Phase I Monitoring of Serially Correlated Nonparametric Profiles By Mixed-Effects Modeling (Running Head: Serially Correlated Nonparametric Profile Monitoring)

Qin Zhou[1] and Peihua Qiu[2]

[1] School of Mathematics and Statistics

Jiangsu Normal University, Xuzhou, Jiangsu, China

[2]Department of Biostatistics

University of Florida, Gainesville, FL, USA

E-mail of the corresponding author: pqiu@ufl.edu

July 12, 2021

Abstract

Profile monitoring is an active research area in statistical process control (SPC) because it has many important applications in manufacturing and other industries. Early profile monitoring methods often impose model assumptions that the mean profile function has a parametric form (e.g., linear), profile observations have a parametric distribution (e.g., normal), and within-profile observations are independent of each other. These assumptions have been lifted in some recent profile monitoring research, making the related methods more reliable to use in various applications. One notoriously challenging task in profile monitoring research is to properly accommodate serial data correlation among profiles observed at different time points, and this task has not been properly addressed in the SPC literature yet. Serial data correlation is common

in practice, and it has been well demonstrated in the literature that control charts are unreliable to use if the serial data correlation is ignored. In this paper, we suggest a novel mixed-effects model for describing serially correlated univariate profile data. Based on this model, a Phase I profile monitoring chart is developed. This chart is flexible in the sense that it does not require any parametric forms for describing the mean profile function and the profile data distribution. It can accommodate both the within-profile and between-profile data correlation. Numerical studies show that it works well in different cases.

Keywords: Data correlation; Local linear kernel smoothing; Mixed-effects modeling; Nonparametric profile monitoring; Statistical process control; Between-profile data correlation.

1 Introduction

Statistical process control (SPC) charts provide a powerful statistical tool for monitoring production lines in manufacturing industries for quality control and management purposes. In some applications, the quality characteristic of a product is reflected in the functional relationship between a response variable and one or more explanatory variables. In these applications, we collect a set of data points of these variables from the sampled product at a given time point. These data points from a single product form a so-called profile, and the related process monitoring problem is called *profile monitoring* in the SPC literature (Jin and Shi 1999, Kang and Albin 2000). This paper aims to develop a flexible and effective new method for univariate profile monitoring.

Due to broad applications of profile monitoring, there have been many existing discussions on this topic in the literature. Early profile monitoring research assumes that the related in-control (IC) profile mean function has a linear form. In such cases, profile monitoring becomes the monitoring of the estimated linear regression coefficients obtained from the individual profile data (cf., Jensen et al. 2008, Kang and Albin 2000, Kim et al. 2003, Mahmoud et al. 2007, Mahmoud and Woodall 2004, Stover and Brill 1998, Wang and Tsung 2005 and Zou et al. 2006). Some researchers have generalized certain linear profile monitoring methods to different cases with parametric nonlinear profiles. See, for instance, Ding et al. (2006), Jensen and Birch (2009),

Williams et al. (2007a,b), and Zou et al. (2007). Of course, the above-mentioned parametric profile monitoring methods depend heavily on the assumed parametric models, and they would not be reliable in cases when such parametric models are invalid. To overcome this limitation, some nonparametric profile monitoring methods have been proposed in the literature. See, for instance, McGinnity et al. (2015), Paynabar et al. (2016), Qiu and Zou (2010), Qiu et al. (2010), Wei et al. (2012), and Zou et al. (2008, 2009). For an overview on profile monitoring, see Qiu (2014, Chapter 10).

In the literature, most existing profile monitoring control charts (e.g., Zou et al. 2008) require a fundamental assumption that observations within a profile and between profiles are independent of each other, which is often invalid in applications. Some methods can accommodate the within-profile data correlation by using various mixed-effects or Gaussian process models for describing the observed profile data (e.g., Abdel-Salam and Brich 2019, Abdel-Salam et al. 2013, Jensen et al. 2008, Jensen and Birch 2009, Qiu et al. 2010, Zhang et al. 2014). But, the observed data between different profiles were still assumed to be independent in all these methods. To handle some different but related research problems, there are some existing discussions on monitoring correlated data. One is about dynamic screening (DS), in which many different dynamic processes need to be monitored separately (Li and Qiu 2016, Qiu and Xiang 2014, Qiu et al. 2018). The DS problem is different from the current profile monitoring problem in that each dynamic process to monitor in the former is a conventional univariate or multivariate process with scalar or vector observations at different time points and thus its observations are not profiles. If each dynamic process in the DS problem is regarded as a profile, then different profiles are always assumed to be independent in the DS literature, although observations within each dynamic process could be correlated (Li and Qiu 2016). Another related problem is spatio-temporal (ST) data monitoring (e.g., Colosimo et al. 2014, Wang et al. 2014, Yan et al. 2018, Yang and Qiu 2020, Zang and Qiu 2018). For monitoring ST data, proper description and estimation of the spatial mean and variance/covariance structures of the observed ST data are the key issues and they are often challenging to address. To simplify the ST data monitoring problem, some existing methods assume that the observed ST data are

independent both spatially and temporally (Yan et al. 2018, Zang and Qiu 2018), while some others assume that the observed ST data follow the Gaussian process models (Colosimo et al. 2014, Wang et al. 2014). These ST data monitoring methods are designed specifically for their own spatial process monitoring problems, and might be appropriate for these specific applications. But, they cannot be used for the current univariate nonparametric profile monitoring problem, because their model assumptions (e.g., ST data independence or Gaussian process models) are invalid here and some of them require a relatively large number of observations at each observation time which may not be available in the current problem. So, we could not find any existing research on monitoring univariate nonparametric profiles with both within-profile and between-profile data correlation in the profile monitoring literature. In practice, however, serial data correlation is common. It has been well demonstrated in the SPC literature that control charts could be unreliable to use if such serial data correlation is not accommodated properly in their construction and design (e.g., Hawkins and Olwell 1998, Qiu 2014). More specifically, it is well known that serial data correlation can cause a dramatic increase or decrease in the false alarm rate of the related control charts when detecting process distributional shifts (e.g., Apley and Lee 2003, Apley and Lee 2008, Li and Qiu 2020). Therefore, it is critically important to address this issue when developing new profile monitoring methods.

This paper aims to develop a Phase I univariate nonparametric profile monitoring chart for monitoring serially correlated profiles, which is notoriously difficult partly because there are no corresponding methods in nonparametric longitudinal data analysis that can accommodate between-profile data correlation (e.g., Chen and Jin 2005, Li 2011, Wu and Zhang 1992, Xiang et al. 2013). To this end, a new mixed-effects model is first suggested for describing nonparametric univariate profiles with both within-profile and between-profile data correlation. Based on this modeling approach, a Phase I univariate nonparametric profile monitoring chart is developed. Besides the flexibility that it can accommodate both within-profile and between-profile data correlation, this chart does not impose any parametric forms on the IC mean profile function and the profile data distribution. Thus, it should be appropriate to use in many univariate profile monitoring applica-

tions.

The remaining parts of the article are organized as follows. The new mixed-effects model and the proposed Phase I profile monitoring chart are described in detail in Section 2. Some simulation studies are presented in Section 3 to evaluate their numerical performance. The proposed chart is applied to a real-data application in Section 4. Several remarks conclude the article in Section 5.

2 Proposed Method

This section consists of two parts. In Subsection 2.1, a nonparametric mixed-effects model is suggested for describing serially correlated univariate profile data. Based on this model, a Phase I profile monitoring chart is suggested in Subsection 2.2.

2.1 Nonparametric mixed-effects modeling of serially correlated univariate profile data

Assume that we have observed data of m IC profiles, denoted as $\{(x_{ij}, y_{ij}), j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, m\}$, where y is a response variable and x is a predictor. To describe the possible within-profile and between-profile data correlation, the following mixed-effects model is considered:

$$y_{ij} = g(x_{ij}) + \xi_i + \gamma(x_{ij}) + \varepsilon_{ij}, \text{ for } j = 1, 2, \ldots, n_i, \ i = 1, 2, \ldots, m, \tag{1}$$

where $g(x)$ is the population mean profile function of x that is assumed to be a continuous function, ξ_i is a random-effects term for describing the variation of the ith profile from $g(x)$, $\gamma(x)$ is a random-effects process that depends on x only for describing the between-profile data correlation at the location x, and ε_{ij} is the random error term. In Model (1), without loss of generality, it is assumed that $x_{ij} \in [0, 1]$, for all i and j. It is further assumed that the three random terms ξ_i, $\gamma(x_{ij})$ and ε_{ij} are all independent of each other, and they have mean 0 and variances σ_ξ^2, σ_γ^2 and σ_ε^2, respectively. Inclusion of the random-effects term ξ_i is for accommodating the within-profile data correlation (cf., Diggle et al. 2002). Similarly, inclusion of the random-effects process $\gamma(x)$ is for accommodating the between-profile data correlation. The idea to suggest the mixed-effects model (1) is motivated

from the perspective of spatio-temporal data modeling, where i is related to the observation time of the ith profile and thus denotes the time domain, and x denotes the space domain. The two random-effects terms ξ_i and $\gamma(x)$ describe the variation of the ith profile function from g in the time and space domains, respectively. As far as we know, there are no nonparametric control charts for monitoring spatio-temporal data that are based on spatio-temporal mixed-effects modeling yet. So, our proposed method in this paper can fill this gap.

In Model (1), the design points $\{x_{ij}\}$ are allowed to be different in all m profiles. Thus, there could be many different values in the set of all design points. In the extreme case when no two design points in the set $\{x_{ij}, j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, m\}$ are the same, the number of random-effect terms in $\{\gamma(x_{ij})\}$ would be the same as the total number of observations. In such a case, Model (1) would be inestimable. To overcome this difficulty and make the estimation of Model (1) simpler, the following model approximation is suggested. Let $N = \sum_{i=1}^{m} n_i$ and $d_q = q/p$, for $q = 1, 2, \ldots, p-1$. Then, the design interval $[0, 1]$ can be divided into p subintervals: $[0, 1] = [0, d_1] \bigcup (d_1, d_2] \bigcup \ldots \bigcup (d_{p-1}, 1]$. The number of subintervals, p, is chosen such that the set of design points in each subinterval is not empty. In each subinterval, it is assumed that the random-effect term $\gamma(x_{ij})$ is the same. Namely, when $x_{ij} \in (d_{q-1}, d_q]$, this random-effect term is denoted as η_q that depends on q only, for $q = 1, 2, \ldots, p$, where $d_0 = 0$ and $d_p = 1$. In such cases, Model (1) can be approximated by

$$y_{ij} = g(x_{ij}) + \xi_i + \eta_q + \varepsilon_{ij}, \text{ for } x_{ij} \in [d_{q-1}, d_q], q = 1, 2, \ldots, p, j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, m. \quad (2)$$

In Model (2), it is further assumed that $\{\xi_i, i = 1, 2, \ldots, m\}$ and $\{\eta_q, q = 1, 2, \ldots, p\}$ are both independent and identically distributed, as in a conventional random-effects model (cf., Diggle et al. 2002). Then, $\{\xi_i, i = 1, 2, \ldots, m\}$ and $\{\eta_q, q = 1, 2, \ldots, p\}$ are for describing the within- and between-profile data correlation, respectively. Obviously, the random process $\gamma(x)$ has been approximated by the step function $\sum_{q=1}^{p} \eta_q I(x \in (d_{q-1}, d_q])$ in (2), which is not continuous in x. Because Model (2) is only used in the intermediate steps for constructing the proposed control chart that will be discussed in the next subsection, this would not be a substantial issue. However, if the

discontinuity of the step function is an issue in some other applications, then it can be replaced by a continuous function that connects the consecutive points in $\{((d_{q-1}+d_q)/2, \eta_q), q = 1, 2, \ldots, p\}$. The resulting function is the smoothing spline of order 1 (cf., Qiu 2005, Section 2.5). We can also consider using smoothing splines of higher orders if a smoother approximation is desired in an application.

