Nonparametric and Semiparametric Mixed Model Methods for Phase I Profile Monitoring

Abdel-Salam G. Abdel-Salam

Department of Statistics, Oklahoma State University, Stillwater, OK 74078, USA.

Faculty of Economics and Political Science, Cairo University, Egypt.

Jeffrey B. Birch

Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0439, USA.

Willis A. Jensen

W. L. Gore & Associates, Inc., Flagstaff, AZ 86003-2400, USA.

Profile monitoring is an approach in quality control best used where the process data follow a profile (or curve). The majority of previous studies in profile monitoring focused on the parametric modeling of either linear or nonlinear profiles, with both fixed and random-effects, under the assumption of correct model specification. Our work considers those cases where the parametric model for the family of profiles is unknown or, at least uncertain. Consequently, we consider monitoring profiles via two methods, a nonparametric (NP) method and a semiparametric procedure that combines both parametric and NP profile fits. We refer to our semiparametric procedure as mixed model robust profile monitoring ($MMRPM$). Also, we incorporate a mixed model approach to both the parametric and NP model fits to account for the autocorrelation within profiles and to deal with the collection of profiles as a random sample from a common population. For each case, we propose two Hotelling's T^2 statistics for use in Phase I analysis to determine unusual profiles, one based on the estimated random effects and one based on the fitted values and obtain the corresponding control limits.

Our simulation results show that our methods are robust to the common problem of model misspecification of the user's proposed parametric model. We also found that both the NP and the semiparametric methods result in charts with good abilities to detect changes in Phase I data, and in charts with easily calculated control limits. The proposed methods provide greater flexibility and efficiency when compared to parametric methods commonly used in profile monitoring for Phase I that rely on correct model specification, an unrealistic situation in many practical problems in industrial applications. An example using our techniques is also presented.

KEY WORDS: T^2 Control Chart, Industrial Application, Model Robust Regression, Model Misspecification, Model Robust Profile Monitoring, P-Spline, Quality Control.

1 INTRODUCTION

Profile monitoring is an approach in statistic process control (SPC) that is important when the product or process quality is best represented by a profile. In most SPC applications, the quality of a process or product is characterized by univariate or multivariate quality characteristics. However, in some applications the quality of a process or product is characterized by a relationship between a response variable and one or more explanatory variables. Kang and Albin (2000) refer to this relationship as *a profile*. In general, the act of using various techniques to statistically monitor the process or product profiles is known as *profile monitoring* (Woodall, 2007; Woodall et al., 2004).

Woodall et al. (2004) presented a literature review on this subject and introduced a general framework for process monitoring using profile data. Woodall (2007) introduced a general strategy for monitoring more complicated parametric models than the simple linear regression model, including, for example, applications to nonlinear models, and discussed several nonparametric (NP) methods, such as wavelets and splines, to fit each profile. Many applications have appeared in the literature. Our focus is on Phase I applications where the aim is to obtain the estimated in-control-limits by analyzing a historical data set (HDS) to gain understanding of the process variation, to determine the process stability, and to remove any outlying samples when setting control chart limits for Phase II analysis. For more details on Phase I and Phase II analysis see Montgomery (2005).

Much of the previous work on profile monitoring has been based on the assumption that the parametric models for profiles are correctly specified. This is often an unrealistic assumption in practice for many types of applications. For example, in plotting the profile data, the researcher may see features such as peaks, dips or local wiggles that are not captured by a parametric profile of any type, linear or nonlinear. While these features may be unique to a single or a few profiles and thus could be considered as outlying profiles, they can often be features that consistently appear in all the profiles and thus should be captured in the model.

The parametric model may be misspecified in different ways, including wrong effect classification (either fixed or random), and/or wrong model matrices. In these situations, the researcher may still want a "function" to describe the profile, although it may not be a parametric one. An incorrectly specified parametric profile model may be improved by using a NP profile model. Such a NP model may be used in estimating profiles with greatly reduced bias in estimating the true profile than achieved by using an incorrectly specified parametric model.

Not all of the previous work in profile monitoring has been completed using parametric models. Several researchers have relied on NP regression or data-driven methods, such as wavelet thresholding, spline regression and local polynomial regression for monitoring profiles (Zou et al. (2008), Zou et al. (2009) and Qiu and Zou (2009)). Recently, Wei et al. (2010) proposed a nonparametric $L-1$ regression location-scale model to screen the shape of the profiles in Phase II analysis. Their method is robust against heavy-tailed distributions that do not have finite second moments. Qiu et al. (2010) introduced a NP procedure by incorporating local linear kernel smoothing in the exponentially weighted moving average control scheme to perform Phase II profile monitoring. Also, see the comment on this paper by Woodall et al. (2010). The goal of this research is to introduce the mixed penalized spline regression approach as an alternative NP method for profile monitoring. Additionally, we propose a

semiparametric approach, referred to as mixed model robust profile monitoring, or $MMRPM$, that combines a parametric fit with a NP fit, in the profile monitoring context. Our NP and semiparametric approaches for fitting profile data are more flexible than a purely parametric approach. Even when a specific functional form appears reasonable, the NP or semiparametric model provides a more robust model alternative that can be useful in the process of model checking and validation. Additionally, our methods can be used to monitor a broad category of profiles, either linear or nonlinear.

Once the profiles are properly modeled, we conclude the Phase I analysis by utilizing several versions of the Hotelling's T^2 statistic to determine outlying profiles, following the example of Jensen et al. (2008) and Jensen and Birch (2009). In our simulations to come, we will introduce step changes to create the outlying profiles, thus the T^2 statistic based on the successive differences estimator of the variance-covariance matrix (Sullivan and Woodall (1996)) is the most appropriate. We could use other versions of the T^2 statistic to detect outlying profiles, such as those based on the MVE or MCD as was done by Amirhossein et al. (2010) but we will not consider those types of outlying profiles here. Our focus is on an alternative approach to modeling the profiles, thus we will focus less on the determination of which type of T^2 statistic is best. This determination was discussed in the studies of Vargas (2003) and Jensen et al. (2007) and the references therein.

As a practical application of this research, we consider the following example from the automotive industry. One of the most important quality characteristics of the automobile engine is the relationship between the torque produced by an engine and the engine speed in revolutions per minute (RPM). A plot of torque versus RPM forms a profile for each engine. Because the observations within an engine may exhibit serial correlation, we propose the use of mixed models to monitor the engine data in oder to account for the correlation structure. It is reasonable to apply our proposed methods on this data set since the parametric mixed model, as illustrated in Section 6, is not able to capture the main characteristics of each engine profile. We apply our proposed methods for profile monitoring on this dataset in Section 6.

The remainder of this paper is organized as follows. Section 2 gives a review for the parametric linear mixed model approach for profile monitoring. We introduce penalized spline regression as a new NP method to estimate the mixed effects profiles in Section 3. In Section 4, we introduce a mixed model robust profile monitoring ($MMRPM$) method. In addition, we present diagnostic tools to determine outlying profile(s) for our new methods. Section 5 contains a simulation study to compare the parametric, NP, and $MMRPM$ methods. In Section 6, we apply the proposed methods and the parametric quadratic mixed model to the automotive industry data set, and Section 7 summarizes the conclusions.

2 PARAMETRIC LINEAR MIXED MODELS APPROACH

A linear mixed (LM) model is defined as a model that contains at least one fixed effect and at least two random effects, including the error term, and is linear in these effects. Linear mixed models are widely used in many

disciplines. Henderson (1950), for example, one of the first authors proposing use of *LM* models, applied them to estimation problems in the animal sciences. Laird and Ware (1982) generalized Henderson's work, and their formulation is commonly known as the Laird-Ware model.

We assume the true mean response function for the i^{th} profile, $f_i(x_{ij})$, can be given by

$$f_i(x_{ij}) = f(x_{ij}) + \xi_i(x_{ij}) + \epsilon_{ij} \qquad i = 1,2,\ldots,m \qquad j = 1,2,\ldots,n_i \tag{1}$$

where $f(x_{ij})$ represents the mean response function for all profiles, the so-called the population average (*PA*) profile, and $\xi_i(x_{ij})$ represents the random effects for the i^{th} profile, and $\xi_i(x_{ij}) \sim N(0,\sigma^2_\xi)$, where x_{ij} represents the j^{th} observation of the single regressor X in the i^{th} profile and m is the number of profiles. The ϵ_i is a $(n_i \times 1)$ vector of errors, where $\epsilon_i^T = (\epsilon_{i1},\epsilon_{i2},\ldots,\epsilon_{in_i})^T$, assumed to follow a multivariate normal distribution with zero mean vector and variance-covariance matrix R_i, $\epsilon_i \sim MN(0,R_i)$. Model 1 can easily be extended to handle multiple regressors but for simplicity we consider here only the one regressor case, a situation fairly common in profile monitoring.

The user's *LM* model can be the Laird-Ware model, a parametric model, for the i^{th} profile of the form

$$\mathbf{y}_i = X_i\beta + Z_i\mathbf{b}_i + \epsilon_i \qquad i = 1,2,\ldots,m \tag{2}$$

where $\mathbf{y}_i$ represents a response vector for the i^{th} profile, β represents a vector of fixed effects common to all profiles with a known X_i, the model matrix. In addition, Z_i is a $(n_i \times q)$ matrix of the predictor variables with random effects $\mathbf{b}_i \sim MN(0,D)$, a $(q \times 1)$ vector of random effects for the i^{th} profile, and D is a $(q \times q)$ variance-covariance matrix. Often Z_i is typically a matrix whose columns are a subset of the columns of X_i. ϵ_i is distributed as stated in the previous paragraph. Let $N = \sum_{i=1}^m n_i$, $i = 1,2,\ldots,m$ denote the total sample size.

The estimated parameter vector for the i^{th} profile ($\hat{\beta}_i^P$) using the parametric approach is

$$\hat{\beta}_i^P = \hat{\beta} + \hat{\mathbf{b}}_i^* \qquad i = 1,2,\ldots,m \tag{3}$$

where $\hat{\beta}$ represents the *PA* coefficients for all the profiles and $\hat{\mathbf{b}}_i^*$ is a $(p \times 1)$ vector containing the elements of

$$\hat{\mathbf{b}}_i = \hat{D}Z_i^T\hat{V}_i^{-1}(\mathbf{y}_i - X_i\hat{\beta}) \qquad i = 1,2,\ldots,m \tag{4}$$

for the columns of Z_i that are equal to those of X_i and zeros otherwise. Consequently $\hat{\mathbf{b}}_i^* = \hat{\mathbf{b}}_i$ when $Z_i = X_i$. The "P" indicates that the parametric approach is used.

The T^2 statistics, based on the estimated parameter vector, are

$$T^2_{Par,i} = (\hat{\beta}_i^P - \bar{\hat{\beta}}^P)^T\hat{\Sigma}^{-1}(\hat{\beta}_i^P - \bar{\hat{\beta}}^P) \qquad i = 1,2,\ldots,m \tag{5}$$

where $\bar{\hat{\beta}}_i^P = \frac{1}{m}\sum_{i=1}^m \hat{\beta}_i^P$, and $\hat{\Sigma}$ represents the estimated variance-covariance matrix for the estimated parameter vector for the i^{th} profile ($\hat{\beta}_i^P$) using the successive differences, $\hat{\Sigma} = \frac{1}{2(m-1)}\sum_{i=1}^{m-1}(\hat{\beta}_{i+1}^P - \hat{\beta}_i^P)(\hat{\beta}_{i+1}^P - \hat{\beta}_i^P)^T$.

