Surveillance of Cardiovascular Diseases Using A Multivariate Dynamic Screening System

Peihua Qiu[1] and Dongdong Xiang[2]

[1]Department of Biostatistics, University of Florida

[2]School of Finance and Statistics, East China Normal University

Abstract

In the SHARe Framingham Heart Study of the National Heart, Lung and Blood Institute, one major task is to monitor several health variables (e.g., blood pressure and cholesterol level) so that their irregular longitudinal pattern can be detected as soon as possible and some medical treatments applied in a timely manner to avoid some deadly cardiovascular diseases (e.g., stroke). To handle this kind of applications effectively, we propose a new statistical methodology called multivariate dynamic screening system (MDySS) in this paper. The MDySS method combines the major strengths of the multivariate longitudinal data analysis and the multivariate statistical process control, and it makes decisions about the longitudinal pattern of a subject by comparing it with other subjects cross-sectionally and by sequentially monitoring it as well. Numerical studies show that MDySS works well in practice.

Key Words: Dynamic screening; LASSO; Multivariate longitudinal data; Process monitoring; Process screening; Standardization; Unequal sampling intervals.

1 Introduction

Motivational example of this research is the SHARe Framingham Heart Study of the National Heart, Lung and Blood Institute. In the study, each of 972 patients was followed

7 times, and some medical indices, including the systolic blood pressure (mmHg), the diastolic blood pressure (mmHg), the total cholesterol level (mg/100ml), and the glucose level (mg/100ml), were recorded during each clinic visit. Among the 972 patients, 945 did not have any strokes during the study, and the remaining 27 patients had at least one stroke. One major goal of the study is to identify patients with irregular longitudinal patterns of the medical indices as early as possible, so that some medical treatments can be taken in a timely manner. In our daily life, this kind of *multivariate dynamic screening (MDS)* problem is popular. For instance, we often need to monitor the quality/performance of a subject or product (e.g., patients, airplanes, cars, health care systems), to make sure that it functions satisfactorily. To this end, observations of multiple variables related to its quality/performance can be collected sequentially over time for statistical analysis, and a signal should be given, as soon as possible, once the quality/performance becomes unacceptably poor. This paper proposes a novel statistical method to solve the MDS problem.

In the literature, there are two types of methods that are relevant to the MDS problem. The first type belongs to the research area of *longitudinal data analysis (LDA)*. By an LDA method, we can construct confidence intervals for the means of the performance variables at different time points based on an observed dataset of some well-functioning subjects. Then, a new subject can be identified to have an irregular longitudinal pattern if its observations fall outside the confidence intervals. There are some existing methods for constructing such confidence intervals [1-9]. This confidence interval approach makes decisions about a subject's pattern at a given time point by comparing the subject with a group of well-functioning subjects cross-sectionally at that time point. It may be inefficient for handling the MDS problem for the following reasons. First, it does not make use of all history data of a subject in question when making decisions about its performance at the current time point. For instance, when we are interested in monitoring a patient's cholesterol levels over time, if that patient's cholesterol levels are consistently above the mean cholesterol levels of the

healthy people for a long time period, then that patient should still be identified as a person who has an irregular cholesterol level pattern, even if his/her observed cholesterol level at any given time point is within the related confidence interval. Obviously, the confidence interval approach cannot achieve that goal. Second, in the MDS problem, it is critical that statistical decisions are made in a dynamical manner, in the sense that a signal should be given as soon as possible once all available observations of a subject up to the current time point have provided enough evidence to support the decision, so that some interventions can be given in a timely manner. The confidence interval approach does not have this dynamic decision-making property either, because it cannot monitor a subject sequentially over time.

The second type of statistical methods relevant to the MDS problem belongs to the research area of *statistical process control (SPC)*. By a SPC control chart, we monitor each subject sequentially, and a signal is given as soon as the chart detects a shift in the subject's longitudinal pattern from an in-control (IC) status to an out-of-control (OC) status (cf., [10-12]). However, a conventional SPC chart cannot be applied to the MDS problem directly for the following reasons. First, a conventional SPC chart is for monitoring a single process, and it makes decisions about the process by comparing its observed data at the current time point with all of its history data. In the MDS problem, if each subject is regarded as a process, then there are many processes involved. To judge whether a specific subject follows a regular longitudinal pattern, we need to compare him/her with a group of well-functioning subjects over time. Second, in a typical SPC problem, the distribution of the process observations is assumed unchanged when the process is IC. In the MDS problem, however, this distribution often changes over time, even for well-functioning subjects (e.g., the mean total cholesterol level of healthy people would change as they age). Therefore, *there is no existing statistical methods that can handle the MDS problem effectively.*

In this paper, we propose a *multivariate dynamic screening system (MDySS)* for handling the MDS problem. Our proposed MDySS method consists of three main steps. First, the reg-

ular longitudinal pattern of the performance variables is estimated from an observed dataset of certain well-functioning subjects, using a nonparametric multivariate LDA method. Second, for a new subject under study, its longitudinal observations are standardized, using the estimated regular longitudinal pattern obtained in the first step. Third, a multivariate SPC control chart is applied to the standardized observations of the new subject for sequential monitoring of its longitudinal pattern, and a signal is given as soon as its observed data suggest a significant shift in its longitudinal pattern from the estimated regular longitudinal pattern. In the next several sections, we will demonstrate that the MDySS method provides an efficient solution to the MDS problem.

The rest part of the paper is organized as follows. In Section 2, we describe our proposed MDySS method in detail. Some of its statistical properties are discussed in Section 3. A simulation study is presented in Section 4 regarding its numerical performance. The MDySS method is applied to the SHARe Framingham Heart Study in Section 5. Several remarks conclude the paper in Section 6. Some technical details and extra numerical results are included in a supplementary file.

2 Proposed MDySS Method

In this section, we describe our proposed MDySS method in two parts. Estimation of the regular longitudinal pattern of the multiple performance variables from an observed dataset of certain well-functioning subjects is discussed in Subsection 2.1. Then, sequential monitoring of individual subjects after their observations are standardized by the estimated regular longitudinal pattern is discussed in Subsection 2.2.

2.1 Estimation of the regular multivariate longitudinal pattern

Let $\boldsymbol{y}$ be the vector of q performance variables that we are interested in monitoring. Assume that there is an observed dataset, called an IC dataset hereafter, of a group of m well-functioning subjects whose observations of $\boldsymbol{y}$ follow the model

$$\boldsymbol{y}(t_{ij}) = \boldsymbol{\mu}(t_{ij}) + \boldsymbol{\varepsilon}(t_{ij}), \qquad \text{for } j = 1, 2, \ldots, J_i,\ i = 1, 2, \ldots, m, \tag{1}$$

where t_{ij} is the jth observation time of the ith subject, $\boldsymbol{y}(t_{ij}) = (y_1(t_{ij}), \ldots, y_q(t_{ij}))'$ is the observed vector of $\boldsymbol{y}$ at t_{ij}, $\boldsymbol{\mu}(t_{ij}) = (\mu_1(t_{ij}), \ldots, \mu_q(t_{ij}))'$ is the its mean vector, and $\boldsymbol{\varepsilon}(t_{ij}) = (\varepsilon_1(t_{ij}), \ldots, \varepsilon_q(t_{ij}))'$ is the q-dimensional error term. In model (1), we assume that observations of different subjects are independent of each other. For simplicity, we further assume that all observation times are within $[0, 1]$.

It should be pointed out that the definition of well-functioning subjects could be different in different research projects. For instance, in the stroke data example discussed in Section 5, all non-stroke patients are regarded as well-functioning subjects. This is reasonable in that example because we can estimate the longitudinal pattern of non-stroke patients from their longitudinal observations. However, the non-stroke patients may not be regarded as well-functioning subjects in a project concerning another disease (e.g., a digestive disease) since they may suffer such a disease.

The regular longitudinal pattern of $\boldsymbol{y}$ is assumed to be described jointly by its mean function $\boldsymbol{\mu}(t)$ and its covariance matrix function $\Sigma(s, t) = \text{Cov}(\boldsymbol{y}(s), \boldsymbol{y}(t))$, for any $s, t \in [0, 1]$. In univariate cases (i.e., $q = 1$), nonparametric estimation of the mean and variance functions of $\boldsymbol{y}$ has been discussed by some papers [1, 2, 4, 7]. In multivariate cases (i.e., $q > 1$), Xiang et al. [8] recently proposed a nonparametric method for estimating model (1), which is briefly described below. For $j = 1, \ldots, J_i$ and $i = 1, \ldots, m$, let $K_i = \text{diag}\{K_{h_l}(t_{ij} - t), j = 1, \ldots, J_i, l = 1, \ldots, q\}$ be a diagonal matrix, and $W_i = \left(K_i^{-\frac{1}{2}} V_i K_i^{-\frac{1}{2}}\right)^{-1}$, where $K_{h_l}(u) = K(u/h_l)/h_l$, $K(\cdot)$ is a kernel function, $\{h_l, l = 1, \ldots, q\}$ are bandwidths, $V_i = \text{Cov}(\boldsymbol{Y}_i)$ with

$\boldsymbol{Y}_i = (y_1(t_{i1}), \ldots, y_1(t_{iJ_i}), \ldots, y_q(t_{i1}), \ldots, y_q(t_{iJ_i}))'$, and the inverse of a matrix in this paper is referred to the Moore-Penrose generalized inverse that always exists. For any given $t \in [0,1]$, $\boldsymbol{\mu}(t)$ is estimated by the following pth order local polynomial kernel smoothing procedure:

$$\min_{\boldsymbol{\beta} \in R^{q(p+1)}} \sum_{i=1}^{m} [\boldsymbol{Y}_i - (I_{q\times q} \otimes X_i)\boldsymbol{\beta}]' W_i [\boldsymbol{Y}_i - (I_{q\times q} \otimes X_i)\boldsymbol{\beta}] \tag{2}$$

where $\otimes$ denotes the Kronecker product, $I_{q\times q}$ is the $q \times q$ identity matrix,

$$\boldsymbol{\beta} = ((\beta_0^{(1)}, \ldots, \beta_p^{(1)}), \ldots, (\beta_0^{(q)}, \ldots, \beta_p^{(q)}))',$$

and

$$X_i = \begin{pmatrix} 1 & (t_{i1} - t) & \ldots & (t_{i1} - t)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (t_{iJ_i} - t) & \ldots & (t_{iJ_i} - t)^p \end{pmatrix}_{J_i \times (p+1)}.$$

The solution of (2) is

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} (I_{q\times q} \otimes X_i)' W_i (I_{q\times q} \otimes X_i)\right]^{-1} \left[\sum_{i=1}^{n} (I_{q\times q} \otimes X_i)' W_i \boldsymbol{Y}_i\right]. \tag{3}$$

Then, the pth order local polynomial kernel estimator of $\boldsymbol{\mu}(t)$ is

$$\widehat{\boldsymbol{\mu}}(t) = \widehat{\boldsymbol{\beta}}' (I_{q\times q} \otimes \boldsymbol{e}_1) \tag{4}$$

where $\boldsymbol{e}_1$ is the $(p+1)$-dimensional vector that has the value of 1 at the 1st position and 0 at all other positions. In the above estimation procedure, K is chosen to be the Epanechnikov kernel function, and the bandwidths $\{h_l, l = 1, \ldots, q\}$ are determined by the conventional cross-validation (CV) procedure [cf., 13, Chapter 2].