It should be pointed out that the random-effects terms in Model (1) are kept simple in this paper to describe the basic within-profile and between-profile data correlation. Actually, they can both be generalized to describe a more complicated data correlation. For instance, the random-effects term $\gamma(x_{ij})$ can be replaced by $\gamma_1(x_{ij}) + \gamma_2(x_{ij})x_{ij}$ or a higher-order polynomial of x_{ij} to describe a complicated between-profile correlation, where $(\gamma_1(x_{ij}), \gamma_2(x_{ij}))^T$ is a random-effects coefficient vector. In such a more general model, $\gamma_1(x_{ij})$ and $\gamma_2(x_{ij})$ can still be approximated by smoothing spline functions of x_{ij}, as discussed above about Model (2). Similarly, the random-effects term ξ_i can be generalized to describe a more complicated within-profile data correlation. For a systematic discussion about such general mixed-effects models, see the book Diggle et al. (2002). Nonparametric profile modeling using such more complicated mixed-effects models will be discussed in our future research.

Next, we discuss estimation of $g(x)$, for $x \in [0, 1]$, and other quantities in Model (2). To this end, consider a neighborhood $[s-h, s+h]$ of a given point $s \in [0, 1]$ with the bandwidth h. Assume that this neightborhood has non-empty intersections with p^* subintervals in $\{(d_{q-1}, d_q], q = 1, 2, \ldots, p\}$, and the corresponding p^*-dimensional vector of the random-effects terms in $\{\eta_q, q = 1, 2, \ldots, p\}$ is denoted as $\boldsymbol{\eta}(s)$. Then, $g(x)$ can be approximated by a linear function $\beta_0 + \beta_1(x - s)$ in the neighborhood $[s-h, s+h]$, and $g(x_{ij})$ can be approximated by $\beta_0+\beta_1(x_{ij}-s)$, for $x_{ij} \in [s-h, s+h]$. To estimate $g(x)$, let us consider the following penalized local kernel least squares estimation problem:

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\eta}} \quad & \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{\beta} - \xi_i \mathbf{1}_{n_i} - \boldsymbol{M}_i(s)\boldsymbol{\eta}(s))^T \boldsymbol{K}_i (\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{\beta} - \xi_i \mathbf{1}_{n_i} - \boldsymbol{M}_i(s)\boldsymbol{\eta}(s)) \\ & + \sum_{i=1}^{m} \frac{\xi_i^2}{\sigma_\xi^2} + \frac{1}{\sigma_\eta^2} \boldsymbol{\eta}(s)^T \boldsymbol{\eta}(s), \end{aligned} \tag{3}$$

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$, $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}^T, \boldsymbol{Z}_{i2}^T, \ldots, \boldsymbol{Z}_{in_i}^T)^T$, $\boldsymbol{Z}_{ij} = (1, x_{ij} - s)^T$, for $j = 1, 2, \ldots, n_i$, $\boldsymbol{M}_i(s)$ is an $(n_i \times p^*)$-dimensional matrix whose (j,q)th element is 1 if x_{ij} is contained in the qth subinterval that has non-empty intersection with $[s-h, s+h]$ and 0 otherwise, for $j = 1, 2, \ldots, n_i$ and $q = 1, 2, \ldots, p^*$, $\mathbf{1}_{n_i}$ is an n_i-dimensional vector with all elements being 1, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_m)^T$, $\boldsymbol{K}_i = \mathrm{diag}[K_h(x_{i1} - s), K_h(x_{i2} - s), \ldots, K_h(x_{in_i} - s)]$, for $i = 1, 2, \ldots, m$, $K_h(\cdot) = K(\cdot/h)/h$, and $K(\cdot)$ is a kernel function with the support $[-1, 1]$. In the minimization problem (3), the first summation is the weighted sum of squares of the estimation errors $\{y_{ij} - [g(x_{ij}) + \xi_i + \gamma(x_{ij})]\}$, after $g(x_{ij})$ is replaced by the linear approximation $\beta_0 + \beta_1(x_{ij} - s)$ in the small neighborhood $[s-h, s+h]$, and the weights are determined by the kernel function $K(\cdot)$. For the two random-effects terms $\xi_i \mathbf{1}_{n_i}$ and $\boldsymbol{M}_i(s)\boldsymbol{\eta}(s)$ included in the summation, two penalty terms $\sum_{i=1}^{m} \xi_i^2/\sigma_\xi^2$ and $\frac{1}{\sigma_\eta^2}\boldsymbol{\eta}(s)^T\boldsymbol{\eta}(s)$ have been used in (3) to regulate the random-effects. A similar penalized local kernel least squares estimation procedure has been discussed in the longitudinal data analysis literature in cases when within-profile data correlation exists but observations from different profiles are assumed independent (Wu and Zhang 2002). The minimization problem (3) is a generalization of the one discussed in Wu and Zhang (2002), where the latter considers only one random-effect term to describe the within-profile data correlation while the former considers two random-effects terms for describing both the within-profile and the between-profile data correlation.

For the minimization problem (3), the solutions to $\boldsymbol{\beta}$, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}(s)$ can be obtained by solving the so-called mixed-model equation (cf., Zhang et al. 1998), and the solutions are

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= [\boldsymbol{B}\boldsymbol{E}^{-1}\boldsymbol{D} - \boldsymbol{A}]^{-1}[\boldsymbol{B}\boldsymbol{E}^{-1}\boldsymbol{F} - \boldsymbol{C}], \\ \widehat{\boldsymbol{\eta}}(s) &= [\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B} - \boldsymbol{E}]^{-1}[\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{C} - \boldsymbol{F}], \\ \widehat{\xi_i} &= p_i(u_i - \boldsymbol{L}_i\boldsymbol{Z}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{L}_i\boldsymbol{M}_i(s)\widehat{\boldsymbol{\eta}}(s)), \qquad \text{for } i = 1, 2, \ldots, m, \end{aligned} \tag{4}$$

where $\boldsymbol{B} = \sum_{i=1}^{m} \boldsymbol{P}_i\boldsymbol{M}_i(s)$, $\boldsymbol{P}_i = \boldsymbol{Z}_i^T\boldsymbol{N}_i$, $\boldsymbol{Q}_i = \boldsymbol{M}_i^T(s)\boldsymbol{N}_i$, $\boldsymbol{N}_i = \boldsymbol{K}_i - p_i\boldsymbol{K}_i\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T\boldsymbol{K}_i$, $\boldsymbol{L}_i = (K_h(x_{i1} - s), K_h(x_{i2} - s), \ldots, K_h(x_{in_i} - s))$, $p_i = (w_i + \frac{\sigma_\varepsilon^2}{\sigma_\xi^2})^{-1}$, $w_i = \sum_{j=1}^{n_i} K_h(x_{ij} - s)$, $u_i = \sum_{j=1}^{n_i} y_{ij}K_h(x_{ij} - s)$,

$$\boldsymbol{A} = \sum_{i=1}^{m} \boldsymbol{P}_i\boldsymbol{Z}_i, \quad \boldsymbol{B} = \sum_{i=1}^{m} \boldsymbol{P}_i\boldsymbol{M}_i(s), \quad \boldsymbol{C} = \sum_{i=1}^{m} \boldsymbol{P}_i\boldsymbol{y}_i,$$

$$\boldsymbol{D}=\sum_{i=1}^{m}\boldsymbol{Q}_i\boldsymbol{Z}_i,\quad \boldsymbol{E}=\sum_{i=1}^{m}\boldsymbol{Q}_i\boldsymbol{M}_i(s)+\frac{\sigma_\varepsilon^2}{\sigma_\eta^2}\boldsymbol{I}_n,\quad \boldsymbol{F}=\sum_{i=1}^{m}\boldsymbol{Q}_i\boldsymbol{y}_i.$$

In (3) and (4), the variances $\sigma_\xi^2, \sigma_\eta^2$, and σ_ε^2 are assumed known. In practice, however, they are usually unknown in advance. In such cases, model (2) can be regarded as a semiparametric model, and we suggest the following iterative procedure for its estimation:

Step 1. At the given point $s \in [0,1]$ considered in (3), set the initial values for σ_ξ^2, σ_η^2 and σ_ε^2, denoted as $\sigma_\xi^2(s)^{(0)}$, $\sigma_\eta^2(s)^{(0)}$ and $\sigma_\varepsilon^2(s)^{(0)}$.

Step 2. In the kth iteration, for $k \geq 0$, compute estimates of $\boldsymbol{\beta}$, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}(s)$ by (4), where the values of σ_ξ^2, σ_η^2 and σ_ε^2 are specified as those obtained in the previous iteration. The resulting estimates are denoted as $\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{\eta}}^{(k)}(s)$ and $\widehat{\boldsymbol{\xi}}^{(k)} = (\widehat{\xi_1}^{(k)}, \widehat{\xi_2}^{(k)}, \ldots, \widehat{\xi_m}^{(k)})^T$.