Jensen et al. (2008) modified the above formula (5) for the *LM* model as a function of the $\hat{\mathbf{b}}_i$ for the i^{th} cluster specific (*CS*) curve as follows:

$$T^2_{Par2,i} = (\hat{\mathbf{b}}_i)^T\hat{\Sigma}^{-1}(\hat{\mathbf{b}}_i) \qquad i = 1,2,\ldots,m \tag{6}$$

In addition, we introduce another formula for the Hotelling's T^2, where the T^2 statistics can be evaluated by comparing the parametric fitted values for the i^{th} CS sample profile, and for the population average (PA), where "P" indicates that the parametric approach is utilized, as follows

$$T^2_{Par1,i} = (\hat{\mathbf{y}}^P_{CS,i} - \hat{\mathbf{y}}^P_{PA})^T \hat{V}^{-1} (\hat{\mathbf{y}}^P_{CS,i} - \hat{\mathbf{y}}^P_{PA}) \qquad i = 1,2,\ldots,m \tag{7}$$

where $\hat{\mathbf{y}}^P_{CS,i}$ is the CS curve fit for the i^{th} profile at $l = 1,2,\ldots,n'$, where n' represents the number of $x-$values used for comparison of the m profiles, based on the parametric LM model. As a general rule, n' should be chosen large enough to obtain adequate representation of all m profiles but smaller than the total sample size, N. The PA curve fit is $\hat{\mathbf{y}}^P_{PA}$ for the historical data set (HDS) obtained parametrically and $\hat{V}$ is an $n' \times n'$ appropriate estimated variance-covariance matrix for the $\hat{\mathbf{y}}^P_{CS,i}$, such as $\hat{V}_D$. It can be shown (Abdel-Salam (2009)) that the $T^2_{Par,i}$ statistics based on the estimated random effects (as in Equation (6)) and based on the fitted values (as in Equation (7)) give the same results for the parametric method but different results for the NP and $MMRPM$ methods.

The previous work on parametric profile monitoring has been based on the assumption that the fitted parametric model adequately describes the profile data, which is an unrealistic situation in many practical problems in real life applications.

3 PROFILE MONITORING FOR MIXED MODELS USING P-SPLINE REGRESSION

In this section, we propose the penalized spline (p-spline) regression method for the mixed model case to estimate the m profiles in the profile monitoring context. Also, we introduce diagnostic tools to determine the presence of outlying profile(s).

3.1 P-spline Estimation for Mixed Effects Profiles

The most applicable and flexible models are those that allow for the CS curves to be estimated by NP methods. See Ruppert et al. (2003), and Wegener and Kauermann (2008) for more details. Recall the true mean response model for the i^{th} profile expressed by (1) is

$$y_{ij} = f(x_{ij}) + \xi_i(x_{ij}) + \epsilon_{ij} \qquad i = 1,2,\ldots,m \qquad j = 1,2,\ldots,n_i \tag{8}$$

where y_{ij} is a response variable for the j^{th} observation on the i^{th} profile, with $\epsilon_{ij} \sim N(0,\sigma^2_\epsilon)$, $f(x_{ij})$ represents the overall PA profile and $\xi_i(x_{ij})$ is a smooth function for the i^{th} profile, representing the random difference between the i^{th} CS curve and the PA curve. Both $f(x_{ij})$ and $\xi_i(x_{ij})$ can be approximated by a p-spline regression.

For example, the polynomial basis of order p can be used to approximate $f(x_{ij})$ and $\xi_i(x_{ij})$, though other basis can be utilized as well, as

$$f(x_{ij}) \approx \beta_0 + \sum_{l=1}^{p} \beta_l x^l_{ij} + \sum_{k=1}^{K_1} u_k (x_{ij} - \kappa_k)^p_+ \qquad i = 1,2,\ldots,m \qquad j = 1,2,\ldots,n_i \tag{9}$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are the non-penalized coefficients for the p^{th} order polynomial. The $u_{p1}, u_{p2}, \ldots, u_{pK}$ are the coefficients for the truncated polynomial basis, the spline portion of the model where $\sum_{k=1}^{K_1} u_k^2 \leq \ constant$. K_1 is the number of knots used in the spline portion of the PA curve. The $(x)_+^p$ are known as truncated power functions. See Ruppert et al. (2003) for thorough discussion on the truncated power basis and other basis functions. A convenient equivalent numerical representation of the p-spline in (9) is to consider (9) as a LM model where the n_i vectors $\mathbf{x_{ij}^T} = (1 \ x_{ij} \ x_{ij}^2 \ldots x_{ij}^p)$ are the n_i rows of X_i in (9), $\beta^T = (\beta_0 \ \beta_1 \ \ldots \beta_p)$, the n_i vectors $\mathbf{z_{ij}^T} = ((x_{ij} - \kappa_1)_+^p, \ldots, (x_{ij} - \kappa_{K_1})_+^p)$ are the n_i rows of Z_i, and $\mathbf{u}^T = (u_1 \ u_2 \ \ldots u_{K_1})$ where $\mathbf{u} \sim MN(0, \sigma_u^2 I)$. This representation allows any p-spline model to be computed efficiently using a basic LM model program contained on most available statistical software packages. Using this equivalent representation, the p-spline for $\xi_i(x_{ij})$ is

$$\xi_i(x_{ij}) \approx b_{i0} + \sum_{l=1}^{p} b_{il} x_{ij}^l + \sum_{k=1}^{K_2} t_{ik}(x_{ij} - \kappa_k)_+^p \qquad i = 1, 2, \ldots, m \qquad j = 1, 2, \ldots, n_i \tag{10}$$

where p is the order of the polynomial basis with $(b_{i0}, b_{i1}, \ldots, b_{ip})^T \sim MN(0, \sum_b)$, and $t_{ik} \sim N(0, \ \sigma_t^2 I)$, for i^{th} profile, and K_2 is the number of knots for the CS curves. In model (8), each CS profile has four parts: a PA component ($f(x_{ij})$) and the "difference between the i^{th} CS curve and the PA curve" component. Each of these components can be approximated by p-spline regression curves consisting of a parametric component and a spline component. The random "difference" component can be approximated by (10), where in model (10), the i^{th} profile has two parts: the random parametric component is $b_{i0} + \sum_{l=1}^{p} b_{il} x_{ij}^l$, and the random spline component is $\sum_{k=1}^{K_2} t_{ik}(x_{ij} - \kappa_k)_+^p$.

Utilizing the relationship between mixed models and p-spline regression, the approximation to model (8) can be described succinctly in the mixed model framework for the i^{th} profile as

$$\mathbf{y}_{CS,i} = X_i \beta + Z_i \mathbf{u} + X_i \mathbf{b}_i + E_i \mathbf{t}_i + \epsilon_i \qquad i = 1, 2, \ldots, m \tag{11}$$

where $\mathbf{b}_i = [b_{i0}, b_{i1}, \ldots, b_{ip}]^T$, and $\mathbf{t}_i = [t_{i1}, t_{i2}, \ldots, t_{iK_2}]^T$. It is not necessary for K_1, the number of knots for the PA profile, to be equal to K_2. Note, our method does not require equal number or location of knots across each of the CS curves. However, for computational convenience both the number and location of knots are chosen to be the same in this paper. In addition, the n_i vectors $\mathbf{e}_{ij} = ((x_{ij} - \kappa_1^i)_+^p, \ldots, (x_{ij} - \kappa_{K_2}^i)_+^p)$ are the n_i row of the $(n_i \times \kappa_{K_2})$ random effects design matrix is given by E_i. Model (11) also can be written in a stacked matrix notation as follows

$$\mathbf{y} = X\beta + Z\mathbf{B} + \epsilon \tag{12}$$

where X and β are as defined in Section 2 and

$$Z = \begin{bmatrix} Z_1 \ X_1 \ 0 \ \ldots \ 0 \ E_1 & \ldots \ldots 0 \\ Z_2 \ 0 \ X_2 \ldots \ 0 \ 0 \ E_2 & \ldots \ 0 \\ \vdots \ \vdots \ \ldots\ldots & \vdots \ \vdots \ldots\ldots\ldots\vdots \\ Z_m \ 0 \ 0 \ \ldots \ X_m \ldots\ldots\ldots & E_m \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{u} \\ \mathbf{b} \\ \mathbf{t} \end{bmatrix},$$

with $\mathbf{u} = [u_1, u_2, \ldots, u_{K_1}]^T$, $\mathbf{b} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m]^T$, $\mathbf{t} = [\mathbf{t}_{1p}, \mathbf{t}_{2p}, \ldots, \mathbf{t}_{mp}]^T$, and

$Cov(\mathbf{B}) \equiv G = \begin{bmatrix} \sigma_u^2 I_{K_1 \times K_1} & 0 & 0 \\ 0 & \sigma_b^2 I_{m \times m} & 0 \\ 0 & 0 & \sigma_t^2 I_{K_2 \times K_2} \end{bmatrix}$ where σ_u^2 controls the amount of smoothing to estimate $f(x_{ij})$, σ_b^2 measures the between profiles variation, σ_ϵ^2 measures the within profile variation, and σ_t^2 controls the amount of smoothing required to estimate $\xi_i(x_{ij})$. Now, $\hat{\beta}$ can be obtained by using the formulas in (3) and $\hat{B}$ as in (13). For example,

$$\hat{\mathbf{B}} = [\hat{\mathbf{u}}, \hat{\mathbf{b}}, \hat{\mathbf{t}}]^T = GZ^T V^{-1}(\mathbf{y} - X\hat{\beta}) \tag{13}$$

with $V = ZGZ^T + \sigma_\epsilon^2 I$ with G as given above.

Then the estimated PA curve using p-spline regression is given by

$$\hat{\mathbf{y}}_{PA}^{PS} = X_i\hat{\beta} + Z_i\hat{\mathbf{u}} \tag{14}$$

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

$$\hat{\mathbf{y}}_{CS,i}^{PS} = X_i\hat{\beta} + Z_i\hat{\mathbf{u}} + X_i\hat{\mathbf{b}}_i + E_i\hat{\mathbf{t}}_i \qquad i = 1, 2, \ldots, m \tag{15}$$

where $\hat{\mathbf{y}}_{CS,i}^{PS}$ is the p-spline regression estimator for the i^{th} profile response.

3.2 Determine Outlying Profile(s)

For the mixed effect p-spline regression approach we introduce two methods for determining abnormal profiles using the T^2 statistic. In the first method the T^2 statistics are based on the fitted profile values as

$$T_{NP1,i}^2 = (\hat{\mathbf{y}}_{CS,i}^{PS} - \hat{\mathbf{y}}_{PA}^{PS})^T \hat{V}^{-1}(\hat{\mathbf{y}}_{CS,i}^{PS} - \hat{\mathbf{y}}_{PA}^{PS}) \qquad i = 1, 2, \ldots, m \tag{16}$$

where $\hat{\mathbf{y}}_{CS,i}^{PS}$ is the CS curve fit for the i^{th} profile obtained at n' evenly spaced regressor locations, beginning with the minimum value of the regressor and ending at the maximum value of the regressor. The "PS" means that, all fits were obtained using the mixed p-spline regression model as given in Equation (15). The nonparametric PA curve fit is $\hat{\mathbf{y}}_{PA}^{PS}$ for all the HDS as given in Equation (14) and $\hat{V}$ is an $(n' \times n')$ appropriately estimated variance-covariance matrix for the $\hat{\mathbf{y}}_{CS,i}^{PS}$, such as $\hat{V}_D$.