In practice, the quantity V_i is usually unknown and needs to be estimated. To this end, Xiang et al. [8] suggested the following estimation method. First, the local linear kernel smoothing procedure is used to provide an initial estimator of $\boldsymbol{\mu}(t)$, denoted as $\widetilde{\boldsymbol{\mu}}(t) = (\widetilde{\mu}_1(t), \ldots, \widetilde{\mu}_q(t))$. Then, we define residuals

$$\widetilde{\varepsilon}_{ijl} = y_{ijl} - \widetilde{\mu}_l(t_{ij}), \qquad j = 1, 2, \ldots, J_i, l = 1, 2, \ldots, q, i = 1, 2, \ldots, m.$$

Finally, the (l_1, l_2)-th element of $\Sigma(s,t) = (\sigma_{l_1,l_2}(s,t))$, for $l_1, l_2 = 1, 2, \ldots, q$, can be estimated by the following kernel estimator

$$\widetilde{\sigma}_{l_1,l_2}(s,t) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{J_i}\sum_{k=1}^{J_i}\widetilde{\varepsilon}_{ijl_1}\widetilde{\varepsilon}_{ikl_2}K\left(\frac{t_{ij}-s}{g_{l_1}}\right)K\left(\frac{t_{ik}-t}{g_{l_2}}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{J_i}\sum_{k=1}^{J_i}K\left(\frac{t_{ij}-s}{g_{l_1}}\right)K\left(\frac{t_{ik}-t}{g_{l_2}}\right)}, \tag{5}$$

where $\{g_l, l = 1, \ldots, q\}$ are bandwidths. In (5), we can still use the Epanechnikov kernel function, and the bandwidths g_l, for $l = 1, \ldots, q$, can be chosen by the CV procedure when estimating the variance of $y_l(t)$ from quantities $\{\widetilde{\varepsilon}^2_{ijl}, j = 1, 2, \ldots, J_i, i = 1, 2, \ldots, m\}$ using the local linear kernel smoothing procedure. From $\widetilde{\Sigma}(s,t) = (\widetilde{\sigma}_{l_1,l_2}(s,t))$, the corresponding estimator of V_i can be computed. The resulting estimator of $\boldsymbol{\mu}(t)$ computed from (3) and (4) is denoted as $\widehat{\boldsymbol{\mu}}(t, \widetilde{\Sigma}) = (\widehat{\mu}_1(t_{ij}, \widetilde{\Sigma}), \ldots, \widehat{\mu}_q(t_{ij}, \widetilde{\Sigma}))'$. Then, the estimator of $\Sigma(s,t)$ is updated by (5), after $\{\widetilde{\varepsilon}_{ijl}, j = 1, 2, \ldots, J_i, l = 1, 2, \ldots, q, i = 1, 2, \ldots, m\}$ are replaced by

$$\widehat{\varepsilon}_{ijl} = y_{ijl} - \widehat{\mu}_l(t_{ij}, \widetilde{\Sigma}), \qquad j = 1, 2, \ldots, J_i, l = 1, 2, \ldots, q, i = 1, 2, \ldots, m.$$

The resulting estimator of $\Sigma(s,t)$ is denoted as $\widehat{\Sigma}(s,t)$.

Model (1) and its estimation procedure described above are quite general. It does not impose any specific structure on the random error term $\boldsymbol{\varepsilon}(t)$, and allows the error covariance matrix $\Sigma(s,t)$ to vary over s and t. As a comparison, the alternative mixed-effects modeling approach [14] usually assumes that $\boldsymbol{\varepsilon}(t)$ consists of two independent parts: one is the random-effects and the other is the pure measurement error, and the variance/covariance of the pure measurement error is time-independent. So, model (1) is more general than most mixed-effects models in the literature. In certain applications, it might be reasonable to specify the correlation structure among observations of $\boldsymbol{y}(t)$ by a parametric model (e.g., a parametric time series model). In such cases, the specified correlation structure can be accommodated when estimating the covariance matrix V_i. For instance, in the SPC literature, it is often assumed that observations of $\boldsymbol{y}(t)$ within a subject are independent of each other at different time points. If that assumption is valid, then the covariance matrix V_i is uniquely determined

by the variance function $\Sigma(t,t)$, for $t \in [0,1]$. Its (l_1,l_2)-th element $\sigma_{l_1,l_2}(t,t)$, for $l_1, l_2 = 1,2,\ldots,q$, can be estimated by

$$\widetilde{\sigma}_{l_1,l_2}(t,t) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{J_i}\widetilde{\varepsilon}_{ijl_1}\widetilde{\varepsilon}_{ijl_2}K\left(\frac{t_{ij}-t}{g_{l_1}}\right)K\left(\frac{t_{ij}-t}{g_{l_2}}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{J_i}K\left(\frac{t_{ij}-t}{g_{l_1}}\right)K\left(\frac{t_{ij}-t}{g_{l_2}}\right)}. \tag{6}$$

2.2 Dynamic screening of irregular longitudinal patterns

The estimated mean function $\widehat{\boldsymbol{\mu}}(t;\widetilde{\Sigma})$ and the estimated covariance matrix function $\widehat{\Sigma}(s,t)$ described in the previous subsection can be used for describing the estimated regular longitudinal pattern of the q-dimensional performance vector $\boldsymbol{y}$. In this subsection, we describe our proposed MDySS method for sequential monitoring of each individual subject in various different cases by using the estimated regular longitudinal pattern, such that a signal is given as soon as possible after the subject's longitudinal pattern becomes irregular.

Assume that observations of $\boldsymbol{y}$ of a new subject under study are obtained at times $t_1^*, t_2^*, \ldots$ in the design interval $[0,1]$. Its longitudinal pattern is called in-control (IC) if its observations follow the model (1). Otherwise, its longitudinal pattern is called out-of-control (OC). So, when the new subject's longitudinal pattern is IC, its observations follow the model

$$\boldsymbol{y}(t_j^*) = \boldsymbol{\mu}(t_j^*) + \Sigma^{\frac{1}{2}}(t_j^*,t_j^*)\boldsymbol{\epsilon}(t_j^*), \qquad \text{for } j = 1,2,\ldots, \tag{7}$$

where $\Sigma^{\frac{1}{2}}(t,t)\boldsymbol{\epsilon}(t)$ equals $\boldsymbol{\varepsilon}(t)$ in model (1). Thus, $\boldsymbol{\epsilon}(t)$ in model (7) has mean $\mathbf{0}$ and covariance matrix $I_{q\times q}$ at each t. For the $\boldsymbol{y}$ observations of the new subject, let us define their *standardized values* by

$$\widehat{\boldsymbol{\epsilon}}(t_j^*) = \widehat{\Sigma}^{-\frac{1}{2}}(t_j^*,t_j^*)\left(\boldsymbol{y}(t_j^*) - \widehat{\boldsymbol{\mu}}(t_j^*;\widetilde{\Sigma})\right), \qquad \text{for } j = 1,2,\ldots. \tag{8}$$

By using these standardized observations of the new subject, we have actually compared its longitudinal pattern cross-sectionally with the estimated regular longitudinal pattern at the time points $t_1^*, t_2^*, \ldots$. In cases when the longitudinal pattern of the new subject is IC, the

mean and variance of the standardized observation $\widehat{\boldsymbol{\epsilon}}(t_j^*)$ are roughly $\mathbf{0}$ and $I_{q\times q}$, respectively, for each j. For the moment, let us assume that we are only interested in monitoring the mean longitudinal pattern of the new subject. In such cases, any shift in the mean of an original observation would result in a shift in the mean of its standardized observation, and vice versa. Therefore, to monitor the mean longitudinal pattern of the new subject, we can simply monitor the means of the standardized observations $\{\boldsymbol{\epsilon}(t_j^*), j = 1, 2, \ldots\}$.