Step 3. Based on $\widehat{\boldsymbol{\beta}}^{(k)}$, $\widehat{\boldsymbol{\xi}}^{(k)}$ and $\widehat{\boldsymbol{\eta}}^{(k)}(s)$, update the estimates of σ_ξ^2, σ_η^2 and σ_ε^2 by

$$\begin{aligned}
\widehat{\sigma}_\varepsilon^2(s)^{(k)} &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i-1}\left(\boldsymbol{y}_i-\boldsymbol{Z}_i\widehat{\boldsymbol{\beta}}^{(k)}-\widehat{\xi}_i^{(k)}\mathbf{1}_{n_i}-\boldsymbol{M}_i(s)\widehat{\boldsymbol{\eta}}^{(k)}(s)\right)^T\times \\
&\quad \boldsymbol{K}_i\left(\boldsymbol{y}_i-\boldsymbol{Z}_i\widehat{\boldsymbol{\beta}}^{(k)}-\widehat{\xi}_i^{(k)}\mathbf{1}_{n_i}-\boldsymbol{M}_i(s)\widehat{\boldsymbol{\eta}}^{(k)}(s)\right), \\
\widehat{\sigma}_\xi^2(s)^{(k)} &= \frac{1}{m-1}(\widehat{\boldsymbol{\xi}}^{(k)})^T\widehat{\boldsymbol{\xi}}^{(k)}, \\
\widehat{\sigma}_\eta^2(s)^{(k)} &= \frac{1}{p-1}(\widehat{\boldsymbol{\eta}}^{(k)}(s))^T\widehat{\boldsymbol{\eta}}^{(k)}(s).
\end{aligned}$$

Step 4. Repeat Steps 2 and 3 until the following condition is met:

$$\max\left(|\widehat{\sigma}_\varepsilon^2(s)^{(k)}-\widehat{\sigma}_\varepsilon^2(s)^{(k-1)}|, |\widehat{\sigma}_\xi^2(s)^{(k)}-\widehat{\sigma}_\xi^2(s)^{(k-1)}|, |\widehat{\sigma}_\eta^2(s)^{(k)}-\widehat{\sigma}_\eta^2(s)^{(k-1)}|\right)\leq\epsilon,$$

where $\epsilon > 0$ is a pre-specified small number (e.g., $\epsilon = 10^{-4}$).

From the above algorithm, we can obtain estimates of $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\eta}(s), \sigma_\xi^2, \sigma_\eta^2$, and σ_ε^2 at the given point s, denoted as $\widehat{\boldsymbol{\beta}}(s)$, $\widehat{\boldsymbol{\xi}}(s)$, $\widehat{\boldsymbol{\eta}}(s)$, $\widehat{\sigma}_\xi^2(s)$, $\widehat{\sigma}_\eta^2(s)$ and $\widehat{\sigma}_\varepsilon^2(s)$. Based on these estimates, we can define the final estimate of $g(s)$ as

$$\widehat{g}(s) = \boldsymbol{e}_1^T\widehat{\boldsymbol{\beta}}(s) \tag{5}$$

where $\boldsymbol{e}_1 = (1,0)^T$. Similarly, the final estimates of $\sigma^2_\xi, \sigma^2_\eta$, and σ^2_ε can be defined as

$$\begin{aligned}
\widehat{\sigma}^2_\varepsilon &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\widehat{\sigma}^2_\varepsilon(x_{ij}), \\
\widehat{\sigma}^2_\xi &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\widehat{\sigma}^2_\xi(x_{ij}), \\
\widehat{\sigma}^2_\eta &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\widehat{\sigma}^2_\eta(x_{ij}).
\end{aligned} \tag{6}$$

After the final estimates $\widehat{\sigma}^2_\varepsilon, \widehat{\sigma}^2_\xi$, and $\widehat{\sigma}^2_\eta$ are obtained, they can be pluged into (4) to obtain estimates of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}(s)$ at each $s \in [0,1]$, denoted as $\widehat{\boldsymbol{\xi}}(s)$ and $\widehat{\boldsymbol{\eta}}(s)$, respectively. Then, the final estimates of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are defined as

$$\begin{aligned}
\widehat{\boldsymbol{\xi}} &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\widehat{\boldsymbol{\xi}}(x_{ij}) \\
\widehat{\eta}_q &= \frac{1}{N_q}\sum_{l=1}^{N_q}\widehat{\eta}_q(s_{q,l}), \quad \text{for } q = 1,2,\ldots,p,
\end{aligned}$$

where $\{s_{q,l}, l = 1,2,\ldots,N_q\}$ are all design points in $\{x_{ij}, j = 1,2,\ldots,n_i, i = 1,2,\ldots,m\}$ whose neighborhoods have non-empty intersections with the qth subinterval $(d_{q-1}, d_q]$, N_q is the total number of such design points, for $q = 1,2,\ldots,p$, and $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \widehat{\eta}_2, \ldots, \widehat{\eta}_p)^T$.

In the minimization problem (3) and its estimation procedure described above, the kernel function $K(\cdot)$ and the bandwidth h should be chosen in advance. Regarding $K(\cdot)$, it can be chosen to be the Epanechnikov kernel function because of its good theoretical properties (Epanechnikov 1969). Namely, we choose $K(x) = 0.75(1-x^2)I(|x| \le 1)$. Regarding the bandwidth h, we suggest using the following empirical formula, as discussed in Qiu et al. (2010):

$$h = c\overline{n}^{-1/5}\sqrt{\text{Var}(x)}, \tag{7}$$

where $\overline{n}$ and $\text{Var}(x)$ are the average number of design points and the variance of the design points within a profile, respectively, which can be estimated from an IC data, and c is a constant. From Qiu et al. (2010), c can be chosen in the interval $[1.0, 2.0]$. In all numerical examples in Sections 3 and 4, we choose $c = 1.5$.

2.2 Phase I monitoring of serially correlated univariate profiles

After the mixed-effects model (2) is estimated, we are ready to construct a control chart for Phase I profile monitoring. Note that besides $g(x_{ij})$, the random-effect term η_q on the right-hand-side of (2) also depends on the design points x_{ij}. Thus, to monitor the ith profile $\{(x_{ij}, y_{ij}), j = 1, 2, \ldots, n_i\}$, it is more reasonable to use $\frac{1}{n_i}\sum_{j=1}^{n_i}[y_{ij} - \widehat{g}(x_{ij}) - \widehat{\eta}_q]^2$ to measure the discrepancy between the observed profile and the estimated profile, compared to $\frac{1}{n_i}\sum_{j=1}^{n_i}[y_{ij} - \widehat{g}(x_{ij})]^2$. Based on this consideration, we suggest the following charting statistic for monitoring a total of m observed profiles at the beginning of Phase I SPC:

$$S_{TM} = \max_{i=1,2,\ldots,m} \frac{1}{n_i}\sum_{j=1}^{n_i}[y_{ij} - \widehat{g}(x_{ij}) - \widehat{\eta}_q]^2\,, \tag{8}$$

and the chart gives a signal when

$$S_{TM} \geq CL, \tag{9}$$

where $CL > 0$ is a control limit, and the subscript of S_{TM} denotes the fact that the mixed-effects model (2) has two mixed-effects terms. This chart will be called TM chart hereafter.

To determine the control limit CL, we need to compute the empirical distribution of S_{TM}. To this end, the following bootstrap procedure is considered, which is similar to the one discussed in Chatterjee and Qiu (2009). First, let us define the residuals $\widehat{\varepsilon}_{ij} = y_{ij} - \widehat{g}(x_{ij}) - \widehat{\xi}_i - \widehat{\eta}_q$, for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, m$. Then, we draw a sample of size $N = \sum_{i=1}^{m} n_i$ with replacement from the set of the original residuals $\{\widehat{\varepsilon}_{ij}, j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, m\}$. The re-sampled residuals are denoted as $\{\widehat{\varepsilon}^*_{ij}\}$, based on which we define the bootstrap observations by $y^*_{ij} = \widehat{g}(x_{ij}) + \widehat{\xi}_i + \widehat{\eta}_q + \widehat{\varepsilon}^*_{ij}$, for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, m$. From the bootstrap observations, we can compute the bootstrap estimates of $g(x_{ij}), \xi_i$, and η_q by (4), denoted as $\widehat{g}^*(x_{ij}), \widehat{\xi}^*_i$, and $\widehat{\eta}^*_q$, respectively, for all i, j, and q. Based on these estimates, the value of the charting statistic S_{TM} can be computed by (8), denoted as S^*_{TM}. The bootstrap re-sampling is then repeated for B times, and B values of S^*_{TM} can be obtained accordingly. The $(1-\alpha)$th quartile of the B values of S^*_{TM} can then be used to approximate the control limit CL, where α is a given IC false alarm rate (FAR).

The entire proposed Phase I profile monitoring procedure is iterative and summarized below.

- At the beginning, model (2) is estimated from a total of m profiles under monitoring. The charting statistic S_{TM} is computed by (8). If (9) is not true, then the entire iterative procedure for Phase I SPC stops and we claim that all m profiles are IC. Otherwise, the profile with the largest value of $\frac{1}{n_i}\sum_{j=1}^{n_i}\left[y_{ij}-\widehat{g}(x_{ij})-\widehat{\eta}_q\right]^2$ is declared to be an outlier and deleted from the subsequent profile monitoring.

- At the kth iteration, for $k>1$, there are $k-1$ profiles deleted in the previous iterations and $m-k+1$ profiles remaining for profile monitoring. Model (2) is estimated from these $m-k+1$ profiles, and the control limit CL is computed from them as well by the bootstrap procedure described above. Then, the charting statistic S_{TM} is computed by (8) based on the $m-k+1$ profiles. If (9) is not true, then the entire iterative procedure stops and we claim that all $m-k+1$ profiles are IC. Otherwise, declare the profile with the largest value of $\frac{1}{n_i}\sum_{j=1}^{n_i}[y_{ij}-\widehat{g}(x_{ij})-\widehat{\eta}_q]^2$ as an outlier and delete it from the subsequent profile monitoring.

- The iterative procedure continues until no outliers can be found.

3 Simulation Study

In this section, we study the numerical performance of the proposed TM chart (8)-(9), using different simulation examples. Without loss of generality, assume that there are $m=25$ profiles to monitor in a Phase I monitoring, all profiles are defined in the design interval $[0,1]$, each profile has $n_i=25$ observations, for $i=1,2,\ldots,m$, and $p=20$ in model (2). Also, the design points x_{ij} of the profiles are generated from the distribution Uniform$(0,1)$, for all $j=1,2,\ldots,n_i, i=1,2,\ldots,m$. To study the impact of within-profile and between-profile data correlation on the performance of the chart (8)-(9), the following four cases are considered for specifying the three random terms in model (2): for $j=1,2,\ldots,n_i$, $i=1,2,\ldots,m$, and $q=1,2,\ldots,p$,

(I) $\xi_i \overset{\text{i.i.d.}}{\sim} N(0,0.5^2)$, $\sigma_\gamma^2=0$, and $\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0,0.5^2)$;

(II) $\sigma_\xi^2=0$, $\eta_q \overset{\text{i.i.d.}}{\sim} N(0,0.5^2)$, and $\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0,0.5^2)$;

(III) $\xi_i \overset{\text{i.i.d.}}{\sim} N(0, 0.1^2)$, $\eta_q \overset{\text{i.i.d.}}{\sim} N(0, 0.5^2)$, and $\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, 0.1^2)$;

(IV) $\xi_i \overset{\text{i.i.d.}}{\sim} t(3)/(10\sqrt{3})$, $\eta_q \overset{\text{i.i.d.}}{\sim} t(3)/(2\sqrt{3})$, and $\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} t(3)/(10\sqrt{3})$.