In the second method, we use the $\hat{\phi}_i$ vectors to calculate the T^2 statistics, where

$$\hat{\phi}_i = [\hat{\mathbf{b}}_i, \hat{\mathbf{t}}_i]^T \qquad i = 1, 2, \ldots, m \tag{17}$$

with $\hat{\mathbf{b}}_i = [\hat{b}_{i0}, \hat{b}_{i1}, \ldots, \hat{b}_{ip}]^T$ and $\hat{\mathbf{t}}_i = [\hat{t}_{i1}, \hat{t}_{i2}, \ldots, \hat{t}_{iK_2}]^T$. If the locations of the regressor values and the number of observations at each location are the same across all m profiles (a common occurrence for profile data obtained from designed experiments), then the T^2 statistic can be based on the estimated predicted random effects, as these represent the possible differences in the m profiles. The T^2 statistics will be denoted by $T_{NP2,i}^2$ and given by

$$T_{NP2,i}^2 = (\hat{\phi}_i - \bar{\phi})^T [\frac{\sum_{i=1}^{m-1}(\hat{\phi}_{i+1} - \hat{\phi}_i)(\hat{\phi}_{i+1} - \hat{\phi}_i)^T}{2(m-1)}]^{-1}(\hat{\phi}_i - \bar{\phi}) \qquad i = 1, 2, \ldots, m \tag{18}$$

where $\bar{\phi} = \frac{\sum_{i=1}^{m} \hat{\phi}_i}{m}$ and $\hat{V}_D$ is the estimated variance-covariance matrix. The unusual profile(s) can be determined by comparing $T^2_{NP1,i}$ or $T^2_{NP2,i}$ with a value from χ^2-distribution where the profile will be marked as outlying if its corresponding T^2 statistic value is large. That is, $T^2_{NP,i} \geq \chi^2_{(df,\alpha)}$ for $i = 1, 2, \ldots, m$, where α represents the significance level, df represents the appropriate degrees of freedom, an issue discussed in Section 5.

Despite the possibility of less bias for the estimated mean response resulting from using the NP methods for mixed models, there is a strong possibility of greater variance in estimating mean response than obtained by a correctly specified parametric model. Mixed model robust regression ($MMRR$), a hybrid combination of the fits obtained by parametric LM and NP mixed models, has been shown to minimize the integrated mean square error of fits when compared to the parametric LM and NP mixed method, while retaining important features of the data (Waterman et al., 2007). For these reasons, we extended the $MMRR$ procedure to the area of profile monitoring in Phase I analysis as discussed in the following section. This second approach is named "mixed model robust profile monitoring ($MMRPM$)" and is obtained by incorporating a mixed model approach to both the parametric and the NP model fits.

4 MIXED MODEL ROBUST PROFILE MONITORING METHOD

Mixed model robust profile monitoring ($MMRPM$) can be considered as an extension of model robust regression (MRR), see Mays et al. (2001), to quality control applications. The development of MMR is motivated by the need to improve upon the shortcomings of the parametric method that may result in high bias due to model misspecification. In addition, the NP method may introduce high variance of fit possibly resulting from an estimated curve that fits the data too closely. MRR has been successfully applied to a broad variety of models and situations including linear normal-theory based regression (Mays et al. (2001)), logistic regression (Nottingham and Birch (2000)), simultaneous modeling of the mean and variance functions (Robinson et al. (2010)), (Pickle et al. (2008)), the multi-response optimization problem (Wan and Birch (2010)), and the LM model (Waterman et al. (2007)).

In $MMRPM$, two separate fits are combined to get the final fits for the PA and the m profiles. In $MMRPM$, we assume that the user has some information about the underlying profiles from which data have been generated and that a parametric model can be formed that provides a reasonable fit to certain portions of the data but fails to adequately fit the data in other parts. That is, the parametric model has been misspecified. Relying on an NP profile entirely results in loss of information about the profile and possibly subjects the results to highly variable fits. Specifically, the $MMRPM$ fit for the PA profile is

$$\hat{\mathbf{y}}^{MMRPM}_{PA} = (1 - \hat{\lambda})\hat{\mathbf{y}}^{P}_{PA} + \hat{\lambda}\hat{\mathbf{y}}^{NP}_{PA} \tag{19}$$

where $\hat{\mathbf{y}}^{P}_{PA}$ is the PA fit from the parametric LM model using an appropriate estimation method. The $\hat{\mathbf{y}}^{NP}_{PA}$ is an NP mixed model fit for the PA profile using the mixed p-spline regression method, as described in Section 3.1 and $\hat{\lambda}$ is the estimated mixing parameter, a number between zero and one. From (19), one can observe that $\hat{\lambda}$ should be

zero for the correctly specified model where $MMRPM$ will reduce to the parametric model. The value of $\hat{\lambda}$ should be small if the parametric model is not badly misspecified and be large for a more misspecified model. $\hat{\lambda}$ should be close to one for a badly misspecified model. In this case, $MMRPM$ will reduce to the NP model. For most practical situations, $\hat{\lambda}$ is an unknown and must be chosen as a function of the data.

The $MMRPM$ fit for the i^{th} CS profile is

$$\hat{\mathbf{y}}_{CS,i}^{MMRPM} = (1-\hat{\lambda})\hat{\mathbf{y}}_{CS,i}^{P} + \hat{\lambda}\hat{\mathbf{y}}_{CS,i}^{NP} \qquad i = 1,2,\ldots,m \tag{20}$$

where $\hat{\mathbf{y}}_{CS,i}^{P}$ is the CS fit from the parametric LM model for the i^{th} profile, and $\hat{\mathbf{y}}_{CS,i}^{NP}$ is an NP mixed model fit for the i^{th} profile. The estimated mixing parameter ($\hat{\lambda}$), as shown in Waterman et al. (2007), can be obtained for the CS fit as

$$\hat{\lambda} = \frac{(\hat{\mathbf{y}}_{i,-i}^{NP} - \hat{\mathbf{y}}_{i,-i}^{P})^T(\mathbf{y} - \hat{\mathbf{y}}_{CS}^{P})}{(\hat{\mathbf{y}}_{CS}^{NP} - \hat{\mathbf{y}}_{CS}^{P})^T(\hat{\mathbf{y}}_{CS}^{NP} - \hat{\mathbf{y}}_{CS}^{P})}, \tag{21}$$

where $\hat{\mathbf{y}}_{i,-i}^{NP}$ and $\hat{\mathbf{y}}_{i,-i}^{P}$ indicate the NP and parametric fits , respectively, for the i^{th} profile without the i^{th} profile. Waterman et al. (2007) showed through a series of Monte-Carlo (MC) simulations that when (21) is used as the estimator of λ, a LM model specified correctly results in a MRR estimate equal to or nearly equal to the parametric fit. If the LM model is badly misspecified, the MRR estimate is the same or nearly equal to the NP fit. The simulation results also showed that the MRR fit is superior to a separate parametric LM model fit or a separate NP model fit under moderate model misspecification in terms of the minimum mean square error of fit criteria. Consequently, we expect the MRR technique to be successful in improving results obtained by a misspecified parametric model when applied to the profile monitoring situation. Our speculation is justified by the MC results shown in Section 5.

The $MMRPM$ T^2 statistic for the fitted semiparametric LM models, using the estimated PA profile ($\hat{\mathbf{y}}_{PA}^{MMRPM}$), the estimated CS profiles ($\hat{\mathbf{y}}_{CS,i}^{MMRPM}$), and the estimated variance-covariance matrix ($\hat{V}$) for the CS profiles, is denoted by $T_{MMRPM1,i}^2$. The $MMRPM$ T^2 statistics are given by

$$T_{MMRPM1,i}^2 = (\hat{\mathbf{y}}_{CS,i}^{MMRPM} - \hat{\mathbf{y}}_{PA}^{MMRPM})^T \hat{V}^{-1} (\hat{\mathbf{y}}_{CS,i}^{MMRPM} - \hat{\mathbf{y}}_{PA}^{MMRPM}) \qquad i = 1,2,\ldots,m \tag{22}$$

where ($\hat{V}$) can be replaced by ($\hat{V}_D$) for the CS fits. The unusual profile(s) can be determined by comparing $T_{MMRPM1,i}^2$ with the appropriate value from the χ^2-distribution. A profile will be marked as outlying if $T_{MMRPM1,i}^2 \geq \chi_{(df_2,\alpha)}^2$ for $i = 1,2,\ldots,m$, where df_2 represents the degrees of freedom associate with the $MMRPM$ fits which is the difference between estimated number of model parameters in the CS_i and PA models. The calculation of df_2 is discussed in Section 5.

Another version for the $MMRPM$ T^2 statistics is based on a convex combination of the estimated random effects from the parametric and NP approaches. Let $\hat{\psi}_i$ refers to the convex combination of the estimated random effects via the estimated mixing parameter ($\hat{\lambda}$) where

$$\hat{\psi}_i = [(1-\hat{\lambda})\hat{\mathbf{b}}_i, \hat{\lambda}\hat{\phi}_i]^T \qquad i = 1,2,\ldots,m \tag{23}$$

with $\hat{\mathbf{b}}_i$ is the estimated random effects obtained parametrically as in Equation (4) and $\hat{\phi}_i$ represents the estimated random effects obtained via the mixed p-spline as given in Equation (17).

The $MMRPM$ T^2 statistics can be obtained based on $\hat{\psi}_i$ as these give the possible differences in the m profiles. We refer to as $T^2_{MMRPM2,i}$ which is given as

$$T^2_{MMRPM2,i} = (\hat{\psi}_i - \bar{\psi})^T [\frac{\sum_{i=1}^{m-1}(\hat{\psi}_{i+1} - \hat{\psi}_i)(\hat{\psi}_{i+1} - \hat{\psi}_i)^T}{2(m-1)}]^{-1}(\hat{\psi}_i - \bar{\psi}) \qquad i = 1,2,\ldots,m \tag{24}$$

where $\bar{\psi} = \frac{\sum_{i=1}^{m}\hat{\psi}_i}{m}$. The abnormal profile(s) can be determined by comparing $T^2_{MMRPM2,i}$ with a value from χ^2-distribution where the profile will be marked as outlying if $T^2_{MMRPM2,i} \geq ((1-\hat{\lambda}) * UCL_{Par} + \hat{\lambda} * UCL_{NP})$ for $i = 1,2,\ldots,m$, where UCL_{Par} represents the upper control limit for the parametric method, and UCL_{NP} represents the upper control limit for the NP method. The estimated mixing parameter is $\hat{\lambda}$ as in Equation (21).

Section 5 gives the results of a Monte-Carlo simulation study to compare the simulated integrated mean square ($SIMSE$) and the simulated probability of signal of the parametric, NP and semiparametric approaches.