To detect mean shifts in the standardized observations $\{\boldsymbol{\epsilon}(t_j^*), j = 1, 2, \ldots\}$ of the new subject, there are some existing control charts in the multivariate SPC literature [14-23]. All these methods assume that observation vectors are independent and normally distributed, and observation times are equally spaced. In order to use these methods, let us first discuss cases when the original observations $\{\boldsymbol{y}(t_j^*), j = 1, 2, \ldots\}$ are independent and normally distributed, and all observation times are equally spaced. In such cases, the standardized observations $\{\widehat{\boldsymbol{\epsilon}}(t_j^*), j = 1, 2, \ldots\}$ are asymptotically i.i.d. with the normal distribution $N(\mathbf{0}, I_{q\times q})$ when the longitudinal pattern of the new subject is IC. To use the method by Lowry et al [20], let us consider the multivariate exponentially weighted moving average (MEWMA) statistic

$$\boldsymbol{E}_{M,j} = \lambda_M \widehat{\boldsymbol{\epsilon}}(t_j^*) + (1 - \lambda_M)\boldsymbol{E}_{M,j-1}, \qquad \text{for } j \geq 1,$$

where $\boldsymbol{E}_{M,0} = \mathbf{0}$, and $\lambda_M \in (0, 1]$ is a weighting parameter. The chart gives a signal when

$$\boldsymbol{E}'_{M,j}\Sigma^{-1}_{0,\boldsymbol{E}_{M,j}}\boldsymbol{E}_{M,j} > h_M, \tag{9}$$

where $h_M > 0$ is a control limit, and $\Sigma_{0,\boldsymbol{E}_{M,j}}$ is the covariance matrix of $\boldsymbol{E}_{M,j}$ when the longitudinal pattern of the new subject is IC. It can be checked that

$$\Sigma_{0,\boldsymbol{E}_{M,j}} = \frac{\lambda_M}{2 - \lambda_M}[1 - (1 - \lambda_M)^{2j}]\Sigma_{0,\widehat{\boldsymbol{\epsilon}}(t_j^*)},$$

where $\Sigma_{0,\widehat{\boldsymbol{\epsilon}}(t_j^*)} \approx I_{q\times q}$ is the IC covariance matrix of $\widehat{\boldsymbol{\epsilon}}(t_j^*)$. When j is big, $[1-(1-\lambda_M)^{2j}] \approx 1$. So, in practice, (9) can be replaced by

$$\frac{2 - \lambda_M}{\lambda_M}\boldsymbol{E}'_{M,j}\boldsymbol{E}_{M,j} > h_M. \tag{10}$$

By using the MEWMA chart (10), we sequentially monitor the new subject and make use of all its history data up to the current time point j, after its longitudinal pattern is compared cross-sectionally with the estimated regular longitudinal pattern in (8). The resulting MDySS method is denoted as MDySS-M, where the last letter "M" denotes MEWMA.

To monitor the standardized observations $\{\widehat{\boldsymbol{\epsilon}}(t_j^*), j = 1, 2, \ldots\}$, an alternative approach is to jointly use q univariate EWMA charts for monitoring q individual components of the observed data. More specifically, let $\widehat{\epsilon}_l(t_j^*)$ be the lth component of $\widehat{\boldsymbol{\epsilon}}(t_j^*)$, and

$$E_{C,j,l} = \lambda_{C,l}\widehat{\epsilon}_l(t_j^*) + (1 - \lambda_{C,l})E_{C,j-1,l}, \qquad \text{for } l = 1, 2, \ldots, q,\ j \geq 1,$$

where $E_{C,0,l} = 0$, and $\lambda_{C,l} \in (0, 1]$ are weighting parameters. Then, the joint monitoring scheme gives a signal at j when there is at least one $1 \leq l \leq q$ such that

$$\sqrt{\frac{2 - \lambda_{C,l}}{\lambda_{C,l}}} E_{C,j,l} > h_{C,l}, \tag{11}$$

where $h_{C,l} > 0$ is a control limit. In (11), because all components of $\widehat{\boldsymbol{\epsilon}}(t_j^*)$ have asymptotic mean 0 and asymptotic variance 1, it is reasonable to choose all $\lambda_{C,l}$ to be the same as λ_C, and all $h_{C,l}$ to be the same as h_C. The resulting MDySS method is denoted as MDySS-C, where the last letter "C" denotes the combination of multiple univariate EWMA charts.

Recently, Zou and Qiu [24] suggested integrating the variable selection method LASSO [25, 26] into a MEWMA chart for solving the conventional SPC problem. They demonstrated that the resulting LASSO-based MEWMA chart can effectively detect shifts of various sizes and directions. Next, we adapt this method to solve the current MDySS problem. Let us first define the MEWMA statistic

$$\boldsymbol{U}_j = \lambda_L \widehat{\boldsymbol{\epsilon}}(t_j^*) + (1 - \lambda_L)\boldsymbol{U}_{j-1}, \quad \text{for } j = 1, 2, \ldots, \tag{12}$$

where $\boldsymbol{U}_0 = \boldsymbol{0}$, and $\lambda_L \in (0, 1]$ is a weighting parameter. Then, for each $\boldsymbol{U}_j$, we compute q LASSO estimators of its mean vector, denoted as $\{\widehat{\boldsymbol{\alpha}}_{j,\widetilde{\gamma}_k}, k = 1, 2, \ldots, q\}$, from the following

minimization procedure

$$\min_{\boldsymbol{\alpha}\in R^q} (\boldsymbol{U}_j - \boldsymbol{\alpha})' (\boldsymbol{U}_j - \boldsymbol{\alpha}) + \widetilde{\gamma}_k \sum_{l=1}^{q} \frac{\alpha_l}{|U_{jl}|},$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_q)'$, $\boldsymbol{U}_j = (U_{j1}, U_{j2}, \ldots, U_{jq})'$, and $\{\widetilde{\gamma}_k, k = 1, 2, \ldots, q\}$ are q values determined from a set of "transition points" by the LARS algorithm of Efron et al. [27]. Then, the LASSO-based MEWMA chart gives a signal at j if

$$Q_j = \max_{k=1,\ldots,q} \frac{W_{j,\widetilde{\gamma}_k} - \mathrm{E}(W_{j,\widetilde{\gamma}_k})}{\sqrt{\mathrm{Var}(W_{j,\widetilde{\gamma}_k})}} > h_L, \tag{13}$$

where $W_{j,\widetilde{\gamma}_k} = \left(\boldsymbol{U}_j' \widehat{\boldsymbol{\alpha}}_{j,\widetilde{\gamma}_k}\right)^2 / (\widehat{\boldsymbol{\alpha}}_{j,\widetilde{\gamma}_k}' \widehat{\boldsymbol{\alpha}}_{j,\widetilde{\gamma}_k})$, and h_L is a control limit. In (13), the quantities $\mathrm{E}(W_{j,\widetilde{\gamma}_k})$ and $\mathrm{Var}(W_{j,\widetilde{\gamma}_k})$ can be approximated by simulation, after using their property that they do not depend on λ_L and j when the process is IC [cf., 24, Proposition 3]. For instance, if λ_L is chosen 1, then $\boldsymbol{U}_j = \widehat{\boldsymbol{\epsilon}}(t_j^*)$. In such cases, if we randomly generate M vectors of $\widehat{\boldsymbol{\epsilon}}(t_j^*)$ from the $N(\mathbf{0}, I_{q\times q})$ distribution, and compute the M corresponding LASSO estimators $\widehat{\boldsymbol{\alpha}}_{j,\widetilde{\gamma}_k}$ and the M corresponding values of $W_{j,\widetilde{\gamma}_k}$, for each k. Then, $\mathrm{E}(W_{j,\widetilde{\gamma}_k})$ and $\mathrm{Var}(W_{j,\widetilde{\gamma}_k})$ can be approximated by the sample mean and the sample variance of the M values of $W_{j,\widetilde{\gamma}_k}$. The MDySS method based on the chart (13) is denoted as MDySS-L, where the last letter "L" denotes LASSO.

In the SPC literature, to evaluate the performance of a control chart, we usually use the IC average run length (ARL), denoted as ARL_0, and the OC ARL, denoted as ARL_1. ARL_0 is defined to be the average number of time points from the beginning of process monitoring to the signal time when the process is IC, and ARL_1 is defined to be the average number of time points from the occurrence of a shift to the signal time after the process becomes OC. In a MEWMA chart (e.g., the chart (12)-(13)), usually the weighting parameter (e.g., λ_L) is specified beforehand. It has been well demonstrated in the literature that large values of the weighting parameter are good for detecting large shifts, and small values are good for detecting small shifts. Commonly used values of the weighting parameter include 0.05, 0.1 and 0.2. The control limit (e.g., h_L in (13)) is then chosen to reach a pre-specified ARL_0

value. The chart performs better for detecting a given shift if its ARL_1 value is smaller. In practice, however, the observation times $\{t_j^*, j = 1, 2, \ldots\}$ may not be equally spaced, as in the SHARe Framingham Heart Study discussed in Section 5. In such cases, ARL_0 and ARL_1 are obviously inappropriate any more for measuring the performance of a control chart, and we need to define new performance measures. To this end, let $\omega > 0$ be a *basic time unit* in a given application, which is the largest time unit that all unequally spaced observation times are its integer multiples (e.g., the basic time unit for clinic visit times could be one day). Define

$$n_j^* = t_j^*/\omega, \qquad \text{for } j = 0, 1, 2, \ldots \tag{14}$$

where $n_0^* = t_0^* = 0$. Then, $t_j^* = n_j^*\omega$, for all j, and n_j^* is the jth observation time in the basic time unit. In cases when the new subject is IC and a control chart (e.g., the chart (12)-(13)) gives a signal at the sth observation time, then n_s^* is a random variable measuring the time to a false signal. Its mean $E(n_s^*)$ measures the IC average time to the signal (ATS), denoted as ATS_0. If the longitudinal pattern of the new subject starts to shift from the regular longitudinal pattern at the τth observation time and the control chart gives a signal at the sth time point with $s \geq \tau$, then the mean of $n_s^* - n_\tau^*$ is called the OC ATS, denoted as ATS_1. Then, to measure the performance of the control chart, its ATS_0 value can be fixed at a certain level beforehand, and the chart performs better if its ATS_1 value is smaller when detecting a shift of a given size. It is obvious that the values of ATS_0 and ATS_1 are just constant multiples of the corresponding values of ARL_0 and ARL_1 in cases when the observation times are equally spaced. In such cases, the two sets of measures are equivalent. In the SPC literature, the concept of ATS has been proposed for measuring the performance of a control chart with variable sampling intervals (VSI) [28, 29, 30]. However, the VSI problem in SPC is completely different from the MDySS problem discussed here. In the VSI problem, the next observation time is determined by the current and all past observations of the process; the next observation is collected sooner if there is more evidence of a shift

based on all available process observations, and later otherwise. Therefore, the VSI scheme is an integrated part of process monitoring, and it is designed by quality engineers. As a comparison, in the proposed MDySS method, observation times are often pre-specified.