Case (I) represents a situation in which within-profile data correlation exists but between-profile data correlation does not exist. Case (II) considers a situation in which within-profile data correlation does not exist but between-profile data correlation exists. Cases (III) and (IV) consider situations when both within-profile and between-profile data correlation exist but the random variables have different distributions. For the IC mean function $g(x)$ in model (2), we consider the following three cases:

Linear profile: $g(x) = 1 + 2x$,

Nonlinear profile 1: $g(x) = 1 + 2x + \sin(2\pi x)$,

Nonlinear profile 2: $g(x) = 1 + 2x + 5x^2 + \sin(2\pi x)$.

Among the above three profiles, the first one is linear, the second one consists of a linear component and a non-linear component, while the third one adds a quadratic term to the second profile. These three cases are denoted as Linear, Nonlinear 1 and Nonlinear 2, respectively, in this section. For evaluating the OC performance of the proposed method, we consider two types of shifts: a step shift δ_1 that can change among 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 and 2.0, and a shift δ_2 in the profile slope which can also change among 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 and 2.0. When a profile has a step shift δ_1, its mean function changes from $g(x)$ to $g(x) + \delta_1$. When a profile has a slope shift δ_2, its mean function changes from $g(x)$ to $g(x) + \delta_2 x$.

As discussed in Section 1, so far we cannot find any existing univariate nonparametric profile monitoring methods that can accommodate between-profile data correlation. But, there are some existing Phase II profile monitoring methods that can accommodate within-profile data correlation. To compare our proposed method with some existing profile monitoring methods, in this section we consider the following three representative Phase II profile monitoring methods that take into account the within-profile data correlation and adapt them to Phase I profile monitoring. The first

existing method was proposed by Jensen et al. (2008) based on the following linear mixed-effects (LM) model:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{10}$$

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$, $\boldsymbol{X}_i$ is the design matrix of the ith profile, $\boldsymbol{\beta}$ is the regression coefficient vector, $\boldsymbol{Z}_i$ is a design matrix of r covariates for the random-effects term, $\boldsymbol{b}_i \sim N_r(0, \boldsymbol{D})$ is the coefficient vector of the random-effects terms with the diagonal covariance matrix $\boldsymbol{D}$, and $\boldsymbol{\varepsilon}_i$ is the vector of the i.i.d. random errors with mean 0 and variance σ_ε^2 of the ith profile. In the above model, we need to specify r covariates for defining the random-effects terms. To make it consistent with the random-effects term in model (2), we specify $\boldsymbol{Z}_i = (1, 1, \ldots, 1)^T$. In such cases, $\boldsymbol{b}_i$ is a scalar random variable. Then, $\boldsymbol{\beta}$ can be estimated by

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i\right),$$

where $\boldsymbol{V}_i = \boldsymbol{X}_i\boldsymbol{D}\boldsymbol{X}_i^T + \sigma_\varepsilon^2\boldsymbol{I}$. So, the charting statistic for Phase I profile monitoring can be defined to be

$$S_{LM} = \max_{i=1,2,\ldots,m} \frac{1}{n_i}\Big(\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}\Big)^T \Big(\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}\Big),$$

and the chart gives a signal when S_{LM} exceeds a control limit. This chart will be labeled as LM, denoting the fact that it is based on a linear mixed-effects modeling.

The second existing method considered here was originally proposed by Zhang et al. (2014), which described the within-profile correlation by the following Gaussian process (GP) model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \theta_i(x_{ij}) + \varepsilon_{ij}, \quad \text{for} \quad j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, m, \tag{11}$$

where $\theta_i(x)$ is a realization of a Gaussian process with mean 0 and covariance function $\text{cov}(\theta_i(x'), \theta_i(x'')) = \sigma_\theta^2 K_{SE}(r_{x',x''})$, for any $x', x'' \in [0, 1]$, where σ_θ^2 is the signal variance, $K_{SE}(r_{x',x''}) = \exp(-r_{x',x''}^2)$, and $r_{x',x''} = |x' - x''|$. The charting statistic for Phase I profile monitoring is defined to be

$$S_{LGM} = \max_{i=1,2,\ldots,m} \frac{1}{n_i} \sum_{j=1}^{n_i} \Big(y_{ij} - \widehat{\beta}_0 - \widehat{\beta}_1 x_{ij}\Big)^2,$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the estimates of β_0 and β_1 defined in Zhang et al. (2014). This chart will be labeled as LGM, denoting the fact that it is based on a linear Gaussian process mixed-effects model.

The third existing method considered here was suggested by Qiu et al. (2010), which was based on the following nonparametric mixed-effects modeling:

$$y_{ij} = g(x_{ij}) + f_i(x_{ij}) + \varepsilon_{ij}, \quad \text{for} \quad j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, m, \tag{12}$$

where $f_i(x)$ is a zero-mean random process for describing the difference between the population profile mean function $g(x)$ and the mean function of the ith profile. The charting statistic for Phase I profile monitoring is defined to be

$$S_{NM} = \max_{i=1,2,\ldots,m} \frac{1}{n_i} \sum_{j=1}^{n_i} \left[y_{ij} - \widehat{g}(x_{ij}) \right]^2,$$

where $\widehat{g}(x)$ is the estimate of $g(x)$ defined in Qiu et al. (2010). This chart will be labeled as NM, denoting the fact that it is based on a nonparametric mixed-effects modeling.

Among these three existing methods, both LM and LGM are for monitoring linear profiles with possible within-profile data correlation, and NM is designed for monitoring nonparametric profiles with possible within-profile data correlation. All these three methods assume that the observed data from different profiles are independent of each other. So, they may not be able to accommodate between-profile data correlation well.

3.1 Performance of the estimated model of (2)

First, we study the performance of the estimated model of (2) by the estimation procedure (3)-(6). The performance is measured by

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[g(x_{ij}) - \widehat{g}(x_{ij}) \right]^2.$$

In cases when $n_i = 25$, $m = 25, 50$ or 100, and $p = 5, 10, 15, 20, 25, 30$ or 50, the MSE values of the estimated model based on 100 replicated simulations are presented in Figure 1. From the plots in the figure, it can be seen that (i) MSE values for estimating linear profile models are generally

smaller than those for estimating nonlinear profile models, (ii) MSE values would become smaller when m increases, and (iii) MSE values are generally smaller when p increases. The first conclusion is intuitively reasonable because nonlinear profile models are generally more difficult to estimate than their linear counterparts, and consequently the related MSE values for estimating nonlinear profile models would be larger. The second conclusion implies that the estimated model would be statistically consistent. The third conclusion is also intuitively reasonable because Model (2) is an approximation of Model (1) (cf., the related discussion in Section 2), and the approximation would be better if p is chosen larger. However, when p is chosen too large, the computation involved in the model estimation would be extensive. Based on this and other simulation examples with different sample sizes (not presented here), it seems that the estimated model of (2) would perform reasonably well when p is chosen to be 20 or larger.

In cases when $m = 25$, $n_i = 25$, for all i, $p = 20$, and the random terms of Model (2) belong to case (II), the 95% pointwise confidence intervals for $g(x)$ are presented in Figure 2 for monitoring both the linear profile (1st panel) and the nonlinear profiles (2nd and 3rd panels). From the first panel, it can be seen that the estimate $\widehat{g}(x)$ is almost identical to the true function $g(x)$ when $g(x)$ is linear, and the 95% pointwise confidence intervals perform well too. From the 2nd and 3rd panels, both the estimates $\widehat{g}(x)$ and the 95% pointwise confidence intervals of $g(x)$ perform well at all places except some places around the peaks and the valleys of the curves of $g(x)$. Basically, the peaks and the valleys of $g(x)$ are smoothed out to a certain degree, which is well known in the local smoothing literature (Qiu 2005, Chapter 2). Even in such cases, the 95% confidence intervals cover $g(x)$ at almost all places. Therefore, $\widehat{g}(x)$ provides a reasonably good estimate of $g(x)$. The confidence intervals in other cases reveal similar patterns, and thus are omitted.

3.2 IC performance of the chart (8)-(9)

In this part, we present some simulation results about the IC performance of the four Phase I profile monitoring methods TM, LM, LGM and NM. To this end, their control limits are determined based on $10,000$ replicated simulations with the nominal IC FAR value of $\alpha = 0.05$ and the bootstrap

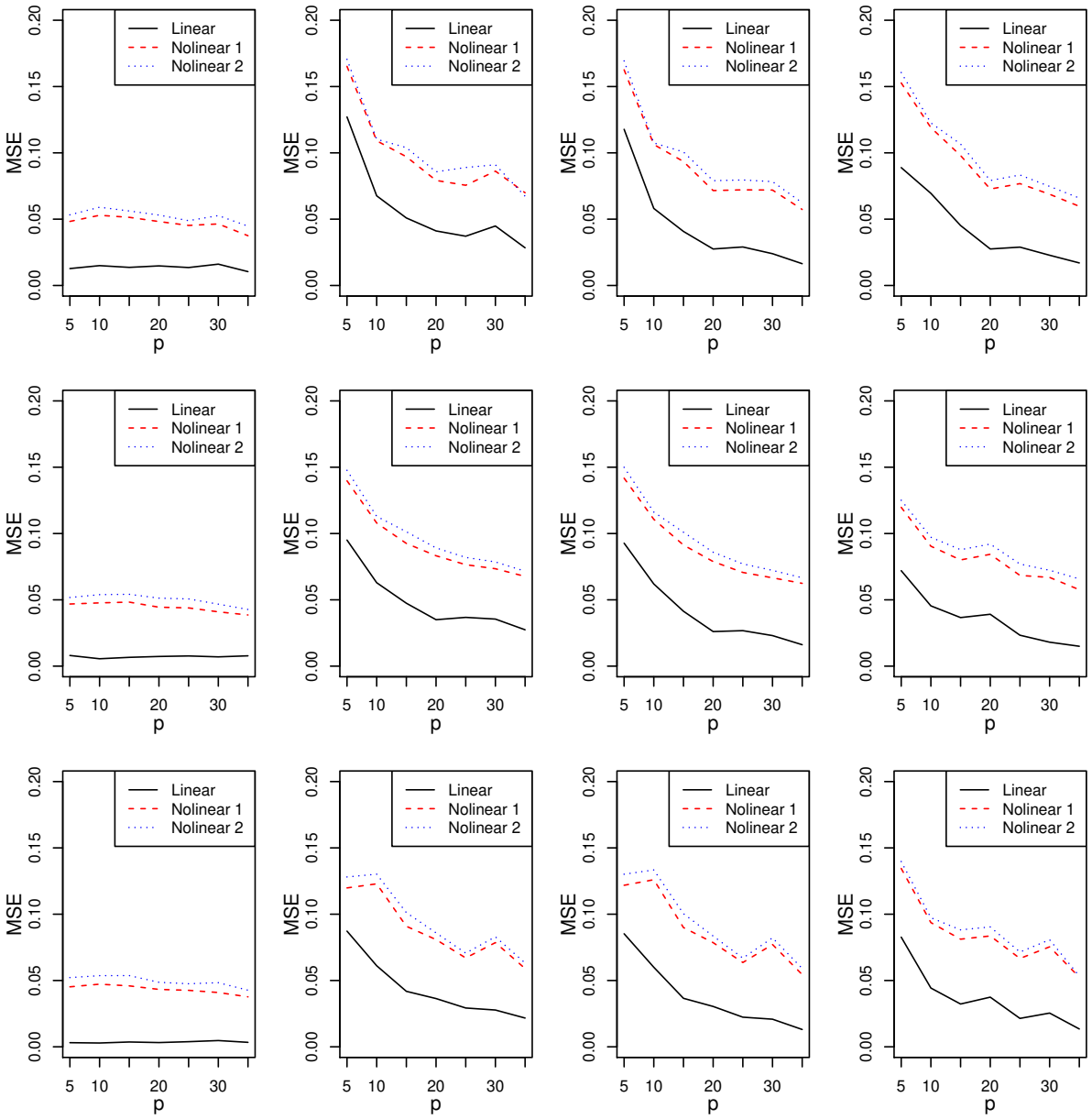

Figure 1: MSE values of the estimated model of (2) in cases when $n_i = 25$, $m = 25$ (1st column), 50 (2nd column), or 100 (3rd column), $p = 5, 10, 15, 20, 25, 30$ or 50, and the random terms are in cases (I) (1st row), (II) (2nd row), (III) (3rd row), and (IV) (4th row).