5 A MONTE-CARLO STUDY

Monte Carlo simulation methods are employed to generate a specific number of correlated and uncorrelated data sets and calculate the desired criteria. In the examples studied here, the model is similar to that used in Waterman et al. (2007) with an extra random effect (b_{i3}) to give more variability for each CS profile. The data generated from the cluster specific (CS) model is

$$y_{ij} = (5 + b_{i1})x_{ij} + (2 + b_{i2})(x_{ij} - 5.5)^2 + \gamma[10\sin(\frac{\pi(x_{ij}-1)}{2.25}) + b_{i3}] + \epsilon_{ij}$$
$$i = 1,2,\ldots,m \qquad j = 1,2,\ldots,n \tag{25}$$

where y_{ij} is the simulated response for the j^{th} observation from the i^{th} profile at x_{ij}. The single regressor X takes on integer values from one to ten, inclusive. The random effects are b_{i1}, b_{i2} and b_{i3}, which are generated independently from the normal distribution with mean zero and variance 0.5. The random errors ϵ_{ij} are assumed to follow a normal distribution with mean zero and variance-covariance matrix (R).The variance-covariance matrix R will take on several forms. For the uncorrelated data case, R will equal $\sigma^2 I$, where σ^2 is the error variance and I is the identity matrix. For the correlated data case, R will incorporate the form of the first-order autocorrelation matrix, $\sigma^2 AR(1)$, with $corr(\epsilon_{ij}, \epsilon_{ij+1}) = \rho$.

The population average (PA) model can be expressed as

$$y_{ij} = 5x_{ij} + 2(x_{ij} - 5.5)^2 + \gamma[10\sin(\frac{\pi(x_{ij}-1)}{2.25})] + \epsilon_{ij} \tag{26}$$

However, the user will assume that the data are generated from the quadratic model

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})x_{ij} + (\beta_2 + b_{i2})x^2_{ij} + \epsilon_{ij} \tag{27}$$

The random effects are b_{i0}, b_{i1} and b_{i2}, are generated independently from the normal distribution with mean zero and variance 0.5. The true model is the one given in Equation (25), where the trigonometric component times γ, the misspecification parameter, serves as the amount of model misspecification. The user's model equals the true model at γ equals zero, the case where there is no model misspecification. Otherwise, the model has been misspecified, and the degree of misspecification increases with γ.

The parametric, nonparametric (NP), and semiparametric approaches were utilized to fit the CS profiles. The parametric approach used a second order polynomial mixed model. The NP approach used the mixed p-spline regression model as described in Section 3. The semiparametric approach utilized our $MMRPM$ method, combining the parametric and NP models via a mixing parameter λ, as described in Section 4. The SAS code for this simulation study has been written by the authors and available upon request.

To keep the number of scenarios reasonably manageable, the study considers different values for the number of profiles ($m = 30, 60$), profile size ($n = 10, 20$), $\rho = (0, 0.2, 0.8)$, and varying degrees of misspecification ($\gamma = 0, 0.25, 0.5, 0.75, 1$). For the $n = 10$ and $n = 20$ cases, the design points are selected as equally space along the interval from 1 to 10.

For comparing our methods with the parametric method, the simulated integrated mean square ($SIMSE$) and the simulated probability of signal have been used. The $SIMSE$ was calculated as

$$SIMSE = \frac{1}{m}(\hat{\mathbf{Y}}_{CS,i} - \mu_{CS,i})^T(\hat{\mathbf{Y}}_{CS,i} - \mu_{CS,i}) \tag{28}$$

at 46 equally spaced design points (values 1 to ten by 0.2), as an arbitrary number of points picked large enough to detect all differences in curves between the minimum and the maximum values for X. The $SIMSE$ quantifies the amount by which an estimator ($\hat{\mathbf{Y}}_{CS,i}$), for the i^{th} profile, differs from the true mean response value of the quantity being estimated ($\mu_{CS,i}$). The $SIMSE$ values were calculated for the CS using the parametric, NP, and $MMRPM$ approaches.

The probability of signal is calculated as the proportion of simulated datasets where a signal occurred. That is, a signal is given when at least one of the mT^2 statistics exceeds the control limit. For the parametric approach, we consider the quadratic polynomial regression model with two random effects, one for linear and one for the quadratic term. Notice that, the true model, as in Equation (25), contains three random effects, but when the user's model includes all three random effects, the third random effect can not be estimated for the data situations considered in our simulation. For the NP approach, the mixed linear p-spline regression model using the first order radial basis function (Ruppert et al. (2003)), with one random effect on the slope is considered.

We consider various amounts of misspecification ranging from γ equal to 0 (no misspecification), 0.25, 0.5, 0.75 and 1. A plot of the PA models using different values for γ is given in Figure 1.

In Figure 1, the sold black line ($\gamma = 0$) occurs when the user's model equals the true model. The dashed gray curve represents the most misspecification in the PA, when $\gamma = 1$. The larger disparity between the $\gamma = 0$ and $\gamma = 1$ models should be reflected in the $SIMSE$ and simulated probability of signal from the simulation study.

In this simulation, uncorrelated and correlated data sets are generated from different scenarios. The parameter estimation techniques, the probability of signals for each method requires substantial computing resources. In

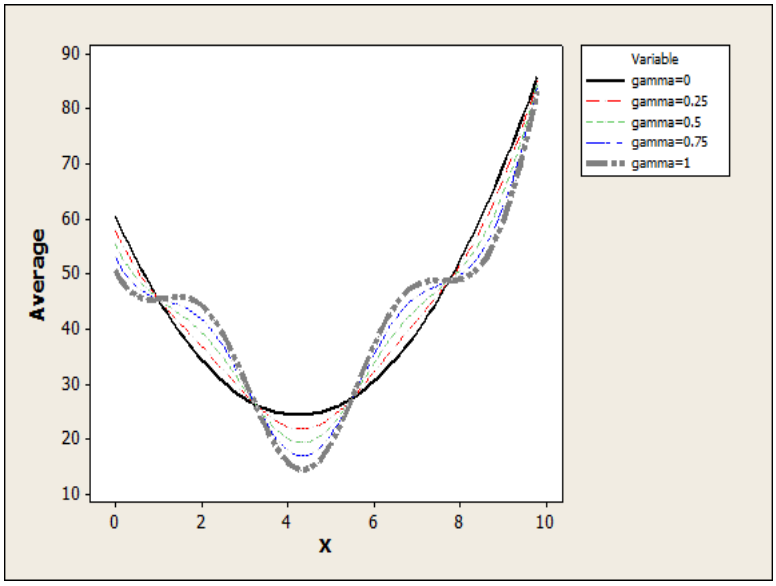

Figure 1: Plot of PA underlying models (gamma (γ) is the misspecification parameter).

an effort to achieve a compromise between extensive use of computational resources and obtaining sufficient precision of Monte Carlo results, 1,000 Monte Carlo replications were employed for each unique combination of our Monte Carlo design factors.

Six T_i^2 statistics, two each for the parametric, NP and semiparametric methods based on the fitted values and based on the estimated random effects, are computed for each Monte Carlo replication. Once the distribution of each T_i^2 statistic is obtained, an upper control limit (UCL) corresponding to an overall probability of false alarm (α) may be calculated. Hence, we need the joint distribution of the T_i^2 statistics. However, the m T_i^2 values within each method are correlated, since each T_i^2 statistic, ($i = 1,2,...,m$) is based on the same estimated mean and variance-covariance matrix, thus making the joint distribution of the T_i^2 value difficult to obtain.

As an alternative Williams et al. (2006), and Jensen et al. (2008) suggested using an approximated joint distribution assuming the T_i^2 statistic values are independent. We follow their suggestion here. Let α be the probability of a false alarm for any individual T_i^2 statistic, then the approximate overall probability of a false alarm for a sample of m independent statistics is $\alpha_{overall} = 1-(1-\alpha)^m$. Thus, for a given overall probability of a false alarm, we use $\alpha = 1-(1-\alpha_{overall})^{\frac{1}{m}}$ in calculation of $UCLs$. Jensen et al. (2008) found that $UCLs$ based on this approach performed well. When using the asymptotic chi-squared distribution of each T_i^2, the UCL is given by

$$UCL_{\chi^2} = \chi^2_{(1-\alpha,df)} \tag{29}$$

where $\chi^2_{(1-\alpha,df)}$ is the $(1-\alpha)$ quantile of a $\chi^2_{(df)}$ distribution, and df represents the degrees of freedom. The df for the parametric approach is given as the number of random effects in the estimated parametric model. For the NP approach, the df is the number of random effects plus the number of knots. The simulation results show that using the chi-square distribution as an approximation to the actual distribution of each T_i^2 along with above mention

values for the df works very well for the parametric and NP approaches. However, for the $MMRPM$ method, the simulation results show that the chi-square approximation works well when the value of df is adjusted to depend on the degree of correlation existing among the error terms. That is, for independent and low correlated errors ($\rho = 0$ or 0.2), df is calculated as the degrees of freedom for the parametric approach plus the degrees of freedom for the NP approach. For high correlated errors, df is computed as a convex combination of the parametric and NP degrees of freedom, where $\hat{\lambda}$ is the coefficient for the NP degrees of freedom. In this study, the probability of signal for the in-control dataset is 0.05, the nominal value.

In the results that follow, the $SIMSE$ and the probability of signal for the in-control and the out-of-control situations for the uncorrelated scenarios with different combinations of m, n, and different levels of misspecification were calculated.

5.1 Simulation Results

In this section, the $SIMSE$ and the simulated probability of signal for the out-of-control scenario are estimated for the uncorrelated error structure situation where $\rho = 0$. The $SIMSE$ from estimating CS curves are obtained using the parametric (Par), the NP, and the semiparametric ($MMRPM$) methods. We show here the $SIMSE$ and the probability of signal for in-control and out-of-control scenarios with various degrees of misspecified parametric models.

Table 1 contains the estimated average mixing parameter ($\bar{\hat{\lambda}}$) and the $SIMSE$ values corresponding to a given degree of misspecification for the independence case. The first and second columns give the number of profiles (m) and profile size (n), respectively. The third column contains the degree of misspecification (γ). The average estimated mixing parameter ($\bar{\hat{\lambda}}$) is given in the fourth column. Columns five through seven contain the $SIMSE$ from the parametric (Par), NP and $MMRPM$ methods, respectively.

The values in bold represent the smallest $SIMSE$ values. The values in brackets represent the Monte Carlo standard errors using 1,000 replications. All of the results in Table 1 were as expected. As the number of observations per profile (n) increases, the $SIMSE$ for each method decreases, as more observations will result in estimates that will be more precise. There is a smaller decrease for the $SIMSE$ from the parametric CS fits as n increases especially for the misspecified model. Of most interest is the behavior of $SIMSE$ as a function of γ. For $\gamma = 0$, the correct model specification case, the $SIMSE$ for the parametric and the $MMRPM$ methods are identical (or nearly so).