Next, we discuss calculation of the control limit h_L of the chart (12)-(13). Calculation of the control limits of the charts (10) and (11) can be discussed similarly. In cases when the IC mean function $\boldsymbol{\mu}(t)$ and the IC covariance matrix function $\Sigma(t,t)$ are both known, $\widehat{\boldsymbol{\epsilon}}(t_j^*)$ in (12) can be replaced by $\boldsymbol{\epsilon}(t_j^*)$ which has the IC distribution $N(\mathbf{0}, I_{q\times q})$. In such cases, for a given ATS_0 value, h_L can be determined by a numerical searching algorithm [31], in which random vectors generated from $N(\mathbf{0}, I_{q\times q})$ can be used in place of $\boldsymbol{\epsilon}(t_j^*)$. For instance, when the sampling rate is fixed at d, which is defined to be the number of observation times every 10 basic time units in this paper, the computed h_L values based on 50,000 simulations are presented in Table 1, in cases when $q = 5$, $\omega = 0.001$, $ATS_0 = 25, 50, 75, 100, 125, 150$, $\lambda_L = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$, and $d = 2, 5, 10$. From the table, it can be seen that h_L increases with ATS_0, d and λ_L. In cases when $\boldsymbol{\mu}(t)$ and $\Sigma(t,t)$ are both unknown and they need to be estimated from an observed dataset of a group of m well-functioning subjects (cf., Section 2.1), h_L can still be computed in the way just described, as long as m is not too small, which will be justified in Section 4.

In the above discussion, we assume that the original observations $\{\boldsymbol{y}(t_j^*), j = 1, 2, \ldots\}$ of a new subject under study are independent and normally distributed. In practice, both the normality and the independence assumptions are usually violated. In cases when these assumptions are violated, control limit values computed based on these assumptions are usually inappropriate to use, because the actual ATS_0 values could be substantially different from the assumed ATS_0 values [31, 32]. Next, we propose a numerical approach to compute the control limit h_L of the chart (12)-(13) in such cases from an IC dataset. Computation of the control limits of the charts (10) and (11) can be discussed similarly. Assume that there is an observed dataset of a group of m well-functioning subjects whose observations of

Table 1: Computed h_L values used in the LASSO-based MEWMA chart (13) in cases when $q = 5$, $\boldsymbol{\epsilon}(t_j^*) \sim N(\mathbf{0}, I_{q\times q})$, the IC mean function $\boldsymbol{\mu}(t)$ and the IC covariance matrix function $\Sigma(t,t)$ are both known, $\omega = 0.001$, $ATS_0 = 25, 50, 75, 100, 125, 150$, $\lambda = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$, and $d = 2, 5, 10$.

		ATS_0					
d	λ	25	50	75	100	125	150
	0.05	-0.815	-0.321	0.026	0.330	0.577	0.798
	0.1	-0.353	0.276	0.717	1.045	1.314	1.533
2	0.2	0.225	0.935	1.375	1.692	1.951	2.154
	0.3	0.575	1.269	1.700	2.000	2.225	2.410
	0.4	0.806	1.495	1.875	2.156	2.393	2.599
	0.5	0.962	1.618	2.006	2.255	2.481	2.635
	0.05	-0.143	0.594	1.075	1.425	1.713	1.944
	0.1	0.500	1.306	1.801	2.250	2.381	2.587
5	0.2	1.134	1.944	2.388	2.712	2.945	3.140
	0.3	1.500	2.231	2.656	2.924	3.163	3.368
	0.4	1.712	2.381	2.812	3.067	3.301	3.457
	0.5	1.869	2.478	2.853	3.153	3.365	3.534
	0.05	0.575	1.425	1.916	2.281	2.559	2.796
	0.1	1.366	2.131	2.622	2.955	3.173	3.394
10	0.2	1.993	2.681	3.120	3.431	3.639	3.837
	0.3	2.213	2.991	3.344	3.635	3.858	4.051
	0.4	2.387	3.056	3.463	3.739	3.944	4.138
	0.5	2.500	3.151	3.519	3.805	4.022	4.185

$\boldsymbol{y}$ follow the model (1). The data of the first m_1 subjects are used for obtaining estimators $\widehat{\boldsymbol{\mu}}(t;\widetilde{\Sigma})$ and $\widehat{\Sigma}(s,t)$, as discussed in Subsection 2.1. Then, the control limit h_L is computed from the remaining $m_2 = m - m_1$ subjects using a block bootstrap procedure [33] described below. (i) Compute the standardized observations of the m_2 well-functioning subjects by

$$\widehat{\boldsymbol{\epsilon}}(t_{ij}) = \widehat{\Sigma}^{-\frac{1}{2}}(t_{ij},t_{ij})\left(\boldsymbol{y}(t_{ij}) - \widehat{\boldsymbol{\mu}}(t_{ij};\widetilde{\Sigma})\right), \qquad \text{for } j = 1,2,\ldots,J_i,\ i = m_1+1, m_1+2,\ldots,m.$$

(ii) It has been justified numerically that both $\mathrm{E}(W_{j,\widetilde{\gamma}_k})$ and $\mathrm{Var}(W_{j,\widetilde{\gamma}_k})$ do not depend on λ_L and j, as in cases when the original observations of the subjects are independent and normally distributed. Then, for the ith subject and for each $k = 1,2,\ldots,q$, we can compute $\{W_{j,\widetilde{\gamma}_k}, j = 1,2,\ldots,J_i\}$. For each k, the quantities $\mathrm{E}(W_{j,\widetilde{\gamma}_k})$ and $\mathrm{Var}(W_{j,\widetilde{\gamma}_k})$ can be estimated by the sample mean and sample variance of $\{W_{j,\widetilde{\gamma}_k}, j = 1,2,\ldots,J_i, i = m_1+1, m_1+2,\ldots,m\}$.

(iii) Randomly select B subjects with replacement from the m_2 well-functioning subjects, and use their standardized observations to compute the value of h_L by a numerical searching algorithm such that a given ATS_0 level is reached. In all numerical examples presented in Sections 4 and 5, we choose $m_1 = m/5$.

3 Statistical Properties

In cases when the covariance matrix function $\Sigma(s,t)$ is assumed known, Xiang et al. [8] proved that the estimated mean function $\widehat{\boldsymbol{\mu}}(t)$ is pointwise L^2 consistent at each $t \in [0,1]$. In this paper, both the estimated mean function $\widehat{\boldsymbol{\mu}}(t;\widetilde{\Sigma})$ and the estimated covariance matrix function $\widehat{\Sigma}(s,t)$ are used for describing the estimated regular longitudinal pattern of $\boldsymbol{y}(t)$ in cases when $\Sigma(s,t)$ is unknown. To justify the legitimacy of this description, we give some new statistical properties of $\widehat{\boldsymbol{\mu}}(t;\widetilde{\Sigma})$ and $\widehat{\Sigma}(s,t)$ in this section. To present our major theoretical results, the following regularity conditions are needed.

(C1) The probability density f of the design points $\{t_{ij}\}$ has two continuous derivatives and is bounded away from 0 in the design space $[0,1]$.

(C2) For any $1 \le l_1, l_2 \le q$, there exists a constant $\delta \in [0,1)$ such that $\sup_t E|y_{l_1}(t)y_{l_2}(t)|^{2+\delta} < \infty$.

(C3) All q components of $\boldsymbol{\mu}(t)$ have $(p+1)$-th continuous derivatives in $[0,1]$.

(C4) For any $k_1, k_2, k_3, k_4 \in \{0,1\}$, $E\{y_{l_1}^{k_1}(t)y_{l_2}^{k_2}(t)y_{l_3}^{k_3}(t)y_{l_4}^{k_4}(t)\}$ has two continuous derivatives in $[0,1]$.

(C5) For $l = 1,2,\ldots,q$, $h_l/h_{\max} \to c_{1l}$ and $J_S h_l^5 \to c_{2l}$, where $h_{\max} = \max\{h_1,\ldots,h_q\}$, $J_S = \sum_{i=1}^m J_i$, c_{1l} and c_{2l} are two positive constants, and "$\to$" denotes convergence when J_S increases.

(C6) For $l = 1,2,\ldots,q$, $g_l/g_{\max} \to c_{3l}$ and $J_S g_l^5 \to c_{4l}$, where $g_{\max} = \max\{g_1,\ldots,g_q\}$, and c_{3l} and c_{4l} are two positive constants.

First, in cases when the covariance matrix function $\Sigma(s,t)$ is assumed known, the estimated mean function $\widehat{\boldsymbol{\mu}}(t)$ defined in (4) has the following uniform consistency result.

Theorem 3.1 Under (C1)-(C5), for any $l = 1, 2, \ldots, q$, we have

$$\sup_{t\in[0,1]} |\widehat{\mu}_l(t) - \mu_l(t)| = O_P\left(h_{\max}^{p+1} + 1/(J_S h_{\max})^{\frac{1}{2}}\right). \tag{15}$$

In cases when $\Sigma(s,t)$ is unknown, it is estimated by $\widetilde{\Sigma}(s,t)$ defined in (5). Regarding $\widetilde{\Sigma}(s,t)$, we have the following result.

Theorem 3.2 Under (C1)-(C6), for any $l_1, l_2 = 1, \ldots, q$, we have

$$\sup_{s,t\in[0,1],s\neq t} |\widetilde{\sigma}_{l_1,l_2}(s,t) - \sigma_{l_1,l_2}(s,t)| = O_P\left(g_{\max}^2 + 1/(J_{SS}^{\frac{1}{2}} g_{\max})\right), \tag{16}$$

$$\sup_{t\in[0,1]} |\widetilde{\sigma}_{l_1 l_2}(t,t) - \sigma_{l_1 l_2}(t,t)| = O_P(g_{\max}^2 + 1/(J_S g_{\max})^{\frac{1}{2}}), \tag{17}$$

where $J_{SS} = \sum_{i=1}^m J_i(J_i - 1)$.

In cases when $\Sigma(s,t)$ is unknown, our final estimator of the mean function $\boldsymbol{\mu}(t)$ is $\widehat{\boldsymbol{\mu}}(t;\widetilde{\Sigma})$ which is proved to be uniformly consistent in the design interval $[0,1]$ in the following theorem.