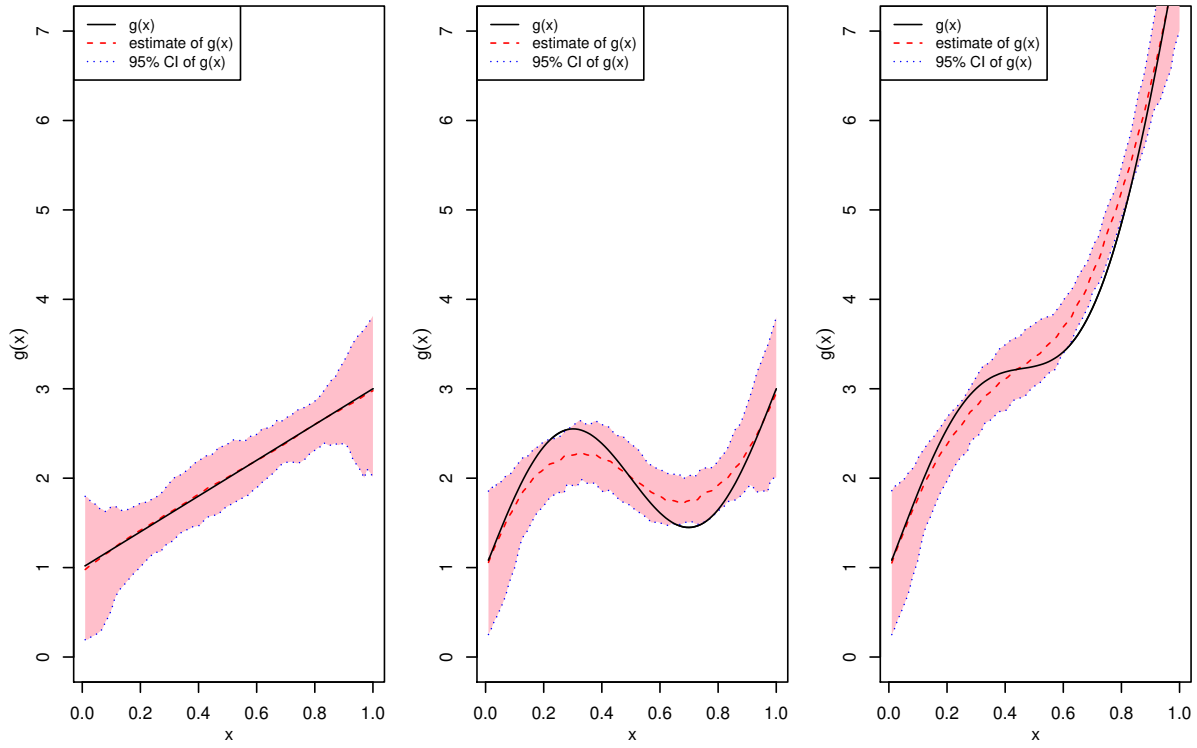

Figure 2: The 95% pointwise confidence intervals for $g(x)$ when $g(x)$ is the linear profile (1st panel), the nonlinear profile 1 (2nd panel) or the nonlinear profile 2 (3rd panel), $m = 25$, $n_i = 25$, for all i, $p = 20$, and the random terms of model (2) belong to case (II).

sample size of $B = 10,000$. Then, the actual IC FAR of each method, defined to be the proportion of the repeated simulations in which at least one signal is obtained, is calculated based on $1,000$ replicated simulations. The actual IC FAR values of the four methods in different cases considered are presented in Table 1. From the table, it can be seen that (i) the actual IC FAR values of LM, LGM and NM are reasonably close to the nominal IC FAR value of $\alpha = 0.05$ in case (I) when there is no between-profile data correlation, (ii) the actual IC FAR values of LM, LGM and NM are far away from $\alpha = 0.05$ in cases (II), (III) and (IV) when there is a substantial between-profile data correlation, and (iii) the actual IC FAR values of TM are reasonably close to $\alpha = 0.05$ in all cases considered. This example shows that the proposed method TM has a reliable IC performance when a substantial between-profile data correlation is present, and the three alternative methods LM, LGM and NM would not be reliable in such cases.

Table 1: Actual IC FAR values of the charts TM, LM, LGM and NM in cases when $\alpha = 0.05$, $m = 25, n_i = 25$, $p = 5, 10$ and 20, and the random terms of model (2) belong to cases (I), (II), (III) and (IIV).

		Linear Profile				Nonlinear Profile 1				Nonlinear Profile 2			
p		LM	LGM	NM	TM	LM	LGM	NM	TM	LM	LGM	NM	TM
	(I)	0.049	0.048	0.046	0.045	0.058	0.057	0.047	0.053	0.061	0.060	0.049	0.056
5	(II)	0.210	0.201	0.223	0.061	0.252	0.250	0.206	0.060	0.258	0.256	0.193	0.062
	(III)	0.314	0.316	0.314	0.071	0.322	0.320	0.264	0.072	0.332	0.332	0.257	0.076
	(IV)	0.183	0.180	0.181	0.032	0.196	0.195	0.182	0.033	0.161	0.161	0.166	0.032
	(I)	0.052	0.053	0.048	0.047	0.061	0.062	0.046	0.052	0.068	0.069	0.052	0.057
10	(II)	0.181	0.178	0.229	0.061	0.204	0.202	0.175	0.062	0.182	0.180	0.194	0.061
	(III)	0.306	0.308	0.308	0.071	0.274	0.274	0.263	0.074	0.227	0.228	0.266	0.075
	(IV)	0.160	0.159	0.189	0.030	0.151	0.152	0.147	0.032	0.141	0.143	0.141	0.031
	(I)	0.056	0.055	0.053	0.046	0.065	0.063	0.047	0.054	0.078	0.075	0.052	0.055
20	(II)	0.119	0.119	0.186	0.065	0.167	0.166	0.175	0.061	0.178	0.177	0.161	0.061
	(III)	0.214	0.232	0.300	0.074	0.263	0.260	0.228	0.075	0.230	0.231	0.218	0.078
	(IV)	0.145	0.139	0.206	0.045	0.147	0.147	0.223	0.041	0.152	0.159	0.224	0.040

3.3 OC performance

In this subsection, we investigate the OC performance of the related control charts. To evaluate the OC performance of a Phase I control chart, it has been well discussed in the literature that potential OC data can contaminate the design of the chart since some OC data could potentially be used in determining the control limit of the chart (cf., Yeh et al. 2009). To overcome this difficulty and objectively measure the OC performance of the chart, some authors suggested using multiple OC performance measures. For instance, Yeh et al. (2009) and Zang and Qiu (2018) used the following three criteria for measuring the OC performance of their suggested Phase I charts: alarm probability (AP), fraction of correctly classified profiles (FCC), and false positive rate (FPR). AP is defined to be the proportion of the repeated simulations in which at least one profile is detected as OC. FCC is defined to be the proportion of all profiles that are correctly classified as IC or OC by a control chart. FPR is defined to be the proportion of the detected OC profiles which are actually IC. From their definitions, it is obvious that the larger the first two criteria or the smaller the third criterion, the better. Also, if some OC profiles are used in determining the control limit of a chart, then the chart would become more conservative in the sense that less OC profiles will be detected. In such cases, AP and FPR would both tend to be smaller, and FPR could be larger or smaller depending on the specific cases. Thus, at least one of these three measures (e.g., at least AP) can catch this scenario. In this subsection, the three OC performance measures AP, FCC and FPR are used, and we consider cases when $n_i = 25$, for all i, $m = 25$, $p = 20$, $\alpha = 0.05$, and the number of OC profiles is 5. The OC profiles in a given set of $m = 25$ profiles for Phase I monitoring are those at the 2nd, 4th, 6th, 8th, and 10th positions. In this study, because the results of LM and LGM are almost identical in all cases considered, the results of LGM will not be presented.

In both the linear and nonlinear profile cases, we let each of the step shift δ_1 and the slope shift δ_2 changes among 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 and 2.0. To make the charts comparable, the control limits of LM and NM have been adjusted in each case considered so that their actual IC FAR values reach the level of $\alpha = 0.05$. The calculated values of AP, FCC and FPR of the three

charts TM, LM and NM in cases (I)-(IV) are presented in Figures 3-6, respectively, which are based on 1,000 replicated simulations. From the figures, it can be seen that (i) the three methods perform almost the same in case (I) when there is no between-profile data correlation, (ii) TM outperforms LM and NM in cases (II)-(IV) when there is a substantial between-profile data correlation, and (iii) NM and LM performs similarly for monitoring linear profiles and NM outperforms LM when monitoring nonlinear profiles. So, this example shows that our proposed method TM outperforms the two representative existing linear and nonparametric profile monitoring methods LM and NM in cases when between-profile data correlation is present.