For large values of γ, the most severe model misspecification case, the NP and $MMRPM$ methods are identical (or nearly so). As indicated by the average $\hat{\lambda}$, the $MMRPM$ method is mostly or all composed of the NP method. For intermediate values of γ, the mild model misspecification case, the $MMRPM$ method performs better than either the parametric or the NP method. Thus, the claim that the $MMRPM$ method is robust to model misspecification is supported by this simulation. Additionally, the main advantage of the $MMRPM$ over either the parametric or the NP methods occurs when the user's model is partially correct and provides a reasonable but not wholly satisfactory fit to the data.

Table 1: $SIMSE$ and average $\hat{\lambda}$ across m, n, and γ. Monte Carlo standard errors in parenthesis. Best values in bold.

m	n	γ	$\bar{\hat{\lambda}}$	Par	NP	MMRPM
30	10	0.00	0.03	2.51 (0.45)	4.04 (0.57)	**2.50** (0.45)
		0.25	0.10	5.45 (0.46)	7.29 (0.65)	**5.40** (0.45)
		0.50	0.57	14.42 (0.53)	12.21 (0.81)	**11.86** (0.61)
		0.75	0.92	29.72 (0.78)	9.58 (0.94)	**9.55** (0.92)
		1.00	1.00	51.71 (1.28)	**11.15** (1.02)	**11.15** (1.02)
30	20	0.00	0.02	**1.25** (0.22)	2.07 (0.32)	**1.25** (0.22)
		0.25	0.20	4.24 (0.25)	5.04 (0.35)	**4.15** (0.25)
		0.50	0.96	13.26 (0.28)	**8.71** (0.46)	**8.71** (0.46)
		0.75	0.98	28.38 (0.37)	**5.44** (0.47)	**5.44** (0.47)
		1.00	1.00	49.73 (0.53)	**6.65** (0.50)	**6.65** (0.50)
60	10	0.00	0.02	**2.50** (0.32)	4.04 (0.42)	**2.50** (0.32)
		0.25	0.10	5.47 (0.35)	7.44 (0.45)	**5.43** (0.33)
		0.50	0.56	14.39 (0.36)	12.20 (0.56)	**11.80** (0.42)
		0.75	0.92	29.65 (0.52)	9.58 (0.87)	**9.57** (0.66)
		1.00	1.00	51.57 (0.85)	**11.16** (0.73)	**11.16** (0.73)
60	20	0.00	0.01	**1.26** (0.17)	2.07 (0.22)	**1.26** (0.17)
		0.25	0.20	4.26 (0.17)	5.03 (0.25)	**4.17** (0.17)
		0.50	0.97	13.27 (0.20)	**8.70** (0.33)	**8.70** (0.33)
		0.75	0.98	28.38 (0.27)	**5.44** (0.35)	**5.44** (0.35)
		1.00	1.00	49.69 (0.39)	**6.61** (0.36)	**6.61** (0.36)

Next, consider the simulated probability of signal for the in-control scenario. The $T^2_{Par1,i}$, $T^2_{NP1,i}$, and $T^2_{MMRPM1,i}$ as in Equations (7), (16) and (22), respectively, are calculated from the parametric, the NP, and the $MMRPM$ methods, respectively, using the fitted values for the CS profile at $n' = 46$ observations for each profile. Also, $T^2_{Par2,i}$, $T^2_{NP2,i}$, and $T^2_{MMRPM2,i}$ as in Equations (6), (18) and (24), respectively, are calculated based on the estimated random effects from all three approaches; parametric, NP, and $MMRPM$, respectively.

Table 2 shows the proportion of the 1,000 datasets that had a signal on the control charts for various T^2 statistics for the in-control scenario and correct model specification ($\gamma = 0$). The UCL were calculated based on the approximate chi-squared distributions with the df chosen as described above. In Table 2, the first and second columns give the number of profiles (m) and the number of observations per profile (n), respectively. The third column presents the true autoregressive coefficient values (ρ). Columns four, five and six contain the simulated probability of signal from the T^2 statistics using the fitted values from the parametric ($T^2_{Par1,i}$), the NP ($T^2_{NP1,i}$), and the semiparametric ($T^2_{MMRPM1,i}$) approaches, respectively. The simulated probability of signals from the T^2

statistics based on the estimated random effects utilizing the parametric ($T^2_{Par2,i}$), the NP ($T^2_{NP2,i}$), and the semi-parametric ($T^2_{MMRPM2,i}$) approaches are given in the seventh, eighth, and ninth columns, respectively.

Table 2: Proportion of data sets with a signal for in-control scenario using the chi-squared distribution based on df (degrees of freedom). The nominal value is 0.05.

			T^2 based on the fitted values			T^2 based on the $eblups$		
m	n	ρ	$T^2_{Par1,i}$	$T^2_{NP1,i}$	$T^2_{MMRPM1,i}$	$T^2_{Par2,i}$	$T^2_{NP2,i}$	$T^2_{MMRPM2,i}$
30	10	0.0	0.034	0.037	0.02	0.034	0.037	0.03
		0.2	0.035	0.032	0.01	0.035	0.034	0.02
		0.8	0.034	0.042	0.034	0.034	0.035	0.037
30	20	0.0	0.035	0.04	0.013	0.035	0.049	0.034
		0.2	0.035	0.046	0.01	0.035	0.044	0.024
		0.8	0.036	0.038	0.037	0.036	0.037	0.041
60	10	0.0	0.047	0.054	0.01	0.047	0.053	0.031
		0.2	0.049	0.048	0.003	0.049	0.047	0.019
		0.8	0.047	0.048	0.047	0.047	0.040	0.047
60	20	0.0	0.047	0.044	0.01	0.047	0.043	0.030
		0.2	0.046	0.046	0.004	0.046	0.043	0.010
		0.8	0.039	0.042	0.039	0.039	0.042	0.039

We see from Table 2 that for the in-control situation, it appears that the $T^2_{Par1,i}$ and $T^2_{Par2,i}$ statistics for the parametric approach based on the fitted values and the estimated random effects, respectively, give the same probability of signal for the in-control situation. This is true for all values of m, n, and ρ. The results are expected since the two $T^2_{Par,i}$ statistics are identical.

We note that the use of $T^2_{Par1,i}$, $T^2_{Par2,i}$, $T^2_{NP1,i}$ and $T^2_{NP2,i}$ statistics give a probability of signal closer to the nominal α level ($\alpha = 0.05$) and that the statistics based on the semiparametric approach have smaller probabilities, sometimes much smaller, than those based on the parametric and NP approaches. Furthermore, the $T^2_{MMRPM1,i}$, and $T^2_{MMRPM2,i}$ statistics give a probability of signal much smaller than the nominal α level for all values of m. Hence, more work is needed in this area to determine a better approximation for the UCL.

Next we consider the simulated probability of signal for the data that is generated from an out-of-control process for the uncorrelated errors case. The power studies were performed by introducing a step change (δ) in the mean vector, β, with different levels of model misspecification parameter (γ). Since the estimated probability of signal is not always 0.05 for the in-control data, the power studies were based on a simulated control limit to ensure that the probability of signal for in-control data will be the same for all charts and equal to the nominal 0.05 level.

For the generated data for m profiles, the first l of them were generated from the in-control distribution using the model as given in Equation (25) with $\beta = (\beta_1, \beta_2)^T = (5,2)^T$, and the last $m - l$ were generated from the same

model and same settings of the design factors, except that $\beta = (\beta_1, \beta_2^*)^T = (5, 2+\delta)^T$, where $\delta = 1, 2$, and 4, with $\delta = 1$ referring to a small shift and $\delta = 4$ representing a large shift. Therefore, we have introduced a step change in the mean coefficient vector (for the quadratic term and consequently in the linear term), causing the last $m - l$ profiles to be shifted away from the first l profiles with different sizes for the shift. In addition, several values of l have been tried, and we found that the probability of signal did not depend much on the value of l. Here, we present the results for $l = 20$ when $m = 30$ and $l = 40$ when $m = 60$.

Tables 3, 4 , 5 , and 6 give the simulated probability of the out-of-control signal by utilizing our six T^2 statistics for different combinations of m and n. In these tables, the first column gives the degree of misspecification (γ) and the size of the shift is given in the second column. The third through the fifth columns give the simulated probability of signal of the out-of-control situation using the parametric (Par), NP, and semiparametric ($MMRPM$) methods based on the fitted values for the CS profiles, respectively. The simulated probability of signal for the out-of-control scenario obtained via Par, NP, and $MMRPM$ methods based on the estimated random effects ($eblups$) are given in the sixth, seventh, and eighth columns, respectively.

Table 3: Simulated probability of signal for out-of-control scenario for independent data set using the six T^2 statistics with different values of misspecification and shifts for $m = 30$, $n = 10$ and $l = 20$. Best values in bold.

		T^2 based on the fitted values			T^2 based on the $eblups$		
γ	Shift	$T^2_{Par1,i}$	$T^2_{NP1,i}$	$T^2_{MMRPM1,i}$	$T^2_{Par2,i}$	$T^2_{NP2,i}$	$T^2_{MMRPM2,i}$
0.00	0	0.050	0.050	0.050	0.050	0.050	0.050
	1	**0.183**	0.176	0.142	**0.183**	0.117	0.153
	2	**0.569**	0.520	0.429	**0.569**	0.258	0.438
	4	**0.979**	0.961	0.904	**0.979**	0.376	0.837
0.25	1	0.184	0.352	**0.400**	0.184	0.274	**0.454**
	2	0.571	0.715	**0.723**	0.571	0.454	**0.777**
	4	0.979	**0.983**	0.976	**0.979**	0.626	0.971
0.50	1	0.184	**0.610**	0.600	0.184	0.526	**0.697**
	2	0.568	**0.874**	0.861	0.568	0.692	**0.930**
	4	0.983	**0.998**	0.994	0.983	0.921	**0.984**
0.75	1	0.182	**0.951**	0.852	0.182	0.926	**0.950**
	2	0.564	**0.996**	0.968	0.564	0.973	**0.991**
	4	0.978	**1.000**	**1.000**	0.978	0.996	**0.998**
1.00	1	0.185	**0.949**	0.856	0.185	**0.932**	0.834
	2	0.558	**0.996**	0.966	0.558	0.972	**0.992**
	4	0.975	**1.000**	**1.000**	0.975	**0.994**	0.993

Comparison of Table 3 to 5 and Table 4 to 6 shows that as the number of profiles (m) increases, the simulated

Table 4: Simulated probability of signal for out-of-control scenario for independent data set using the six T^2 statistics with different values of misspecification and shifts for $m = 30$, $n = 20$ and $l = 20$. Best values in bold.