Theorem 3.3 Under (C1)-(C6), for any $l = 1, 2, \ldots, q$, we have

$$\sup_{t\in[0,1]} \left|\widehat{\mu}_l(t,\widetilde{\Sigma}) - \mu_l(t)\right| = O_P\left(h_{\max}^{p+1} + 1/(J_S h_{\max})^{\frac{1}{2}}\right). \tag{18}$$

In cases when $\Sigma(s,t)$ is unknown, our final estimator of $\Sigma(s,t)$ is $\widehat{\Sigma}(s,t) = (\widehat{\sigma}_{l_1,l_2}(s,t))$, where $\widehat{\sigma}_{l_1,l_2}(s,t)$ is its (l_1,l_2)-th element, for $l_1, l_2 = 1, \ldots, q$. For $\widehat{\Sigma}(s,t)$, we have the following result.

Theorem 3.4 Under (C1)-(C6), for any $l_1, l_2 = 1, \ldots, q$, we have

$$\sup_{s,t\in[0,1],s\neq t} |\widehat{\sigma}_{l_1,l_2}(s,t) - \sigma_{l_1,l_2}(s,t)| = O_P\left(g_{\max}^2 + 1/(J_{SS} g_{\max})\right), \tag{19}$$

$$\sup_{t\in[0,1]} |\widehat{\sigma}_{l_1,l_2}(t,t) - \sigma_{l_1,l_2}(t,t)| = O_P\left(g_{\max}^2 + 1/(J_S g_{\max})^{\frac{1}{2}}\right). \tag{20}$$

The proofs of these four theorems are given in a supplementary file.

4 Numerical Study

In this section, we present some simulation results to investigate the numerical performance of the proposed MDySS procedures described earlier. In estimating the mean function $\boldsymbol{\mu}(t)$ and the covariance matrix function $\Sigma(s,t)$, p is fixed at 1, the kernel function is chosen to be the Epanechnikov kernel $K(u) = 0.75(1-u^2)I(|u| \leq 1)$, and all bandwidths are chosen by the CV procedure. In all examples, q is fixed at 5.

First, we consider cases when observation vectors within a subject are independent and normally distributed, the IC mean function $\boldsymbol{\mu}(t)$ and the IC covariance matrix function $\Sigma(t,t)$ are assumed known to be

$$\boldsymbol{\mu}(t) = (0, t, 1+0.2t+0.3t^2, 1-\exp(-10t), \cos(t))', \quad t \in [0,1],$$

$$\begin{aligned}\Sigma(t,t) &= \operatorname{diag}\{1, \exp(t), \frac{1}{1+t}, 2, \log(t+5)\} \times \begin{pmatrix} 1 & 0.8 & 0.8^2 & 0.8^3 & 0.8^4 \\ 0.8 & 1 & 0.8 & 0.8^2 & 0.8^3 \\ 0.8^2 & 0.8 & 1 & 0.8 & 0.8^2 \\ 0.8^3 & 0.8^2 & 0.8 & 1 & 0.8 \\ 0.8^4 & 0.8^3 & 0.8^2 & 0.8 & 1 \end{pmatrix} \\ &\times \operatorname{diag}\{1, \exp(t), \frac{1}{1+t}, 2, \log(t+5)\}, \quad t \in [0,1].\end{aligned}$$

Note that different components of the above IC mean function $\boldsymbol{\mu}(t)$ change over time in different patterns; so do the components of the IC covariance matrix function $\Sigma(t,t)$. In such cases, the standardized observations of a new subject to monitor can be computed by $\boldsymbol{\epsilon}(t_j^*) = \Sigma^{-\frac{1}{2}}(t_j^*, t_j^*)(\boldsymbol{y}(t_j^*) - \boldsymbol{\mu}(t_j^*))$, for $j = 1, 2, \ldots$, and the control limits of the LASSO-based

MEWMA chart (13) presented in Table 1 can be used for the MDySS-L procedure. The control limits of the MEWMA charts (10) and (11) can be computed easily by a numerical algorithm in the same way as we compute the control limit of the chart (13). Now, let us consider a mean shift occurring at the initial time point from $\boldsymbol{\mu}(t)$ to

$$\boldsymbol{\mu}_1(t) = \boldsymbol{\mu}(t) + \Sigma(t,t)\boldsymbol{\delta}, \text{ for } t \in [0,1],$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3, \delta_4, \delta_5)'$ denotes the standardized shift size. Next, we consider 20 standardized shift sizes listed in Table 2, labeled from 1 to 20. The non-shift case (i.e., $\boldsymbol{\delta} = \mathbf{0}$) is included in the table as well, labeled by 0, for investigating the IC performance of the related procedures. These shift sizes contain many different cases with different numbers of shifted components. The basic time unit in this example is $\omega = 0.001$, and the sampling rate d is chosen to be 2, 5 or 10. The observation times are generated by randomly choosing d times without replacement from every 10 basic time points. We first consider the MDySS-L procedure, in which the ATS_0 value is fixed at 100 and the weighting parameter λ_L is chosen to be 0.05, 0.1 or 0.2 in the chart (13). The corresponding h_L values can be found from Table 1. The ATS_0 and ATS_1 values of the chart computed based on 10,000 replicated simulations are shown in Figure 1. From the figure, it can be seen that: (i) the chart with relatively small λ_L values performs better for detecting small shifts, and it performs better for detecting large shifts when λ_L is chosen relatively large, and (ii) the ATS_1 values tend to be smaller when d is larger. The first result is generally true for the LASSO-based MEWMA charts as discussed in [24], and the second result is intuitively reasonable because more observations are available for process monitoring when d is larger and consequently a shift can be detected faster.

Next, we consider a more realistic situation when the IC mean function $\boldsymbol{\mu}(t)$ and the IC covariance matrix function $\Sigma(s,t)$ are both unknown and they need to be estimated from an IC dataset of a group of m well-functioning subjects. For each subject, it is still assumed that the sampling rate is d, and the basic time unit is $\omega = 0.001$. After $\boldsymbol{\mu}(t)$ and $\Sigma(s,t)$

Table 2: Twenty shift sizes $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3, \delta_4, \delta_5)'$ considered in the numerical study.

$\boldsymbol{\delta}$	δ_1	δ_2	δ_3	δ_4	δ_5
0	0.00	0.00	0.00	0.00	0.00
1	0.00	0.25	0.00	0.00	0.00
2	0.00	0.00	0.50	0.00	0.00
3	0.00	0.00	0.00	0.75	0.00
4	0.00	0.00	0.00	0.00	1.00
5	0.25	0.00	0.25	0.00	0.00
6	0.00	0.25	0.50	0.00	0.00
7	0.00	0.25	0.00	0.75	0.00
8	0.00	0.50	0.00	0.00	0.50
9	0.50	0.00	0.00	1.00	0.00
10	0.75	0.75	0.00	0.00	0.00
11	1.00	0.00	0.00	0.00	1.00
12	0.25	0.25	0.25	0.00	0.00
13	0.25	0.00	0.50	0.00	1.00
14	0.50	0.00	0.00	0.50	0.50
15	0.50	0.00	1.00	0.50	0.00
16	1.00	0.00	1.00	0.00	1.00
17	0.50	0.50	0.50	0.50	0.50
18	0.25	0.25	0.50	0.75	1.00
19	0.25	0.25	0.75	0.50	0.50
20	0.75	1.00	0.50	1.00	0.75

are estimated from the IC dataset, observations of a new subject can be standardized by (8) for online monitoring. The standardized observations $\{\widehat{\boldsymbol{\epsilon}}(t_j^*)\}$ are asymptotically i.i.d. with the common distribution $N(\mathbf{0}, I_{q\times q})$ when the new subject is IC. The control limit h_L is still chosen to be those in Table 1. Therefore, the performance the chart (13) should depend on the IC sample size m. When m gets larger, its performance should be more reliable. For each generated IC dataset, the ATS_0 or ATS_1 value of the chart (13) is computed based on 10,000 replicated simulations. Then, the entire process, starting from the generalization of the IC dataset to the computation of the actual ATS_0 or ATS_1 value, is repeated 100 times. The averaged actual ATS_0 and ATS_1 values, are presented in Figure 2, in cases when the nominal ATS_0 is chosen 100, λ_L= 0.2, d= 2, 5 or 10, and m=30, 50 or 70. From the figure, it can be seen that: (i) the actual ATS_0 values are all within 10% of the nominal ATS_0 value

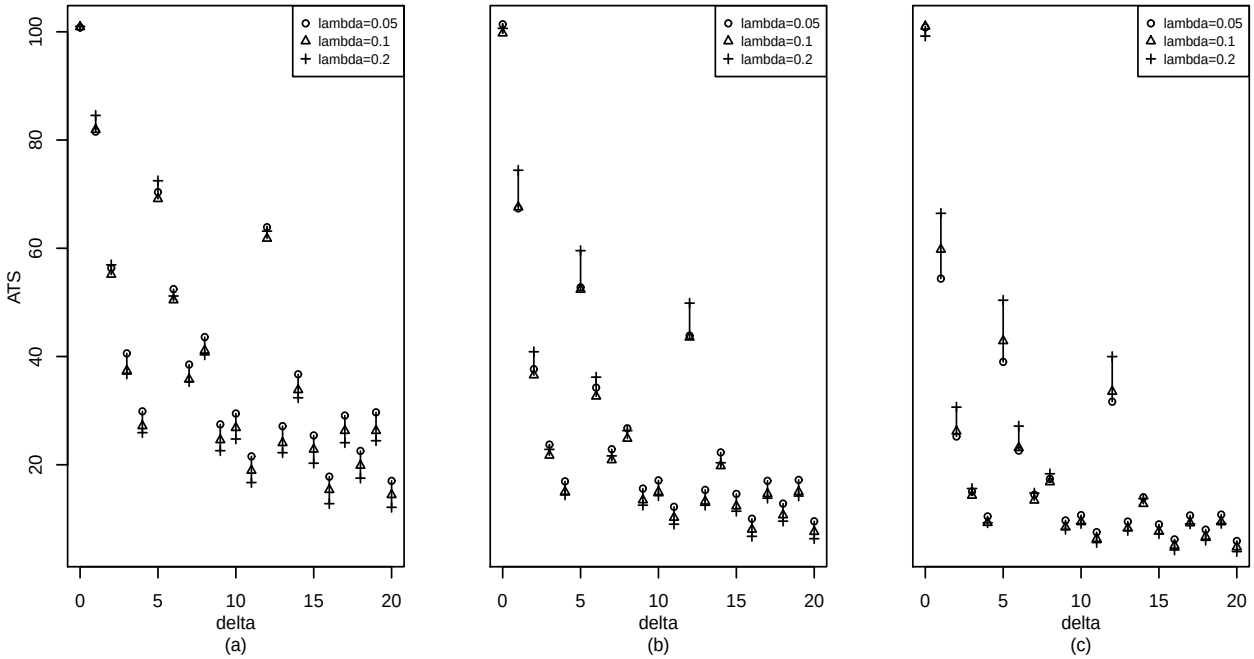

Figure 1: Actual ATS values of the chart (13) for detecting step mean shifts of size $\boldsymbol{\delta}$ occurring at the initial time point, in cases when observation vectors within a subject are independent and normally distributed, the IC mean function $\boldsymbol{\mu}(t)$ and the IC covariance matrix function $\Sigma(s,t)$ are assumed known, $d = 2$ (plot (a)), 5 (plot (b)) or 10 (plot (c)), $\omega = 0.001$, $\lambda_L = 0.05$, 0.1 or 0.2, and the nominal ATS_0 is 100.