4 A Real Data Example

In this section, we illustrate the application of the proposed Phase I profile monitoring chart TM, using the Vertical Density Profiles (VDP) dataset discussed in Walker and Wright (2002). This dataset is about a manufacturing process of particle boards, whose density properties are critical to their machinability and thus require careful control and monitoring. Density readings are collected by a laser device at fixed vertical depths of a board. So, observations from a single board can be regarded as a profile. In the dataset, there are 24 profiles. In each profile, design points are fixed at $x_j = 0.002 \times j$ inches, for $j = 0, 1, \ldots, 313$. The original dataset is shown in the left panel of Figure 7. A more detailed introduction about this dataset and the manufacturing of particle boards can be found in Walker and Wright (2002) and Young et al. (1999). This dataset has been analyzed previously by several authors. Walker and Wright (2002) compared multiple VDP profiles using additive modeling and smoothing splines. Williams et al. (2007) focused on the Phase I analysis using the nonlinear regression curve estimation. As indicated by these authors, the VDP data exhibit a significant amount of within-profile positive autocorrelation, although they have not considered such autocorrelation in their analysis. None of these papers consider the possible between-profile data correlation. To explore the possible between-profile data correlation, we present the observations of the 24 consecutive profiles at $x_j = 0.1, 0.3$ and 0.5 in the right panel

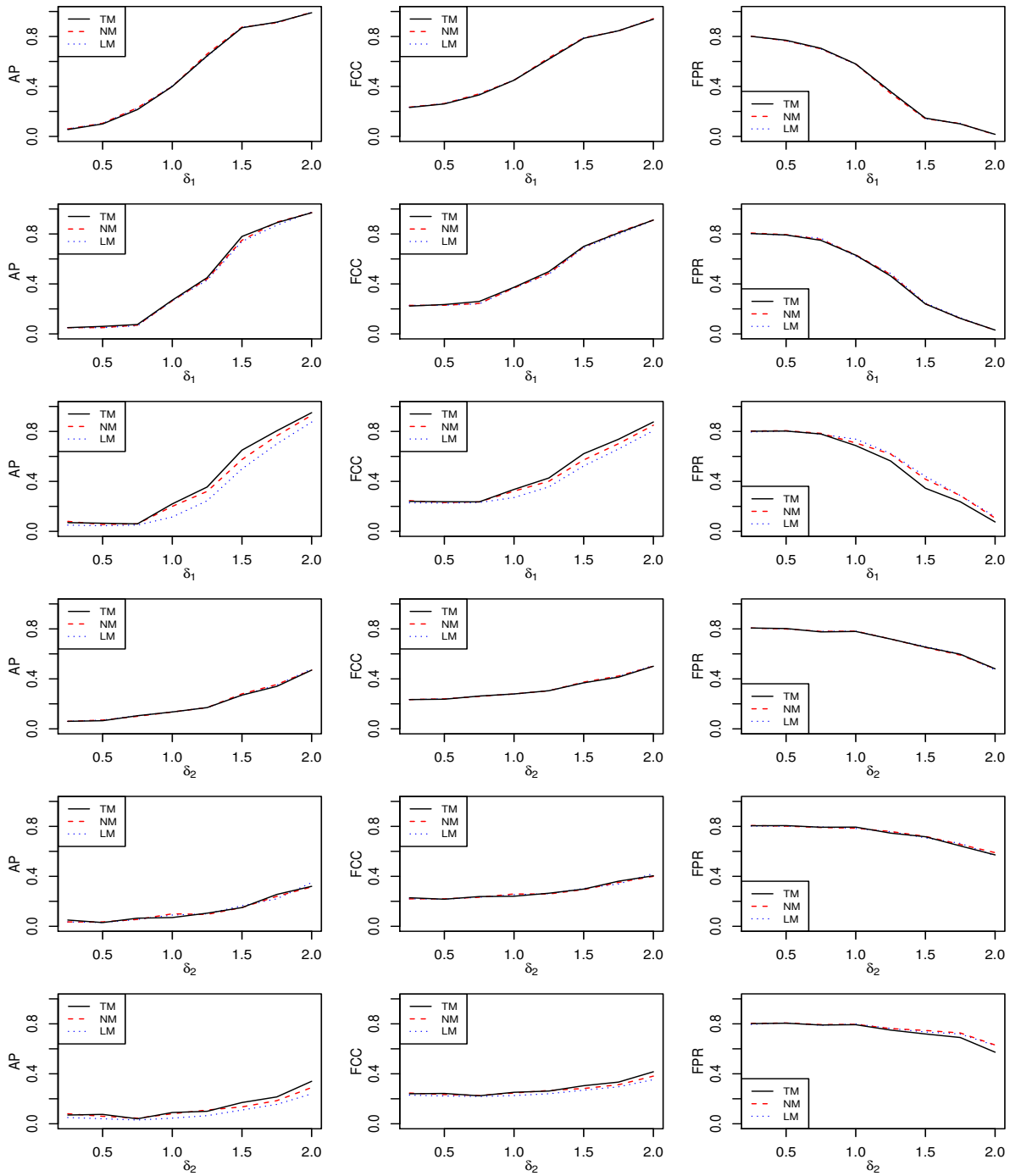

Figure 3: Calculated values of AP, FCC and FPR of the three charts TM, LM and NM in Case (I) for detecting step shifts (first three rows) of size δ_1 and slope shifts (last three rows) of size δ_2 in the linear (1st and 4th rows), nonlinear profile 1 (2nd and 5th rows), and nonlinear profile 2 (3rd and 6th rows) models.

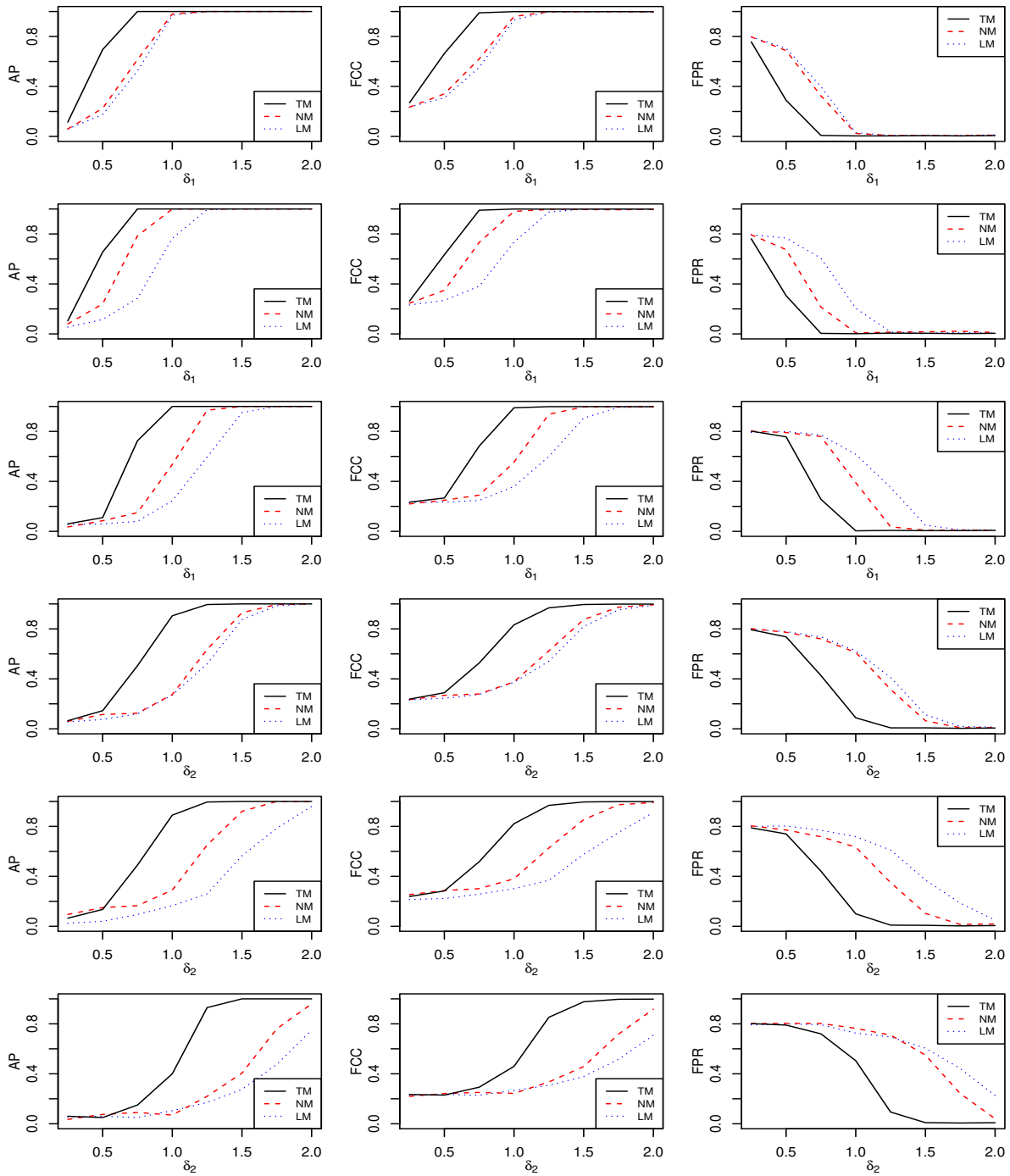

Figure 4: Calculated values of AP, FCC and FPR of the three charts TM, LM and NM in Case (II) for detecting step shifts (first three rows) of size δ_1 and slope shifts (last three rows) of size δ_2 in the linear (1st and 4th rows), nonlinear profile 1 (2nd and 5th rows), and nonlinear profile 2 (3rd and 6th rows) models.

