.0061	0.9989	0.3586	0.7441
0.3	0.8566	0.1763	0.0044	0.9987	0.5724	0.9156

Table L: Average performance metrics based on nonparametric non-cluster-based method with $\gamma = 4$

Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6671	0.3329	0.4187	0.9988	0.0034	0.0828
0.1	0.6686	0.3314	0.2938	0.998	0.0098	0.1424
0.15	0.6751	0.3269	0.1736	0.9967	0.0319	0.3442
0.2	0.6874	0.3181	0.1137	0.9954	0.0713	0.5324
0.25	0.7061	0.3041	0.0938	0.9932	0.1318	0.7212
0.3	0.7242	0.2902	0.0786	0.9919	0.1889	0.7888