of 100, except the case when $m = 30$ and $d = 10$, (ii) the ATS_1 values are generally smaller if the shift size $\boldsymbol{\delta}$ is larger or d is larger, and (iii) the ATS_1 values do not depend on the value of m much, especially when $\boldsymbol{\delta}$ and d are large. The corresponding cases when λ_L= 0.05 or 0.1 and the covariance matrix $\Sigma(s,t)$ is unexchangeable are discussed in the supplementary file. The results show that similar conclusions can be made.

Next, we compare the three procedures MDySS-M, MDySS-C, and MDySS-L in the setup of the above example when m is fixed at 70. In the three procedures, the weighting parameters λ_M, λ_C, and λ_L are all chosen to be 0.2, and the nominal ATS_0 is fixed at 100. The other parameters are kept to be the same as those in the above example. The computed actual ATS_0 and ATS_1 values of the three procedures are presented in Figure 3. From this figure and Table 2, we can have the following conclusions. First, the MDySS-L performs well in cases when the shift in one component is much larger than the shifts in the remaining

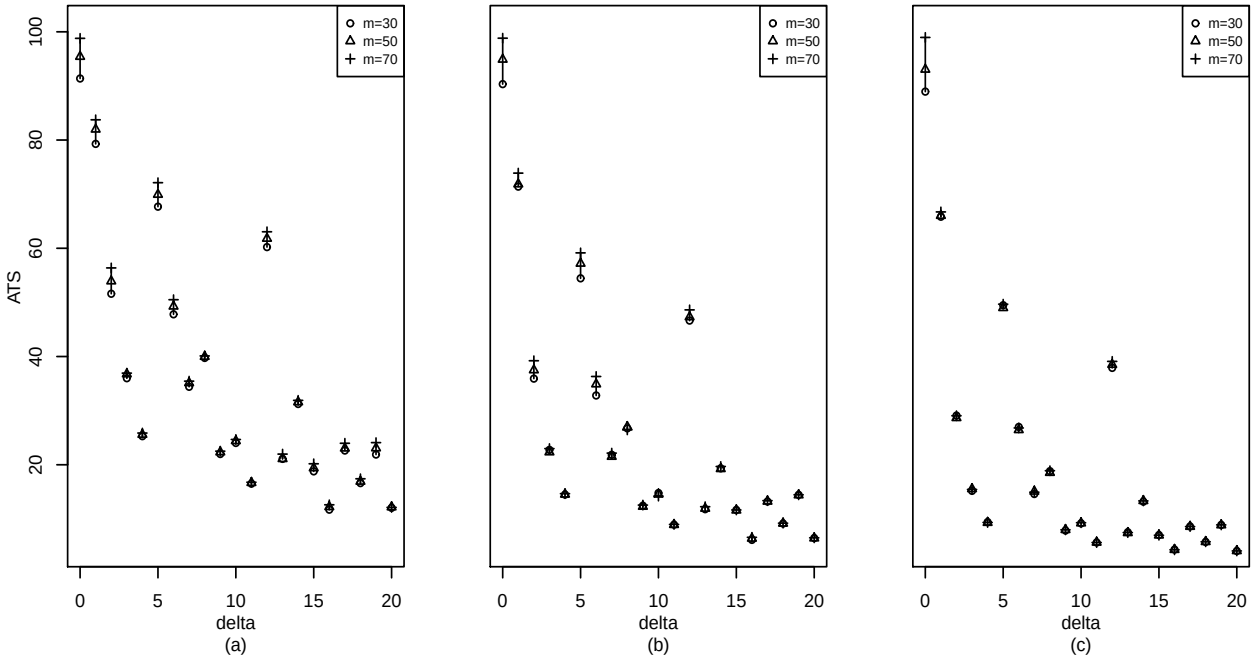

Figure 2: Actual ATS values of the chart (13), for detecting step mean shifts of the size $\boldsymbol{\delta}$ occurring at the initial time point, in cases when the IC mean function $\boldsymbol{\mu}(t)$ and the IC covariance matrix function $\Sigma(s,t)$ are estimated from an IC dataset with m subjects, $d = 2$ (plot (a)), 5 (plot (b)) or 10 (plot (c)), $m = 30, 50$ or 70, $\omega = 0.001$, $\lambda_L = 0.2$, and the nominal ATS_0 value is 100.

components. Second, the MDySS-C procedure performs well in cases when all components have shifts and the componentwise shifts are all quite large. Third, the performance of the procedure MDySS-M is generally between the performance of the other two procedures. So, if we know in advance that the potential mean shift occurs in most components and the shift sizes are quite large, then we can consider using the MDySS-C procedure. If the potential mean shift can only affect a small number of components, then we can consider using the MDySS-L procedure. If there is no such prior information about the potential mean shift, then the MDySS-M procedure can be considered. We also performed simulations in cases when $\lambda_M = \lambda_C = \lambda_L =$0.05 or 0.1, and the results are given in the supplementary file. Similar conclusions can be made. Results in a case when the covariance structure is not exchangeable are given in the supplementary file as well. The three procedures have a reasonably good performance in that case too.

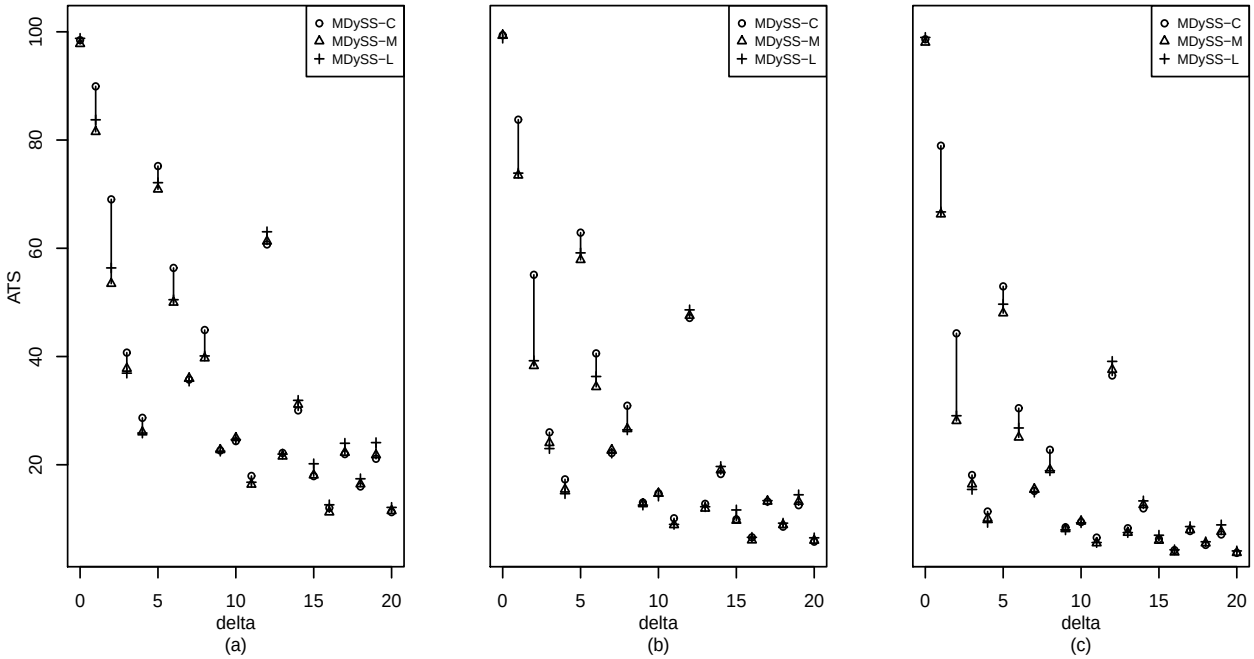

Figure 3: Actual ATS values of the procedures MDySS-M, MDySS-C, and MDySS-L, for detecting step mean shifts of the size $\boldsymbol{\delta}$ occurring at the initial time point, in cases when the IC mean function $\boldsymbol{\mu}(t)$ and the IC covariance matrix function $\Sigma(s,t)$ are estimated from an IC dataset with m subjects, $m = 70$, $d = 2$ (plot (a)), 5 (plot (b)) or 10 (plot (c)), $\omega = 0.001$, $\lambda_M = \lambda_C = \lambda_L = 0.2$, and the nominal ATS_0 is 100.