γ	Shift	T^2 based on the fitted values			T^2 based on the $eblups$		
		$T^2_{Par1,i}$	$T^2_{NP1,i}$	$T^2_{MMRPM1,i}$	$T^2_{Par2,i}$	$T^2_{NP2,i}$	$T^2_{MMRPM2,i}$
0.00	0	0.050	0.050	0.050	0.050	0.050	0.050
	1	**0.183**	0.141	0.132	**0.183**	0.110	0.151
	2	**0.572**	0.474	0.388	**0.572**	0.218	0.413
	4	**0.982**	0.958	0.893	**0.982**	0.323	0.843
0.25	1	0.184	0.310	**0.409**	0.184	0.237	**0.429**
	2	0.570	0.665	**0.751**	0.570	0.393	**0.796**
	4	0.982	0.988	**0.993**	0.982	0.584	**0.996**
0.50	1	0.183	**0.539**	0.401	0.183	**0.441**	0.417
	2	0.566	**0.847**	0.757	0.566	0.634	**0.695**
	4	0.983	**1.000**	**1.000**	0.983	**0.990**	0.986
0.75	1	0.185	**0.919**	0.839	0.185	**0.903**	0.813
	2	0.566	**0.991**	0.977	0.566	**0.958**	0.908
	4	0.982	**1.000**	**1.000**	0.982	**0.991**	0.989
1.00	1	0.188	**0.920**	0.843	0.188	**0.898**	0.876
	2	0.564	**0.993**	0.975	0.564	**0.957**	0.943
	4	0.982	**1.000**	**1.000**	0.982	**0.995**	0.989

probability of signal of the out-of-control situation increases for all the methods that are presented here for most of the γ and shift values, especially for the moderate and large shifts. The NP method based on the fitted values gives a higher probability of signal than the one based on the estimated random effects. This result is due to the greater amount of information regarding differences between each CS curve and the PA curve contained in the fitted values than contained in estimated random effects.

Another important observation from the above Tables 3, 4, 5 and 6 is that the performances of the $MMRPM$ methods based on either the fitted values or the eblups are much closer to each other and are superior to the parametric and NP methods for the various degrees of model misspecification. In addition, the performance of the NP method improves as the number of observations per profile increases, since the NP fits become more accurate when the response variable is observed at more values of the regressor variable. By comparing Tables 5 with 6, it can be seen that, as n increases, the simulated probability of signal for the out-of-control situation increases over most of the combinations of γ and shift.

From the simulation study results for the uncorrelated data scenario, we conclude that the method that has the best fits, as measured by $SIMSE$, has the highest probability of signal values. Consequently, the parametric

Table 5: Simulated probability of signal for out-of-control scenario for independent data set using the six T^2 statistics with different values of misspecification and shifts for $m = 60$, $n = 10$ and $l = 40$. Best values in bold.

		T^2 based on the fitted values			T^2 based on the $eblups$		
γ	Shift	$T^2_{Par1,i}$	$T^2_{NP1,i}$	$T^2_{MMRPM1,i}$	$T^2_{Par2,i}$	$T^2_{NP2,i}$	$T^2_{MMRPM2,i}$
0.00	0	0.050	0.050	0.050	0.050	0.050	0.050
	1	**0.159**	0.130	0.125	**0.159**	0.087	0.134
	2	**0.627**	0.532	0.462	**0.627**	0.225	0.471
	4	**0.999**	0.998	0.982	**0.999**	0.429	0.967
0.25	1	0.159	0.281	**0.363**	0.159	0.206	**0.405**
	2	0.622	0.721	**0.790**	0.622	0.417	**0.849**
	4	**0.999**	0.998	0.998	**0.999**	0.671	**0.999**
0.50	1	0.155	0.492	**0.594**	0.155	0.406	**0.615**
	2	0.620	0.871	**0.912**	0.620	0.640	**0.947**
	4	0.999	0.999	**1.000**	0.999	0.964	**1.000**
0.75	1	0.156	**0.907**	0.812	0.156	0.858	**0.925**
	2	0.621	**0.988**	0.974	0.621	0.948	**0.999**
	4	0.999	**1.000**	**1.000**	0.999	0.999	**1.000**
1.00	1	0.154	**0.900**	0.811	0.154	**0.862**	0.844
	2	0.616	**0.988**	0.983	0.616	**0.955**	0.954
	4	0.999	**1.000**	**1.000**	0.980	**0.999**	0.994

method has the highest probability of signal values when γ is zero, regardless of the size of the shift. The $MMRPM$ is competitive with the parametric, and NP is competitive when using the fitted values but not when using the eblups. When γ is large, $MMRPM$ and NP are very similar in probability of signal values across all values of shift, and values of n, m, and type of T^2. The parametric has very poor probability of signal values, especially for small shifts, being competitive only if the size of the shift is large. When γ is intermediate, $MMRPM$ is clearly superior over the NP and parametric methods, although when the size of the shift is large, all methods give very similar values of probability of signal. Apparently, if the size of the shift is large enough, all these methods work about the same. Consequently, even if the model fits the data poorly the shift can still be detected. Once again, the main advantage of the $MMRPM$ is its ability to utilize information contained in the parametric model with information picked up by the NP model to work well in terms of fitting the data and well in terms of making decisions about the data regardless of the degree of model misspecification in the parametric model. The impact of changing m or n, over the values we have chosen is minimal compared to the shift effect and the γ effect.

A power study was performed using $AR(1)$ structure on the random errors with a low correlation ($\rho = 0.2$) and a high correlation ($\rho = 0.8$) to compare the parametric, the NP and the semiparametric approaches. The user's

Table 6: Simulated probability of signal for out-of-control scenario for independent data set using the six T^2 statistics with different values of misspecification and shifts for $m = 60$, $n = 20$ and $l = 40$. Best values in bold.

		T^2 based on the fitted values			T^2 based on the $eblups$		
γ	Shift	$T^2_{Par1,i}$	$T^2_{NP1,i}$	$T^2_{MMRPM1,i}$	$T^2_{Par2,i}$	$T^2_{NP2,i}$	$T^2_{MMRPM2,i}$
0.00	0	0.050	0.050	0.050	0.050	0.050	0.050
	1	**0.157**	0.135	0.117	**0.157**	0.118	0.147
	2	**0.631**	0.553	0.454	**0.631**	0.268	0.470
	4	**0.999**	0.997	0.993	**0.999**	0.502	0.970
0.25	1	0.156	0.291	**0.403**	0.156	0.264	**0.417**
	2	0.635	0.736	**0.820**	0.635	0.482	**0.877**
	4	0.999	**1.000**	**1.000**	0.999	0.756	**1.000**
0.50	1	0.160	**0.490**	0.407	0.160	0.497	**0.517**
	2	0.630	**0.880**	0.815	0.630	0.711	**0.741**
	4	0.999	**1.000**	**1.000**	**0.999**	0.901	0.996
0.75	1	0.153	**0.904**	0.850	0.153	**0.902**	0.890
	2	0.630	**0.989**	0.982	0.630	**0.979**	0.969
	4	0.999	**1.000**	**1.000**	**0.999**	**0.999**	0.997
1.00	1	0.154	**0.904**	0.850	0.154	**0.904**	0.897
	2	0.628	**0.988**	0.983	0.628	**0.979**	0.971
	4	0.999	**1.000**	**1.000**	**0.999**	**0.999**	0.998

model also incorporates this $AR(1)$ structure. Due to the space limitations, only the key results will be discussed here. One concern in the correlated data scenario is whether the misspecification term influences the estimated simulated probability of signal for the out-of-control situations to the same degree as for the uncorrelated scenario.

There are some key differences between the uncorrelated and correlated scenarios. On average, the $SIMSE$ values, for correlated error cases, increase as the correlation increases. For example, in the $m = 30$ with $n = 20$ case, the $SIMSE$ value for the parametric model at $\gamma = 0$ for $\rho = 0$, 0.2 and 0.8 are 1.25, 1.74 and 6.25, respectively. Recall that, the mean square error of fits is the sum of the squared bias plus the variance of the fits when conditioned on the values of the random effects. In the $\gamma = 0$ case, the bias is zero for the parametric and $MMRPM$ methods, hence the mean square error is the variance of the fit. If the sample size n remains fixed, and ρ increases, the effective sample size decreases (as n remains fixed), so the variance must increase as ρ increases.

There are similarities between the uncorrelated case and the correlated errors structure cases. For $\gamma = 0$, the parametric methods have the highest simulated probability of signal for the out-of-control and the $MMRPM$ procedure obtains the simulated probability of signal values very close to the parametric simulated probability of signal values. For $\gamma = 1$, the NP method have the highest simulated probability of signal and the $MMRPM$ gives

simulated probability of signal very close to the NP values. As γ increases from zero to one, the simulated probability of signal values for the $MMRPM$ method are either the highest values (for low to moderate model misspecification), or are very close in value to the "best" simulated probability of signal values. Also, it is observed that as the number of profiles (m) increases the simulated probability of signal from the NP and $MMRPM$ methods increase for different shift sizes and various degrees of model misspecification. As the number of observations per each profile (n) increases the power of the $MMRPM$ method to detect the step shift gets better. In addition, by increasing the degree of autocorrelation between errors the simulated probability of signal for the NP and $MMRPM$ methods increases. While as ρ increases the simulated probability of signal for the parametric method decreases especially for the small and moderate shift sizes. In general, as ρ increases the performance of the NP and $MMRPM$ methods increases for all values of m and n.

In summary, when the degree of misspecification increases the semiparametric and NP approaches show greater superiority over the parametric approach. We see that the semiparametric approach is always at least equivalent to the parametric approach (for the correctly specified model) and often far superior. The $MMRPM$ and the NP approaches maintain a high probability of signal for data sets with correlated errors structures, a condition common in longitudinal data sets.

6 THE AUTOMOBILE ENGINE APPLICATION

In this section, we apply our proposed methods to a "real world" profile monitoring problem. This application from Amirhossein et al. (2010) concerns the automotive industry. The application is for the engine type $TU3$ which are assembled for a French automobile, the Peugeot. In the study, the engine is run at different RPM values and the corresponding torque values obtained. The torque produced by the engine is considered as response variable and the corresponding speed RPM is considered as the explanatory variable. The profiles that describe the relationship between torque and RPM should be similar, when the manufacturing process is in-control. An engine with mechanical defects or any other issues will result in an outlying engine profile. Because there are multiple RPM values obtained for each engine it is natural to try to apply a multivariate quality control procedure to discover engines which would not be acceptable.

In this data set, we have a total of twenty-six engines as our historical data set (HDS) for Phase I analysis. For each engine, 14 engine speed values are set to 1500, 2000, 2500, 2660, 2800, 2940, 3500, 4000, 4500, 5000, 5225, 5500, 5775, and 6000 RPM and the corresponding torque values are reported. Therefore, 14 points are collected for each engine and form the engine profile of interest. The raw data set, with straight line segments connecting the torque values, is shown in Figure 2. Figure 2 illustrates a typical automobile engine profiles. The data set contains 26 profiles each of which consists of 14 measurements along the profile. Within each engine, there is no replication which means that there is a single measurement at each RPM. The raw data set contains some peaks and dips, not easily captured by a purely parametric model. These curves are clearly nonlinear and thus new techniques to

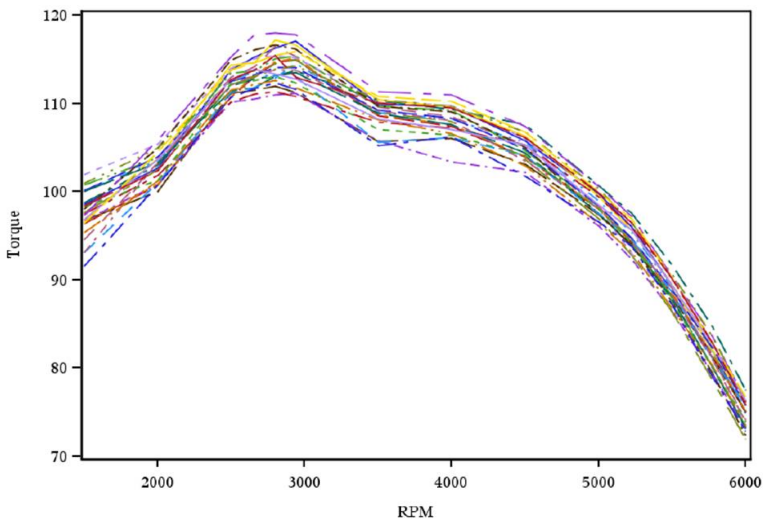

Figure 2: The raw data set for 26 automobile engines (Torque vs. RPM).

monitor these curves are desirable. Therefore, our two new approaches (NP and semiparametric) will be applied to this data set.