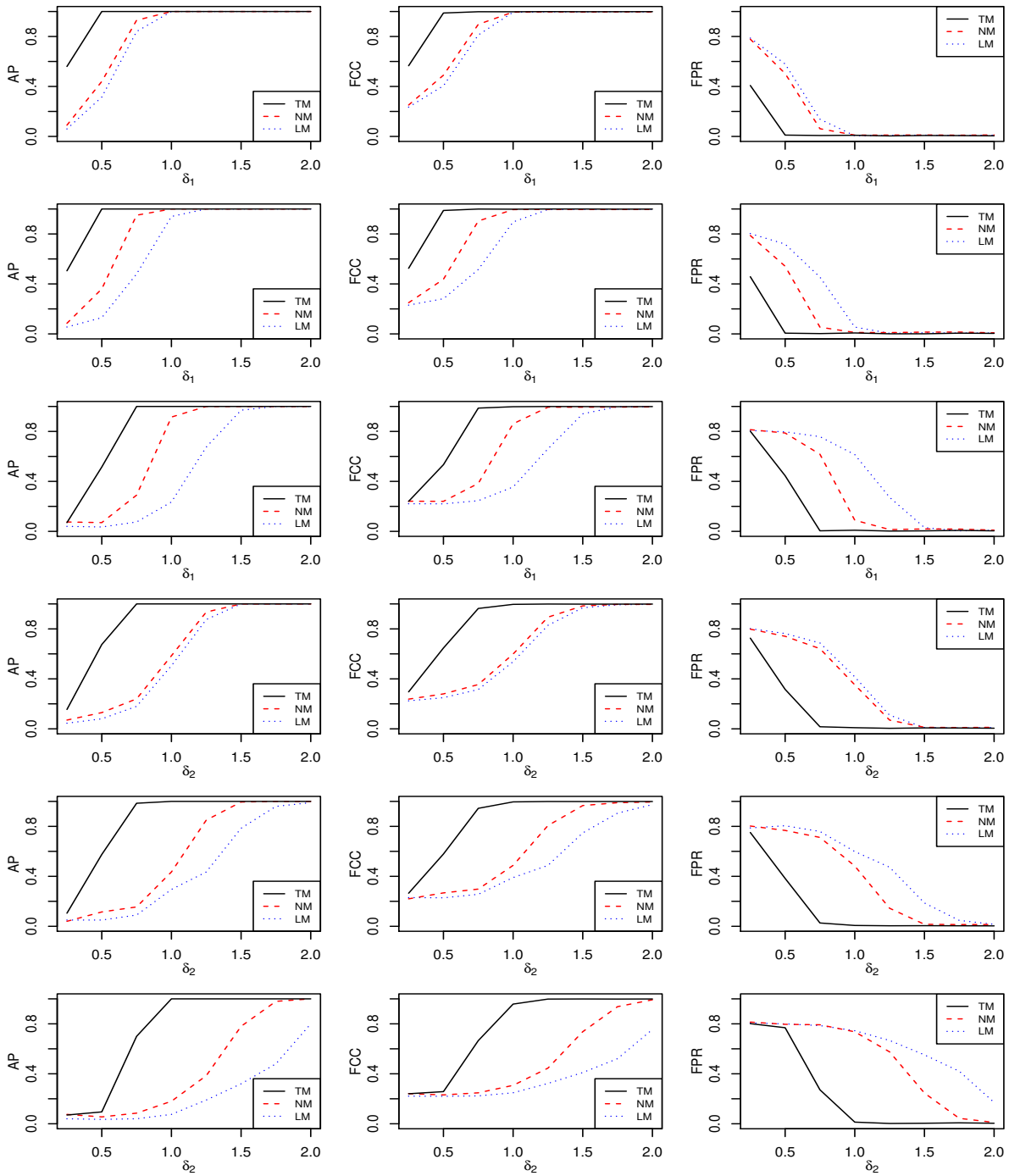

Figure 5: Calculated values of AP, FCC and FPR of the three charts TM, LM and NM in Case (III) for detecting step shifts (first three rows) of size δ_1 and slope shifts (last three rows) of size δ_2 in the linear (1st and 4th rows), nonlinear profile 1 (2nd and 5th rows), and nonlinear profile 2 (3rd and 6th rows) models.

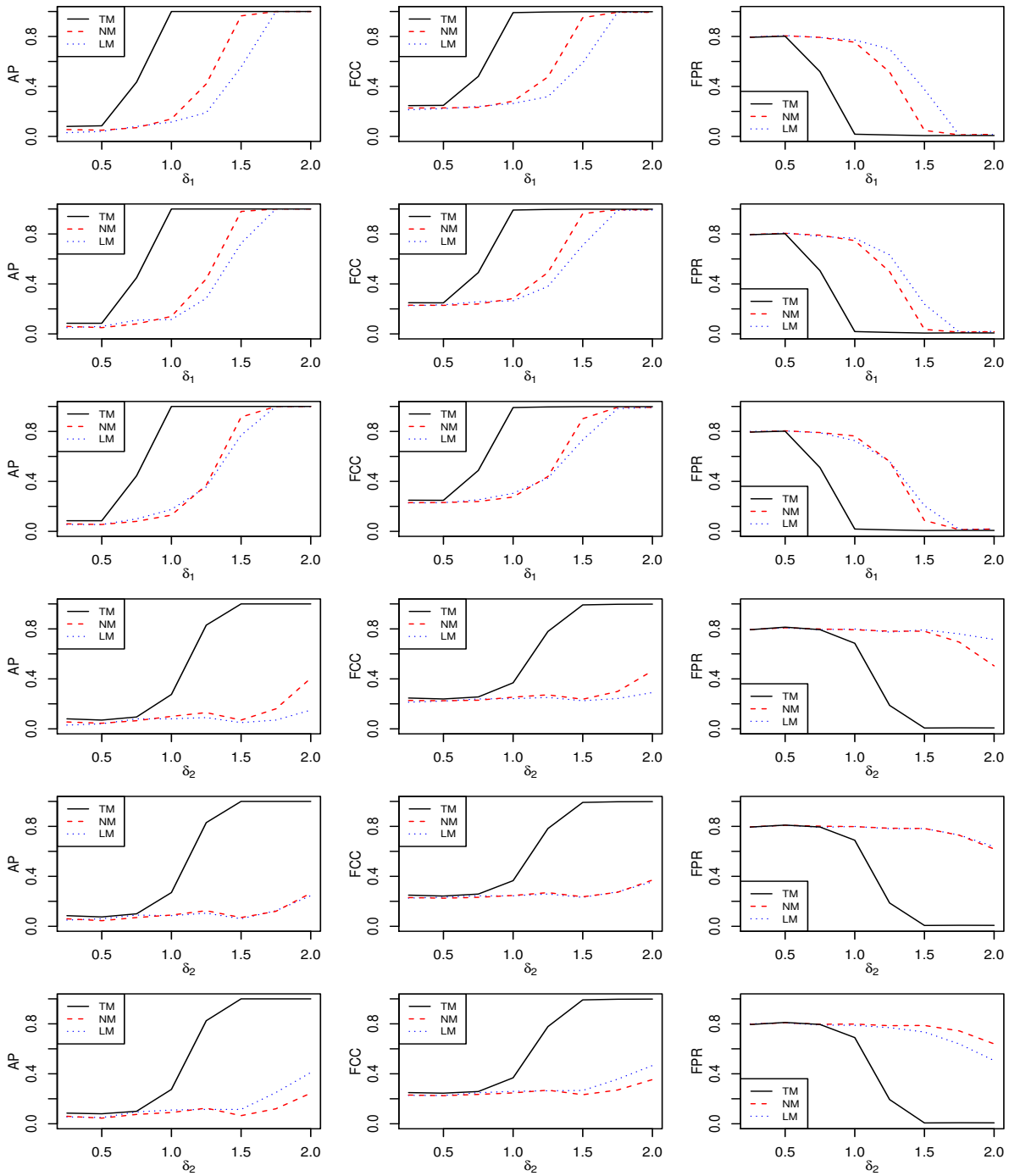

Figure 6: Calculated values of AP, FCC and FPR of the three charts TM, LM and NM in Case (IV) for detecting step shifts (first three rows) of size δ_1 and slope shifts (last three rows) of size δ_2 in the linear (1st and 4th rows), nonlinear profile 1 (2nd and 5th rows), and nonlinear profile 2 (3rd and 6th rows) models.

of Figure 7, where the y-axis denotes the density reading and the x-axis is the profile index. So, for instance, the solid curve in the plot denotes the observed density readings of the 24 profiles at the vertical depth of $x_j = 0.1$ (i.e., $j = 50$). From the plot, it seems that there is some evidence of autocorrelation (e.g., a data point on the solid curve tends to be relatively large if its previous data point on the same curve is relatively large). To investigate autocorrelation of each curve numerically, a total of 24 data points on the curve might be too small. To overcome this difficulty, we consider the set of all pairs $\mathcal{P} = \{(y_{i-1,j}, y_{ij}),\ \text{for all possible } i, j\}$. Namely, for a given observation y_{ij} on a specific profile, it is paired with the corresponding observation on the previous profile. Then, we compute the Pearson's correlation coefficient for $\mathcal{P}$ to be $r = 0.799$, which indicates a quite strong positive correlation between y_{ij} and $y_{i-1,j}$. The 95% confidence interval of the correlation coefficient computed by a bootstrap re-sampling approach with the bootstrap sample size of 10,000 is $[0.790, 0.808]$, which is completely above 0. Thus, it seems that between-profile data correlation is quite significant in this VDP data.

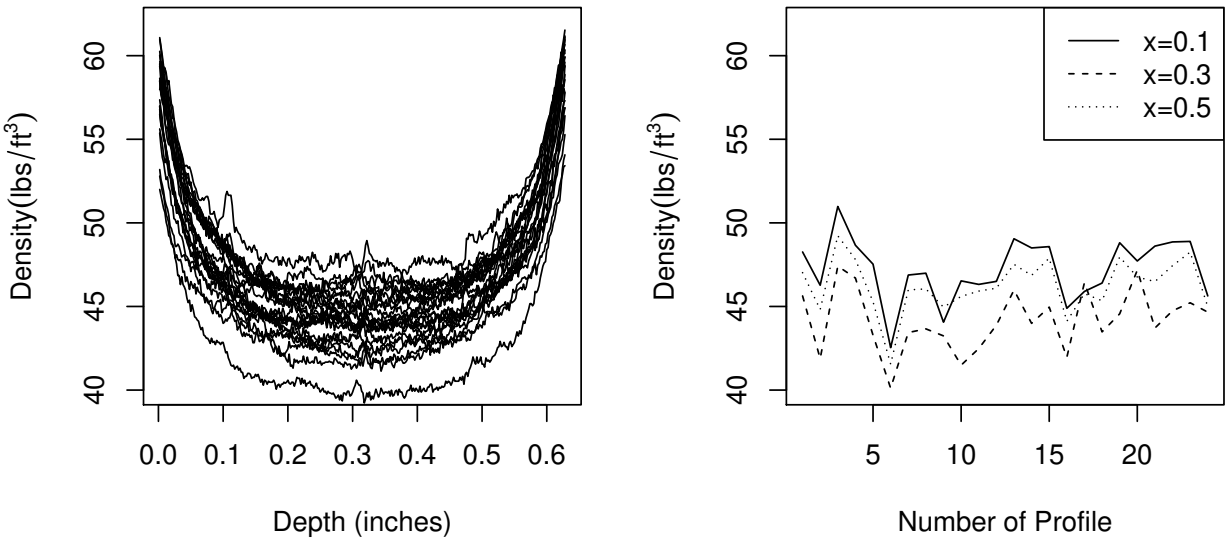

Figure 7: VDP data with 24 profiles of density readings at different vertical depths of particle boards (left panel), and density readings of the 24 profiles at the vertical depths of 0.1, 0.3 and 0.5 inches (right panel).

From the left panel of Figure 7, it seems that the profile at the bottom is quite different from

the other profiles, and the remaining profiles are quite similar. The profile at the bottom turns out to be the 6th profile in the VDP data. So, the last 12 profiles in the data are used as an IC dataset for IC model estimation and design of the related control charts, and the first 12 profiles are used for testing. Then, model (2) is estimated from the IC data, in which $m = 12$ and $n_i = 314$ for all i. As suggested in the previous section, we choose $p = 20$. Then, the estimated mean profile function $\widehat{g}(x)$ is shown in Figure 8 by the thin solid curve, along with the 95% pointwise confidence intervals for $g(x)$ that are shown by the shaded region around $\widehat{g}(x)$. The estimates of σ_ξ^2, σ_η^2 and σ_ε^2 are computed by (6) to be $1.681, 1.772$, and 0.548, respectively. Next, we use a bootstrap procedure with the bootstrap sample size of $B = 100$, in which the bootstrap samples are generated in the way as described in the paragraph immediately below the expression (9). Then, 100 bootstrap estimates of σ_η^2 can be obtained from the 100 bootstrap samples. The 95% bootstrap confidence interval for σ_η^2 is [1.574,2.010], which is completely above 0. So, we can conclude that σ_η^2 is significantly larger than 0, or there is a significant between-profile data correlation. This result is consistent with the test for the Pearson's correlation coefficient discussed in the previous paragraph. The control limits of the charts TM, LM, LGM and NM are computed by the bootstrap procedure with the bootstrap sample size of $B = 100$ as well. They turn out to be 13.498, 21.991, 14.485 and 16.348, respectively. Then, the four control charts are used to monitor the first 12 profiles in the VDP data that are shown in Figure 8 by the gray curves. Both the TM and NM methods detect the 6_{th} profile (bold solid curve in the plot) as an outlier, and the LM and LGM methods detect both the 6_{th} profile and the 10_{th} profile (bold dashed curve in the plot) as outliers. In this example, the results from the TM and NM methods are the same, while both methods LM and LGM detect an extra outlier. Because the methods LM and LGM assume that the population mean profile function is linear, which is obviously invalid in this example, we believe that the results from TM and NM are more reliable.