Next, we consider an example in which $q = 5$, and the within-subject observation vectors are neither independent nor normally distributed. More specifically, it is assumed that the IC mean function $\boldsymbol{\mu}(t)$ is the same as that in the above example. The error term $\boldsymbol{\varepsilon}(t_{ij})$ in model (1) follows the time series model

$$\boldsymbol{\varepsilon}(t_{ij}) = 0.5\boldsymbol{\varepsilon}(t_{i,j-1}) + \boldsymbol{e}(t_{ij}), \qquad \text{for any } i, j,$$

where $\boldsymbol{e}(t_{ij})$'s are independent and identically distributed random vectors having the common

multivariate t distribution with mean $\mathbf{0}$, degrees of freedom 4, and the covariance matrix

$$\Sigma_t = \begin{pmatrix} 1 & 0.8 & 0.8^2 & 0.8^3 & 0.8^4 \\ 0.8 & 1 & 0.8 & 0.8^2 & 0.8^3 \\ 0.8^2 & 0.8 & 1 & 0.8 & 0.8^2 \\ 0.8^3 & 0.8^2 & 0.8 & 1 & 0.8 \\ 0.8^4 & 0.8^3 & 0.8^2 & 0.8 & 1 \end{pmatrix}.$$

In such cases, it can be checked that the IC covariance matrix function of model (1) is

$$\Sigma(s,t) = \begin{cases} \frac{8}{3}\Sigma_t, & \text{when } s = t \\ \frac{4}{3}\Sigma_t, & \text{when } s \neq t. \end{cases}$$

Now assume that there is an IC dataset with observations from m well-functioning subjects. Part of the IC dataset is used for obtaining estimators $\widehat{\boldsymbol{\mu}}(t;\widetilde{\Sigma})$ and $\widehat{\Sigma}(s,t)$ of the IC mean and covariance matrix functions, as discussed in Subsection 2.1, and the remaining part is used for computing the control limits of the three control charts by the block bootstrap procedure with $B = 10,000$, as discussed at the end of Section 2. Let us first consider the procedure MDySS-L in cases when $\omega = 0.001$, $d = 2$, $m_1 = m/5$ =30, 50 or 70, $\lambda_L = 0.2$, and the 20 shifts presented in Table 2 are considered, which are assumed to occur at the initial time point. For each simulated IC dataset, the actual ATS_0 and ATS_1 of the chart are first computed based on 10,000 replicated simulations of online monitoring. Then, the entire process, including the generation of the IC data and the computation of the actual ATS values, are repeated 100 times. The averaged actual ATS_0 and ATS_1 values and the corresponding standard errors are presented in Figure 4. From the figure, it can be seen that: (i) the actual ATS_0 values are within about 10% of the nominal ATS_0 value of 100 in cases when $m_1 \geq 50$, (ii) the ATS_1 values do not depend on the m_1 value much when $m_1 \geq 50$, and (iii) the procedure MDySS-L performs worse in the current situation, compared to its performance in cases when the within-subject observations are assumed independent and normally distributed (cf., plot (a) in Figure 2).

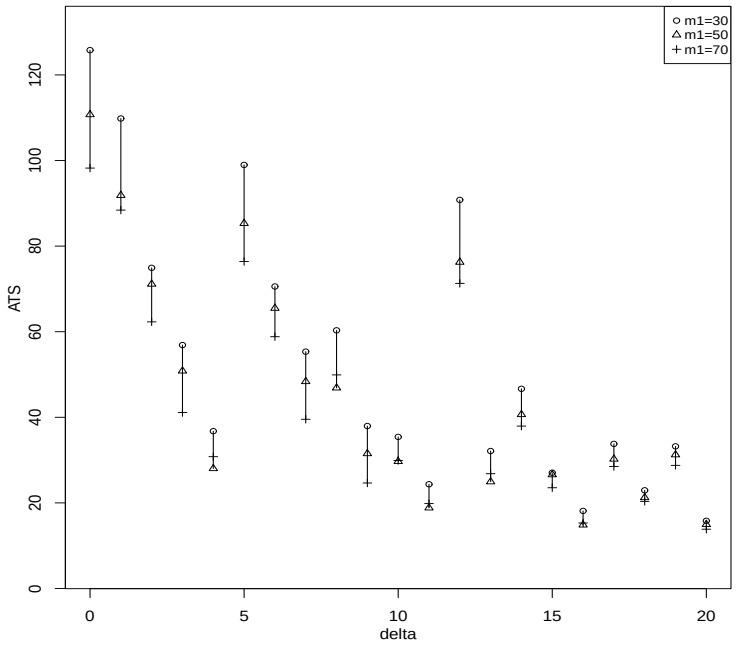

Figure 4: Actual ATS values of the chart (13), for detecting step mean shifts of the size $\boldsymbol{\delta}$ occurring at the initial time point, in cases when the IC mean function $\boldsymbol{\mu}(t)$ and the IC covariance matrix function $\Sigma(s,t)$ are estimated from an IC dataset with m_1 subjects, $m_1 = 30, 50$ or 70, $d = 2$, $\omega = 0.001$, $\lambda_L = 0.2$, and the nominal ATS_0 is 100.

Next, we compare the three procedures MDySS-M, MDySS-C, and MDySS-L in the setup of the above example when m_1 is fixed at 70. In the three procedures, the weighting parameters λ_M, λ_C, and λ_L are all chosen to be 0.2, and the nominal ATS_0 is fixed at 100. The other parameters are kept to be the same as those in the above example. The computed actual ATS_0 and ATS_1 values of the three procedures are presented in Table 3. From the table, it can be seen that similar conclusions to those from Figure 3 can be made here, and all procedures perform worse in the current situation, compared to their performance in cases when the within-subject observations are assumed independent and normally distributed (cf., plot (a) in Figure 3).

At the end of this section, we would like to point out that the computation involved in the proposed MDySS procedure is actually quite easy because the mean function $\mu(t)$ and the covariance function $\Sigma(s,t)$ are estimated by *local* smoothing procedures. For instance,

Table 3: Actual ATS_0 (in the first row) and ATS_1 values of the procedures MDySS-M, MDySS-C, and MDySS-L, along with their standard errors (in parentheses), for detecting step mean shifts of the size $\boldsymbol{\delta}$ occurring at the initial time point, in cases when the IC mean function $\boldsymbol{\mu}(t)$ and the IC covariance matrix function $\Sigma(s,t)$ are estimated from an IC dataset with m_1 subjects, $m_1 = 70$, $d = 2$, $\omega = 0.001$, $\lambda_M = \lambda_C = \lambda_L = 0.2$, and the nominal ATS_0 is 100.

$\boldsymbol{\delta}$	MDySS-C	MDySS-M	MDySS-L
0	101.53(0.58)	101.62(0.67)	98.23(0.58)
1	96.35(0.62)	94.45(0.58)	88.41(0.50)
2	74.19(0.49)	67.19(0.39)	62.30(0.41)
3	51.02(0.35)	45.61(0.43)	41.11(0.42)
4	32.49(0.16)	34.21(0.19)	30.80(0.18)
5	80.85(0.56)	81.69(0.52)	76.40(0.47)
6	60.84(0.40)	62.91(0.40)	58.83(0.37)
7	44.63(0.34)	43.51(0.40)	39.54(0.39)
8	50.50(0.29)	52.38(0.31)	49.91(0.28)
9	28.37(0.20)	26.64(0.25)	24.65(0.24)
10	27.59(0.14)	31.09(0.15)	29.92(0.14)
11	20.16(0.10)	20.54(0.10)	19.84(0.09)
12	68.45(0.42)	74.90(0.47)	71.29(0.40)
13	25.09(0.12)	28.14(0.17)	26.82(0.16)
14	35.06(0.17)	39.03(0.24)	37.95(0.24)
15	22.91(0.10)	24.23(0.12)	23.54(0.11)
16	14.12(0.06)	15.23(0.08)	15.30(0.08)
17	26.32(0.09)	28.43(0.13)	28.50(0.14)
18	18.80(0.07)	20.59(0.09)	20.36(0.09)
19	26.87(0.10)	29.25(0.15)	28.77(0.14)
20	11.78(0.04)	13.51(0.06)	13.85(0.06)

in the example of Figure 1 when $d = 10$, $q = 5$ and $\omega = 0.001$, it takes about 30 minutes to compute the estimators of $\mu(t)$ and $\Sigma(s,t)$ using an Intel CORE i5-4210U 2.4-GHz CPU. The computation time for finding each control limit of the charts (10)-(12) is about 20 minutes in the set up of that example. Because all these computations are required for only once, before the Phase II online process monitoring, this computation cost is regarded as small. For Phase II online process monitoring, the proposed MDySS procedure is as computationally efficient as most existing multivariate control charts. For instance, in the set up of Figure 1, the averaged computing time of each online run using the chart (12) is about 0.003 seconds. This speed is acceptable for most applications. MatLab codes for implementing the proposed

procedure are available from the authors upon request.

5 Analysis of the Stroke Data

In this section, we apply our proposed MDySS method to the stroke data of the SHARe Framingham Heart Study. As described in Section 1, there were 945 patients who did not experience any strokes in the study (i.e. $m = 945$), and 27 patients experienced at least one stroke. Each patient was followed 7 times (i.e., $J_i = 7$ for each i), and four medical indices (i.e., $q = 4$), including the systolic blood pressure (mmHg), diastolic blood pressure (mmHg), total cholesterol level (mg/100ml), and glucose level (mg/100ml), were recorded at each time. For a more complete description of this data, see [34]. Because these four medical indices, denoted as $\boldsymbol{y}$, are important risk factors of stroke, it is important to detect their irregular longitudinal patterns so that some medical interventions can be made in a timely manner to avoid strokes.

To apply the proposed MDySS method, the 945 non-stroke patients are used as the IC data. The analysis in [8] confirmed that within-subject observations were correlated and not normally distributed. Therefore, the IC data will be used for estimating the regular longitudinal pattern of $\boldsymbol{y}$ and for designing the MDySS procedure as well. In this example, because we expect a small number of the four medical indices would shift from their regular longitudinal patterns at the initial stage of a vascular disease, the MDySS-L procedure is considered. Then, the first 514 patients in the IC dataset are used for estimating the IC mean function $\boldsymbol{\mu}(t)$ and the IC covariance matrix function $\Sigma(s,t)$, and the remaining 431 patients are used for designing the MDySS-L procedure, as discussed at the end of Section 2. In this dataset, all observation times range from 16 to 83 years old, and the natural basic time unit is 1 year old. In the MDySS-L procedure, we choose $\lambda_L = 0.2$ and $ATS_0 = 20$. Then, the control limit h_L is computed to be 1.19. The LASSO-based MEWMA chart (13)

is shown in Figure 1, after it is applied to the 27 stroke patients to monitor their longitudinal patterns of $\boldsymbol{y}$. From the figure, it can be seen that it gives signals for 26 out of the 27 stroke patients. The signal times, defined as the differences between patients' ages when signaling and their ages at the beginning of process monitoring, are listed in Table 4. The average signal time, computed from the 26 stroke patients who get signals, is 11.84.