In the next section, the relationship between torque and RPM of an engine will be modeled by the parametric, NP and semiparametric ($MMRPM$) methods. In addition, the mean square error (MSE) is calculated for each method.

6.1 Phase I Analysis

In a Phase I analysis, we have m quality profiles in the HDS. The main goal of the Phase I analysis is to estimate process parameters based on an in-control process. The proposed approaches that we consider for this data set will account for the autocorrelation within engines. These approaches (parametric, NP and semiparametric) can be considered as methods for data reduction that enable the determination of outlying profiles.

The parametric approach using the mixed second order quadratic model in one regressor is

$$\mathbf{y}_i = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})\mathbf{x}_i^* + (\beta_2 + b_{i2})\mathbf{x}_i^{*2} + \epsilon_i \qquad i = 1, 2, \ldots, 26 \tag{30}$$

where $\mathbf{y}_i$ is the vector containing the torque values for the i^{th} engine. The $\mathbf{x}_i^*$ is a vector containing the centered RPM values, $\mathbf{x}_i^* = \mathbf{x}_i - \bar{\mathbf{x}}$. Since the $PRMs$ are fixed for each engine, we denote the RPM, x_{ij} by simply $\mathbf{x}_i$. The ϵ_i is the vector containing the errors for the i^{th} engine. The parameters, β_0, β_1 and β_2 are the fixed effects that are common for all profiles. The b_{i0}, b_{i1} and b_{i2} are the random effects for the i^{th} engine profile, and are unique for each engine profile.

Because the random errors for each profile are likely to be correlated, we must account for their correlation in the analysis. An $AR(1)$ structure for the variance-covariance matrix $\mathbf{V}$ is utilized to model the errors within a profile, as assumed by Amirhossein et al. (2010) for the linear mixed model (the parametric approach). The NP approach utilizing the linear mixed p-spline regression model, as described in Section 3, with $p = 1$ and utilizing

the radial basis. In addition, the semiparametric approach via the $MMRPM$ method, as described in Section 4, is applied to this data set.

The fitted profiles using the parametric, NP, and $MMRPM$ approaches are illustrated in Figure 3 with the raw data (displayed as in Figure 2) as well for the twenty-six engines. One can see that the parametric approach is able to capture the main characteristics in the raw data set but does not capture several dips and wiggles that may be important, suggesting that the quadratic model is misspecified. On the other hand, the NP and semiparametric approaches capture most of the main features in the data set. Furthermore, the MSE values are 4.58, 0.44 and 0.43 for the parametric, NP and $MMRPM$ fits, respectively. In addition, the semiparametric approach relies more on the NP approach than the parametric model indicating that the parametric model is incorrectly specified. The estimated mixing parameter ($\hat{\lambda}$) value is approximately 0.98. Thus, 98% of the semiparametric fit comes from the NP fit while only 2% comes from the parametric fit. Based on the MSE criteria, we can say that the NP and semiparametric approaches are superior to the parametric approach for fitting this data set.

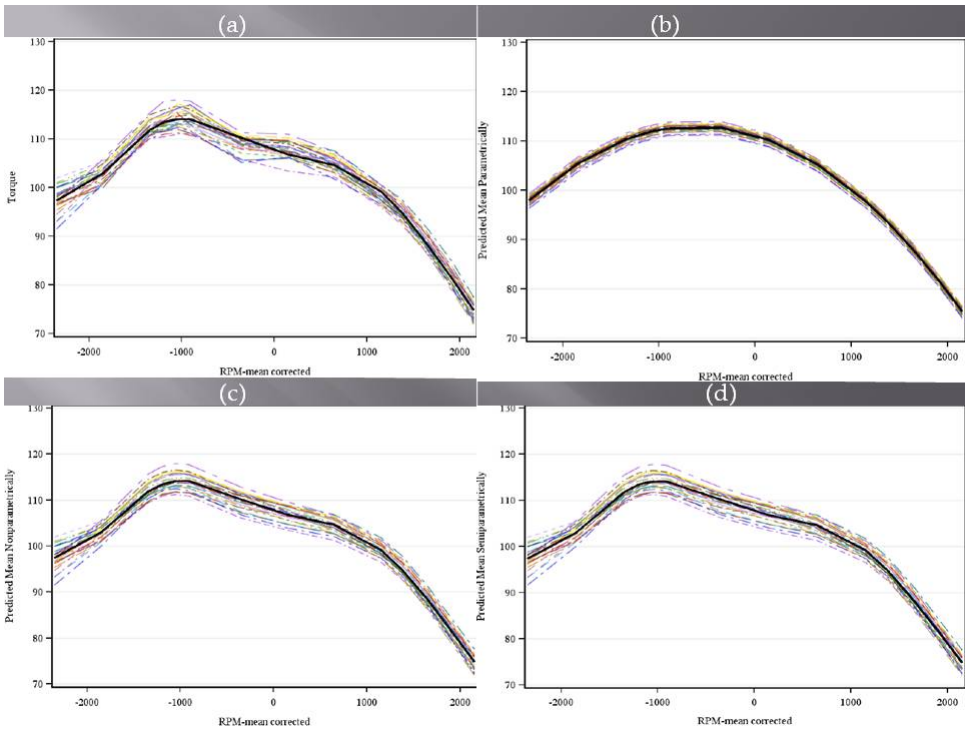

Figure 3: (a) The raw data set for 26 automobile engines with the PA curve (solid black) (b) Fitted profiles using a second order polynomial mixed model (Parametric Approach) with the PA curve (solid black) (c) Fitted profiles using linear mixed p-spline regression (NP Approach) with the PA curve (solid black) and (d) Fitted profiles using MMRPM method (Semiparametric Approach) with the PA curve (solid black).

Since all approaches considered in this paper assume that the random errors are normally distributed, the first step in the Phase I analysis is to assess the viability of this assumption. To examine if the normality is a reasonable assumption for the parametric, NP and $MMRPM$ methods, we show in Figure 4 a histogram of the residuals from

each method as an exploratory measure. The residuals have been computed from each method for the 26 engines using the $AR(1)$ structure for the variance-covariance matrix. Traditional tests of normality, such as the Anderson-Darling test, are not valid here since the residuals are correlated. Therefore, we use these histograms as graphical aids to check the normality assumption.

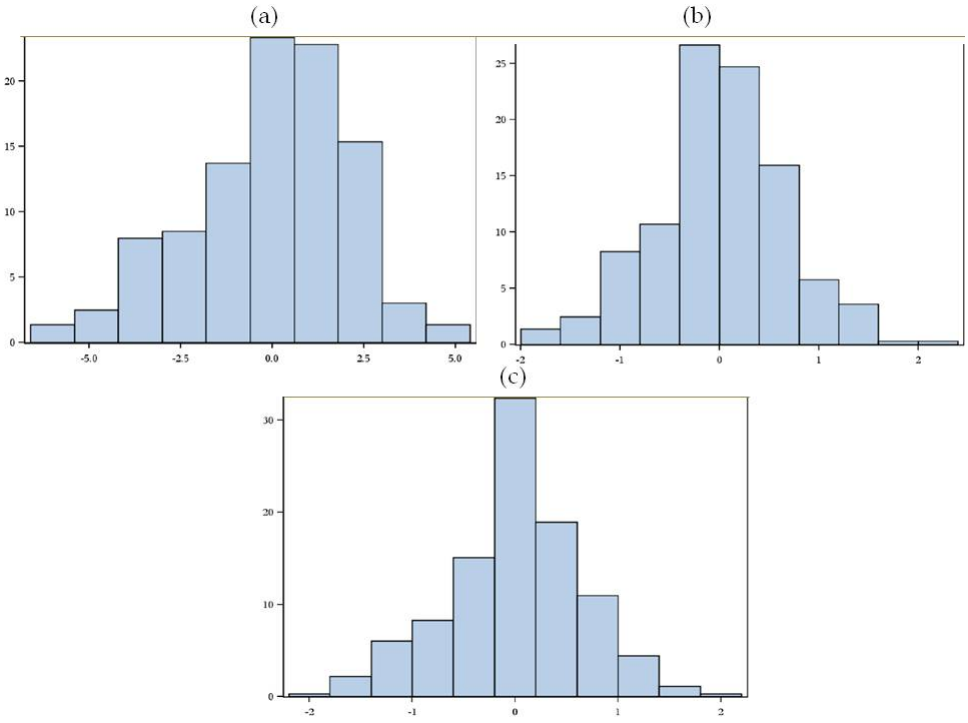

Figure 4: (a) Histogram for the residuals from the fitted profiles using the parametric approach (b) Histogram for the residuals from the fitted profiles using the NP approach and (c) Histogram for the residuals from the fitted profiles using the semiparametric approach.

Figure 4, gives the histogram for the residuals from the parametric approach (sub-figure (a)). It can be seen that, the residuals are centered around zero but there is some amount of left-skewness in the data. This result is in agreement with the one from Amirhossein et al. (2010). The histogram for the residuals from the NP approach (sub-figure (b)) looks very symmetric around zero and the range for the residual values from two to negative two. The residuals from the semiparametric approach (sub-figure (c)) strongly resemble data from a normal distribution with mean zero. Hence, the residuals resulting from the NP and $MMRPM$ methods appear to more strongly support the normality assumption than those from the parametric method.

In the following section, we use a T^2 based control procedure to investigate the stability of the process as well as determine if there are outlying profiles using the parametric, NP and semiparametric approaches.

6.2 The T^2 Control Chart

In order to determine if a profile is an outlier or if the profile has shifted, we propose two methods for each approach: one, by analyzing the vector of the fitted values and the second by utilizing the vector of the estimated random effects from the parametric, NP and $MMRPM$ approaches. In Phase I analysis, we are interested in using the HDS to estimate the mean vector μ_b and variance-covariance matrix Σ_b of the $\hat{\mathbf{b}}_i$ vectors, after we have removed the out-of-control profiles.

One method of identifying out-of-control profile is use of the Hotelling's T^2 statistics. The simulation results of Section 5 indicate that the T^2 statistics based on the fitted values and the estimated random effects are identical for the parametric approach and nearly equal for the other two approaches. Hence, we only focus on the T^2 statistics based on the estimated random effects from each one of our methods as given in Equations (6), (18), and (24).

The T^2 control chart consists of a plot of T_i^2 statistics by i for all profiles in the HDS. If a T_i^2 statistic exceeds the UCL associated with the chart, then that estimated profile is identified as a possible outlier to be removed from the HDS. As pointed out in the simulation section, use of the asymptotic distribution for determining the $UCLs$ for the NP and semiparametric is a conservative approach, providing $UCLs$ that are slightly larger than they should be resulting in probability of signals slightly smaller than the nominal value.