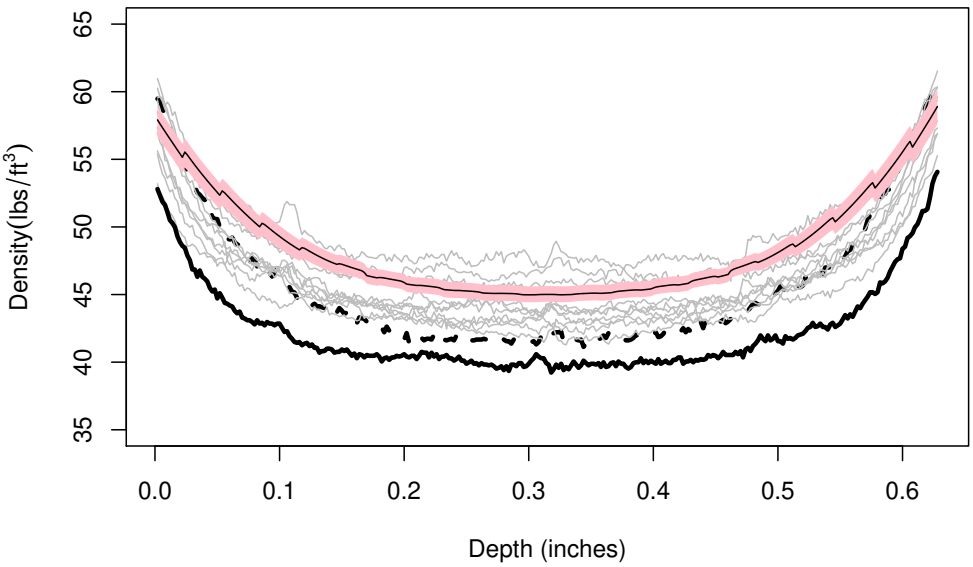

Figure 8: The first 12 profiles in the VDP data. The 6_{th} and 10_{th} profiles are shown by the bold solid and bold dashed curves, respectively. The estimate $\widehat{g}(x)$ is shown by the thin solid curve, and the 95% pointwise confidence intervals for $g(x)$ are shown by the shaded region around $\widehat{g}(x)$.

5 Concluding Remarks

Profile monitoring has many important applications. In practice, observed profile data often have between-profile and within-profile data correlation. In the SPC literature, however, we could not find any existing univariate nonparametric profile monitoring methods that can accommodate the between-profile data correlation, although there are some existing nonparametric methods that can accommodate within-profile data correlation. In the previous sections, we have described our proposed new method for Phase I profile monitoring, which can accommodate both between-profile and within-profile data correlation. Our new method is based on the nonparametric mixed-effects modeling approach that is discussed in Section 2.1, in which between-profile data correlation is described by a random-effect term that depends on the design points of a profile. Numerical examples in Sections 3 and 4 show that the proposed method performs well in different cases considered. As discussed in Section 2, the random-effects model (2) is kept simple, and there

are many possible generalizations to describe more complicated within-profile and between-profile data correlation. Also, the current paper focuses on Phase I monitoring of univariate profile data. It might be possible for this method to be generalized for Phase II monitoring of univariate or multivariate profiles when both between-profile and within-profile data correlation is present. The current method depends on the parameter p in Model (2). Although an empirical guideline is provided in Section 3.1 (cf., Figure 1) about its selection, a theoretical study about its dependence on the curvature of the random-effects process $\gamma(x)$ in Model (1) might be helpful. These issues will be addressed in our future research.

Acknowledgments: We thank a referee for many constructive comments and suggestions about the paper, which improved its quality greatly. Zhou's research is supported in part by the National Natural Science Foundation of China grant 11671178 and the project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions. Qiu's research is supported in part by the NSF grant DMS-1914639.

References

Abdel-Salam, G.A.-S., Brich, 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.

Abdel-Salam, G.A.-S., and Brich, J.B. (2019), "A semiparametric nonlinear mixed model approach to phase I profile monitoring," *Communications in Statistics - Simulation and Computation*, **48**, 1677–1693.

Apley, D.W., and Lee, H.C. (2003), "Design of exponentially weighted moving average control charts for autocorrelated processes with model uncertainty," *Technometrics*, **45**, 187–198.

Apley, D.W., and Lee, H.C. (2008), "Robustness comparison of exponentially weighted moving average charts on autocorrelated data and on residuals," *Journal of Quality Technology*, **40**,

428–447.

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

Chen, K., and Jin, Z. (2005), "Local polynomial regression analysis of clustered data," *Biometrika*, **92**, 59–74.

Chen, Y., Birch, J.B., and Woodall, W.H. (2015), "Cluster-based profile analysis in Phase I," *Journal of Quality Technology*, **47**, 14–29.

Colosimo, B.M., Cicorella, P., Pacella, M., and Blaco, M. (2014), "From profile to surface monitoring: SPC for cylindrical surfaces via Gaussian processes," *Journal of Quality Technology*, **46**, 95–113.

Epanechnikov, V. A. (1969). "Non-parametric estimation of a multivariate probability density," *Theory of Probability and its Applications*, **14**, 153–158.

McGinnity, K., Chicken, E., and Pignatiello, J.J., Jr. (2015), "Nonparametric changepoint estimation for sequential nonlinear profile monitoring," *Quality and Reliability Engineering International*, **31**, 57–73.

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.

Diggle, P.J., Heagerty, P.J., Liang, K.-Y., and Zeger, S.L. (2002), *Analysis of Longitudinal Data (2nd edition)*, Oxford, UK: Oxford University Press.

Hawkins, D.M., and Olwell, D.H. (1998), *Cumulative Sum Charts and Charting for Quality Improvement*, New York: Springer-Verlag.

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.

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

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

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

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

Li, Y. (2011), "Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation," *Biometrika*, **98**, 355–370.

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

Li, W., and Qiu, P. (2020), "A general charting scheme for monitoring serially correlated data with short-memory dependence and nonparametric distributions," *IISE Transactions*, **52**, 61–74.

Kim, K., Mahmoud, M.A., and Woodall, W.H. (2003), "On the monitoring of linear profiles," *Journal of Quality Technology*, **35**, 317–328.

Mahmoud, M.A., and Woodall, W.H. (2004), "Phase I analysis of linear profiles with calibration applications," *Technometrics*, **46**, 380–391.

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

Paynabar, K., Qiu, P., and Zou, C. (2016), "A change point approach for phase-I analysis in multivariate profiles monitoring and diagnosis," *Technometrics*, **58**, 191–204.

Qiu, P. (2005), *Image Processing and Jump Regression Analysis*, New York: John Wiley & Sons.

Qiu, P. (2014), *Introduction to Statistical Process Control,* Boca Raton, FL: Chapman Hall/CRC.

Qiu, P., and Xiang, D. (2010), "Univariate dynamic screening system: an approach for identifying individuals with irregular longitudinal behavior," *Technometrics*, **56**, 248–260.

Qiu, P., and Zou, C. (2010), "Control chart for monitoring nonparametric profiles with arbitrary design," *Statistica Sinica*, **20**, 1655–1682.

Qiu, P., Zou, C., and Wang, Z. (2010), "Nonparametric profile monitoring by mixed effects modeling (with discussions)," *Technometrics*, **52**, 265–293.

Qiu, P., Zi, X., and Zou, C. (2018), "Nonparametric dynamic curve monitoring," *Technometrics*, **60**, 386–397.

Stover, F.S., and Brill, R.V. (1998), "Statistical quality control applied to ion chromatography calibrations," *Journal of Chromatography*, **804**, 37–43.

Walker, E., and Wright, S.P. (2002), "Comparing curves using additive models," *Journal of Quality Technology*, **34**, 118–129.

Wang, K., and Tsung, F. (2005), "Using profile monitoring techniques for a data-rich environment with huge sample sizes," *Quality and Reliability Engineering International*, **21**, 677–688.

Wang, A., Wang, K., and Tsung, F. (2014), "Statistical surface monitoring by spatial-structure modeling," *Journal of Quality Technology*, **46**, 359–376.

Wei, Y., Zhao, Z., and Lin, D.K.J. (2012), "Profile control charts based on nonparametric L1 regression methods," *Annals of Applied Statistics*, **6**, 409–427.

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 Agriculture, 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.

Wu, H., and Zhang, J. (2002), "Local polynomial mixed-effects models for longitudinal data," *Journal of the American Statistical Association*, **97**, 883–897.

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

Yan, H., Paynabar, K., and Shi, J. (2018), "Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition." *Technometrics*, **60**, 181–197.

Yang, K., and Qiu, P. (2020), "Online sequential monitoring of spatio-temporal disease incidence rates," *IISE Transactions*, **52**, 1218–1233.

Yeh, A.B., Huwang, L., and Li, Y.M. (2009), "Profile monitoring for a binary response," *IIE Transactions*, **41**, 931–941.

Zang, Y., and Qiu, P. (2018), "Phase I monitoring of spatial surface data from 3D printing," *Technometrics*, **60**,169–180.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998), "Semiparametric stochastic mixed models for longitudinal data," *Journal of the American Statistical Association*, **93**, 710–719.

Zhang, Y., He, Z., Zhang, C., and Woodall,W. (2014), "Control charts for monitoring linear profiles with within-profile correlation using Gaussian process models," *Quality and Reliability Engineering International*, **30**, 487–501.

Zhou, S., and Jin, J. (2005), "Automatic feature selection for unsupervised clustering of cycle-based signals in manufacturing processes," *IIE Transactions*, **37**, 569–584.

Zou, C., Zhang, Y., and Wang, Z. (2006), "Control chart based on change-point model for monitoring linear profiles," *IIE Transactions*, **38**, 1093–1103.

Zou, C., Tsung, F., and Wang, Z. (2007), "Monitoring general linear profiles using multivariate exponentially weighted moving average schemes," *Technometrics*, **49**, 395–408.

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

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