Table 4: Signal times (STs) of the 27 stroke patients by the LASSO-based MEWMA chart (13). The symbol "-" denotes no-signal.

ID	ST	ID	ST	ID	ST
1	20	10	8	19	12
2	12	11	0	20	19
3	0	12	12	21	20
4	15	13	12	22	13
5	7	14	8	23	7
6	15	15	13	24	12
7	8	16	16	25	-
8	7	17	8	26	12
9	12	18	15	27	25

6 Concluding Remarks

We have described a multivariate dynamic screening system for identifying irregular multivariate longitudinal patterns. This method combines the strengths of multivariate longitudinal data analysis and multivariate SPC. It makes decisions about the longitudinal pattern of a subject by comparing it with other subjects cross-sectionally and by sequentially monitoring it as well. Numerical examples presented in the previous two sections have shown that it is effective in various cases.

There are still some issues that need to be addressed in our future research. For instance, when the within-subject observations are correlated, some parametric time series models might be appropriate in some cases, although time series modeling is generally challenging in cases with irregularly spaced observation times. If a parametric time series model is

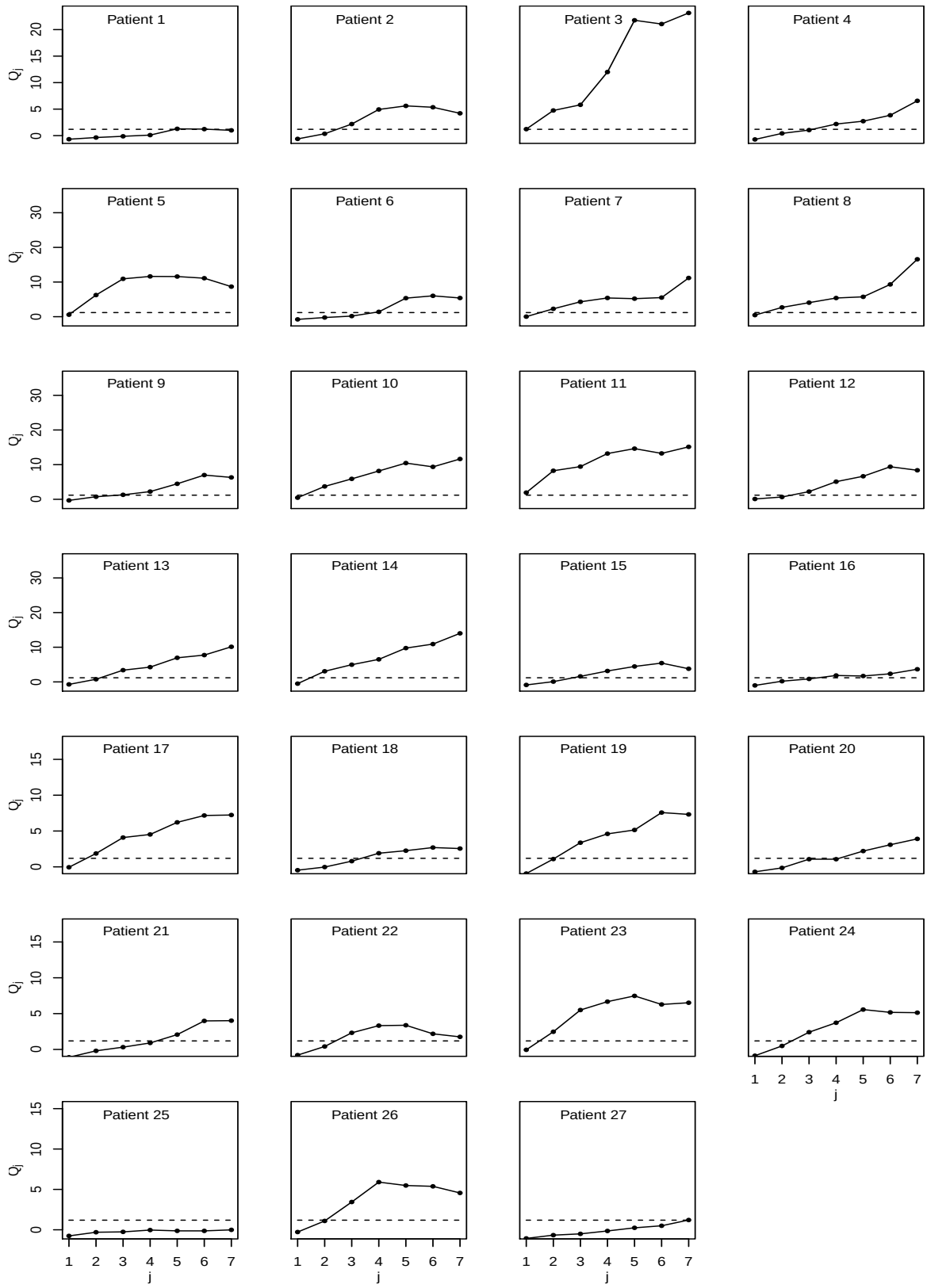

Figure 5: The LASSO-based MEWMA chart (13) when it is applied to the 27 stroke patients to monitor their longitudinal patterns of $\boldsymbol{y}$ in cases when we choose $\lambda_L = 0.2$ and $ATS_0 = 20$. The dashed horizontal line in each plot denotes the control limit.

confirmed to be appropriate in a given application, then our MDySS method should be more effective after accommodating such a parametric time series model. In the current version of the MDySS method, such parametric time series modeling is ignored completely. Also, when the observation vectors are not normally distributed, some nonparametric multivariate SPC charts might be more appropriate to use [31, 35, 36], compared to the MEWMA charts (10), (11) and (13) considered in the paper.

Acknowledgments: We thank the two referees for some constructive comments and sugegstions which greatly improved the quality of the paper.

References

1. Chen K, Jin Z. Local polynomial regression analysis of clustered data. *Biometrika* 2005; **92**: 59–74.

2. Li Y. Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika* 2011; **98**: 355–370.

3. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.

4. Lin X, Carroll R. Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association* 2000; **95**: 520–534.

5. Lin X, Carroll R. Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* 2001; **96**: 1045–1056.

6. Ma S, Yang L, Carroll R. A simultaneous confidence band for sparse longitudinal regression. *Statistica Sinica* 2012; **22**: 95–122.

7. Wang N. Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* 2003; **90**: 43–52.

8. Xiang D, Qiu P, Pu X. Nonparametric regression analysis of multivariate longitudinal data. *Statistica Sinica* 2013; **23**: 769–789.

9. Zhao Z, Wu W. Confidence bands in nonparametric time series regression. *Annals of Statistics* 2008; **36**: 1854–1878.

10. Hawkins DM, Olwell DH. *Cumulative Sum Charts and Charting for Quality Improvement*. New York: Springer-Verlag, 1998.

11. Montgomery DC. *Introduction To Statistical Quality Control (6th edition)*. New York: John Wiley & Sons, 2009.

12. Qiu P. *Introduction to Statistical Process Control*, London: Chapman & Hall/CRC, 2014.

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

14. Qiu P, Zou C, Wang, Z. Nonparametric profile monitoring by mixed effects modeling (with discussions). *Technometrics* 2010; **52**: 265–277.

15. Capizzi G, Masarotto G. A least angle regression control chart for multidimensional data. *Technometrics* 2011; **53**: 285–296.

16. Croisier RB. Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics* 1988; **30**: 243–251.

17. Hawkins DM. Multivariate quality control based on regression-adjusted variables. *Technometrics* 1991; **33**: 61–75.

18. Hawkins DM, Maboudou-Tchao EM. Self-starting multivariate exponentially weighted moving average control charting. *Technometrics* 2007; **49**: 199–209.

19. Healy JD. A note on multivariate CUSUM procedure. *Technometrics* 1987; **29**: 409–412.

20. Lowry CA, Woodall WH, Champ CW *et al.* Multivariate exponentially weighted moving average control chart *Technometrics* 1992; **34**: 46–53.

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

22. Woodall WH, Ncube MM. Multivariate CUSUM quality-control procedures. *Technometrics* 1985; **27**: 285–292.

23. Zamba KD, Hawkins DM. A multivariate change-point for statistical process control. *Technometrics* 2006; **48**: 539–549.

24. Zou C, Qiu P. Multivariate statistical process control using LASSO. *Journal of the American Statistical Association* 2009; **104**: 1586–1596.

25. Tibshirani, RJ. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society (Series B)* 1996; **58**: 267–288.

26. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 2006; **101**: 1418–1429.

27. Efron B, Hastie T, Johnstone I *et al.* Least angle regression. *Annals of Statistics* 2004; **32**: 407–489.

28. Costa AFB. Joint X and R charts with variable parameters. *IIE Transactions* 1998; **30**: 505–514.

29. Reynolds MR, Amin RW, Arnold JC. CUSUM charts with variable sampling intervals. *Technometrics* 1990; **32**: 371–384.

30. Wu Z, Zhang S, Wang P. A CUSUM scheme with variable sample sizes and sampling intervals for monitoring the process mean and variance. *Quality and Reliability Engineering International* 2007; **23**: 157–170.

31. Qiu P. Distribution-free multivariate process control based on log-linear modeling. *IIE Transactions* 2008; **40**: 664–677.

32. Pan X, Jarrett J. Applying state space to SPC: monitoring multivariate time series. *Journal of Applied Statistics* 2004; **31**: 397-418.

33. Lahiri SN. *Resampling Methods for Dependent Data*. New York: Springer, 2003.

34. Cupples LA et al. The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Medical Genetics* 2007; **8**(Suppl 1): S1.

35. Qiu P, Hawkins DM. A rank-based multivariate CUSUM procedure. *Technometrics* 2001; **43**: 120–132.

36. Qiu P, Hawkins DM. A nonparametric multivariate CUSUM procedure for detecting shifts in all directions. *Journal of the Royal Statistical Society (Series D) - The Statistician* 2003; **52**: 151–164.