For this analysis, the asymptotic distribution of T_i^2 using the parametric approach is χ^2_{df} for all $i = 1,2,\ldots,m$. The UCL is calculated as $UCL_{Par} = \chi^2_{(1-\alpha,q)}$ where $\chi^2_{(1-\alpha,q)}$ is the $(1-\alpha)^{th}$ quantile from a χ^2 distribution with q degrees of freedom where α is adjusted as described in Section 5, and q replaced by the number of estimated random effects. The UCL for the T_i^2 statistics based on the NP estimated random effects is obtained as $UCL_{NP} = \chi^2_{(1-\alpha,df_{NP})}$ with df_{NP} replaced by the number of estimated random effects plus the number of knots. Thus $df_{NP} = p + K_1 + K_2$. For this application $p = 1, K_1 = 2$ and $K_2 = 2$.

In addition, the estimated first-order autocorrelation, estimated from the $MMRPM$ residuals, is 0.5 approximately. Thus, using our results from the simulation section, the UCL for the T_i^2 statistics based on the $MMRPM$ estimated random effects is obtained as in Equation (31), as a convex combination from the parametric UCL and the NP UCL via the estimated mixing parameter $\hat{\lambda}$ as given in Equation (21).

$$UCL_{MMRPM} = (1-\hat{\lambda}) * UCL_{Par} + \hat{\lambda} * UCL_{NP} \tag{31}$$

For this analysis, we choose $\alpha_{overall} = 0.05$ which corresponds to

$$\alpha = 1-(1-\alpha_{overall})^{\frac{1}{m}} = 1-(1-0.05)^{\frac{1}{26}} = 0.0019709 \tag{32}$$

The upper control limits are calculated as $UCL_{Par} = 12.459$, $UCL_{NP} = 18.942$ and $UCL_{MMRPM} = 18.755$.

The T^2 chart for the $m = 26$ profiles is used to identify profiles with abnormal estimated random effects values. The T_i^2 values for the parametric, NP and $MMRPM$ as given in Equations (6), (18), and (24) are computed and the corresponding control charts are given in Figure 5.

From Figure 5, we see that the use of the T^2 charts based on the parametric approach (sub-figure (a)) does not produce a signal. This result is in agreement with that of Amirhossein et al. (2010). In addition, the NP approach

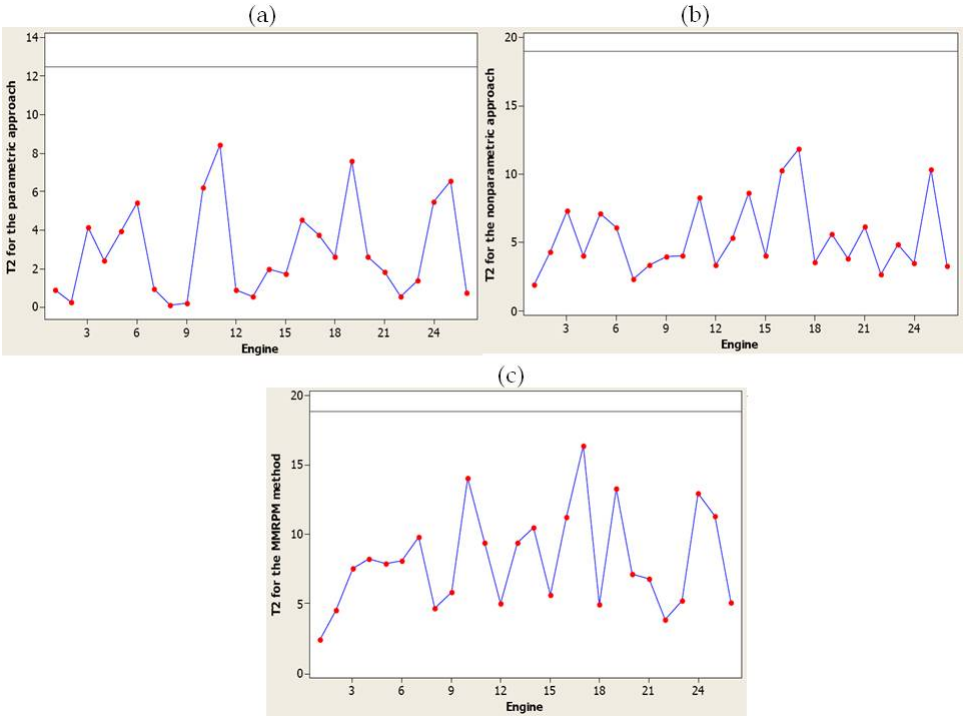

Figure 5: (a) T^2 control chart based on the estimated random effects parametrically. (b) T^2 control chart based on the estimated random effects nonparametrically (c) T^2 control chart based on the estimated random effects semiparametrically.

(sub-figure (b)) and $MMRPM$ approach (sub-figure (c)) also do not produce a signal. It is interesting to note that the parametric, NP and semiparametric methods agree on this decision even thought the NP and $MMRPM$ methods give much better fits for the engine profiles in term of smaller mean square errors.

Our MSE results along with the plots shown in Figure 3 indicate that the mixed quadratic parametric model is not able to capture the main features in this data set. Also, the parametric model produced the largest MSE in comparison with the NP and $MMRPM$ methods. Our simulation results indicate that using the T^2 approach with a misspecified parametric model is not as powerful at detecting step-shifts in profiles away for the normative profiles as the NP or $MMRPM$ method. For these reasons the user should strongly consider using either the NP or $MMRPM$ results for this data. Despite the fact that all three methods reach the same conclusion regarding the "in-control" status of each profile, the NP and $MMRPM$ results provide a better description of the actual behavior of each engine profile. Thus, the NP and $MMRPM$ methods give the user greater ability to properly interpret the true relationship between engine speed and torque for this type of engine and an increased likelihood of detecting unusual engines in future production. As a result of this analysis, we conclude that all 26 engines can be used to obtain the parameters estimates on which the Phase II control charts will be based. Our conclusion agrees with that expressed in Amirhossein et al. (2010) where they found no unusual profiles using the quadratic model.

7 SUMMARY

In this paper, we have developed statistical procedures to monitor a product or process whose quality is measured across a continuum where a plot of the resulting response forms a profile. This research studied the mixed effects models by introducing two new techniques for profile monitoring in Phase I analysis. The first proposed technique was the NP approach via a mixed p-spline regression. The NP approach provides flexible fits to PA and CS profiles. The second proposed technique is a semiparametric procedure via the $MMRPM$ method in which we combine both parametric and NP profile fits via a mixing parameter to gain advantages from both. For each case, we formulated two Hotelling's T^2 statistics, one based on the estimated random effects and one based on the fitted values, and describe how the corresponding UCL values are obtained.

A Monte Carlo study and a real dataset were performed to compare the $SIMSE$ and the simulated probability of signal of the parametric, NP, and semiparametric approaches. Both correlated and uncorrelated errors structure scenarios were evaluated for varying amounts of model misspecification, number of profiles, number of observations per profile, shift location, and in- and out-of-control situations. The $MMRPM$ method for uncorrelated and correlated scenarios was competitive and, often, clearly superior with the parametric and NP over all levels of misspecification. Therefore, this research supports the claim that the NP and $MMRPM$ methods are robust to model misspecification.

We also found that both the NP and the semiparametric methods result in charts with good abilities to detect changes in Phase I data, and in charts with easily calculated though conservative control limits. The proposed methods provide greater flexibility and efficiency than current parametric methods used in profile monitoring that rely on correct model specification.

References

Abdel-Salam, A.-S. G. (2009). *Profile Monitoring with Fixed and Random Effects using Nonparametric and Semiparametric Methods*. PhD thesis, Department of Statistics, Virginia Polytechnic Institute & State University, Blacksburg, VA.

Amirhossein, 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*, 26-5:509–520.

Henderson, C. (1950). Estimation of genetic parameters. *Annals of Mathematical Statistics*, 21:309–310.

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.

Kang, L. and Albin, S. L. (2000). On-line monitoring when the process yields a linear profile. *Journal of Quality Technology*, 32:418–426.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963–974.

Mays, J. E., Birch, J. B., and Starnes, B. A. (2001). Model robust regression: Combining parametric, nonparametric, and semiparametric methods. *Journal of Nonparametric Statistics*, 13:245 – 277.

Montgomery, D. C. (2005). *Introduction to Statistical Quality Control,* 5^{th} *Edition*. John Willy&Sons, Inc.

Nottingham, Q. and Birch, J. (2000). A semiparametric approach to analysing dose-response data. *Statistics in Medicine*, 19:389–404.

Pickle, S. M., Robinson, T. J., Birch, J. B., and Anderson-Cook, C. (2008). A semi-parametirc approach to robust parameter design. *Journal of Statistical Planning and Inference*, 138:114–131.

Qiu, P. and Zou, C. (2009). Control chart for monitoring nonparametric profiles with arbitrary design. *Statistica Sinica*, 19:1337–1357.

Qiu, P., Zou, C., and Wang, Z. (2010). Nonparametric profile monitoring by mixed effects modeling. With Discussion. *Technometrics*, 52(3):265–287.

Robinson, T., Birch, J. B., and Starnes, A. (2010). A semi-parametric approach to dual modeling when no replication exists. *Journal of Statistical Planning and Inference*, 140:2860–2869.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, NY.

Sullivan, J. H. and Woodall, W. H. (1996). A comparison of multivariate control charts for individual observations. *Journal of Quality Technology*, 28:398–408.

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

Wan, W. and Birch, J. B. (2010). A semiparametric technique for the multi-response optimization problem. *Quality and Reliability Engineering International, (in press)*, http://www.web-e.stat.vt.edu/dept/web-e/tech_reports/TechReport09_1.pdf.

Waterman, M. J., Birch, J. B., and Schabenberger, O. (2007). Linear mixed model robust regression. *Technical Report, Department of Statistics, Virginia Polytechnic Institute & State University, Blacksburg, VA. http://www.web-e.stat.vt.edu/dept/web-e/tech_reports/TechReport07-3.pdf*.

Wegener, M. and Kauermann, G. (2008). Examining heterogeneity in implied equity risk premium using penalized splines. *Advances in Statistical Analysis*, 92:35–56.

Wei, Y., Zhao, Z., and Lin, D. K. (2010). A general class of nonparametric L-1 regression with its application to profile control charts. To appear in *Annals of Applied Statistics*.

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 covariance matrix estimator. *Journal of Quality Technology*, 38:217–229.

Woodall, W. H. (2007). Current research on profile monitoring. *Producao*, 36:309–320.

Woodall, W. H., Birch, J. B., and Du, P. (2010). Comment on "nonparametric profile monitoring by mixed effects modeling". *Technometrics*, 52(3):285–287. doi:10.1198/TECH.2010.09165.

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.

Zou, C., Qiu, P., and Hawkins, D. (2009). Nonparametric control chart for monitoring profiles using the change point formulation. *Statistica Sinica*, 19:1337–1357.

Zou, C., Tsung, F., and Wang, Z. (2008). Monitoring profiles based on nonparametric regression methods. *Technometrics*, 50(4):512–526.