An Efficient On-line Monitoring Method for High-dimensional Data Streams

Changliang Zou[1], Wei Jiang[2], Zhaojun Wang[1] and Xuemin Zi[3]

[1]*Institute of Statistics, Nankai University, Tianjin, China*

[2]*Antai College of Economics and Management, Shanghai Jiaotong University, Shanghai, China*

[3] *School of Science, Tianjin University of Technology and Education, China*

Abstract

Monitoring high-dimensional data streams has become increasingly important for real-time detection of abnormal activities in many data-rich applications. We are interested in detecting an occurring event as soon as possible, but we do not know which subset of data streams is affected by the event. By connecting to the problem of detecting heterogeneous mixtures, a new control chart is developed based on a powerful goodness-of-fit test of the local cumulative sum statistics from each data stream. Numerical results show that the proposed method is able to balance the detection of various fractions of affected streams, and generally outperforms existing methods.

Keywords: Change-point detection; CUSUM; Goodness-of-fit test; Higher criticism; Multiple testing; Sequential detection; Statistical process control

1 Introduction

Statistical approaches to continual surveillance of multiple data streams are greatly needed in today's industrial, clinical, and epidemiological environments. Among them, the problem of global on-line monitoring a large number of independent data streams through sequential observations has become increasingly important. For example, multi-sensor change-point detection, where sensors are distributed to monitor emergence of an event signal, has attracted considerable attentions recently. See Tartakovsky and Veeravalli (2008), Guerriero et al. (2009), Woodall and Montgomery (2014) and the references therein. Another typical example is the monitoring of multistage processes. Many manufacturing operations, such as semiconductor manufacturing and automotive body assembly, comprise a large number of stages. A commonly used method is to globally monitor the so-called one-step forecast errors (residuals) obtained from all the stages. Those residuals are usually independent under certain assumptions. See Li and Tsung (2009) for detailed illustration. Similar applications include monitoring business processes with high-dimensional transactions and detecting fraudulent records among them (c.f., Tsung et al. 2007; Jiang et al. 2007), and the biological problem of detecting recurrent DNA copy number variants in multiple samples (c.f., Zhang et al. 2010). The main objective is to detect changes in the process, which occur at an unknown time point as early as possible after the occurrence, and at the same time control the global rate of false alarms.

Suppose we are monitoring p data streams, observing the kth data stream X_{kt} over time $t = 1, 2, \ldots$. We firstly assume that the data streams are mutually independent and then discuss the dependent cases later. Under the null hypothesis (in-control; IC), the observations X_{kt}'s are independently and identically distributed (i.i.d.) normal random variables with mean μ_{0k} and variance σ_{0k}^2. Without loss of generality, we assume $\mu_{0k} = 0$ and $\sigma_{0k}^2 = 1$. Otherwise, we can always use $(X_{kt} - \mu_{0k})/\sigma_{0k}$ to standardize X_{kt}. Under the alternative hypothesis (out-of-control; OC), certain data streams incur mean changes at some unknown change-points. We denote the affected and unaffected subsets of data streams as $\mathcal{A}_a$ and $\mathcal{A}_a^c$, respectively. Let p_a denote the cardinality of $\mathcal{A}_a$, i.e., the number of affected streams. After the change-point τ_k, for $k \in \mathcal{A}_a$, the mean of X_{kt} changes from 0 to μ_k. For simplicity, here we assume that all the affected data streams incur changes at the same time point $\tau_k = \tau$; however, the methods discussed below are also applicable when τ_k's are different.

The problem is to raise an alarm as quickly as possible after the change occurs. In the change-point detection problem, a detection scheme is a stopping time T with respect to the observed data sequences $\mathbf{X}_t = \{X_{1t}, \ldots, X_{pt}\}_{t\geq 1}$. We use an alarm system consisting of two parts at stage s: an alarm statistic $a(\{\mathbf{X}_t\}_{t=1}^{s})$ and an alarm limit $g(s)$. The time of an alarm, T, is defined as $T = \min\{s; a(\{\mathbf{X}_t\}_{t=1}^{s}) \geq g(s)\}$. That is, the decision $T = s$ depends only on the first s observations of dimension p, and $T = s$ means that an alarm is triggered at time s to indicate that a change has occurred somewhere in the first s observations. Consistent with the literature, we focus on using a series of one-sided charts to detect changes of mean levels in some data streams, but the two-sided charts can be constructed without difficulty. As a convention in the practice of monitoring methods, we first develop our procedure under the assumption that the true values of $\mu_k, k \in \mathcal{A}_a$ are all known. In a later section, we will consider how to extend the proposed method to more general cases when μ_k's cannot be specified before monitoring.

Directly applying single-stream detection procedures for each sensor, such as cumulative sum (CUSUM) or exponentially weighted moving average (EWMA) control scheme, suffers from high false alarm rate and accordingly a global decision procedure that employs observations from all streams would usually be desired. Considerable research has been developed on designing an efficient global monitoring procedure. If the affected streams were known, we could use the likelihood ratio test to formulate a CUSUM and to only monitor those p_a streams. Such a scheme possesses certain optimal properties due to the use of likelihood procedure (Moustakides 1986; Veeravalli 2001), but it is generally not applicable because $\mathcal{A}_a$ is unknown in practice. A practical procedure is the one proposed by Tartakovsky et al. (2006) in which it is assumed that there is exactly $p_a = 1$ out of p data streams being affected. Specifically, the stopping time is defined as $T_{\max} = \inf\{t : \max_{k=1,\ldots,p} S_k(t) \geq L\}$, where $S_k(t)$ is the individual CUSUM statistic for the kth data stream, taking the following recursive form

$$S_k(t) = \max\left\{0, S_k(t-1) + \mu_k(X_{kt} - \mu_k/2)\right\}. \tag{1}$$

In other words, this procedure is based on the maximum of individual CUSUM statistics and it is in fact the likelihood procedure corresponding to the scenario when one knows exactly $p_a = 1$. For the general case, theoretically we can derive likelihood ratio detection procedures incorporating observations from all sensors using standard statistical techniques. However, a drawback of these procedures is that they incur high computational complexity when p is large. See Section 3 in Mei (2010) for detailed formulations and discussion.

Alternatively, a robust scheme is proposed by Mei (2010) based on the sum of the local CUSUM statistics from each individual data stream, taking the form of

$$T_{\text{sum}} = \inf \left\{ t : \sum_{k=1}^{p} S_k(t) \geq L \right\}. \tag{2}$$

However, as shown by Mei (2010) through simulations, $T_{\max}$ is more effective than T_{sum} when potential changes occur in only a few data streams. On the other hand, $T_{\max}$ would be outperformed by T_{sum} by a large margin when the change occurs in a moderate or large number of streams. This is not surprising by noticing the "max" and "sum" operators are employed in $T_{\max}$ and T_{sum} respectively. See Section 4 for further numerical evidence. The goal of this paper is to propose a reasonably simple scheme which is capable of efficiently detecting changes no matter how many data streams are affected, i.e., a control chart which could balance the detection ability between $T_{\max}$ and T_{sum}.

As a side note, the problem of monitoring p-dimensional data streams is analogous to using a multivariate control chart on p variables. The T_{sum} and $T_{\max}$ procedures are essentially similar to the Croisier's (1988) multivariate CUSUM charting procedure and the Hawkins's (1991) regression-adjusted control chart, respectively, assuming the covariance matrix of observation vectors is a p-dimensional identity matrix. In the related literature of statistical process control (SPC), Wang and Jiang (2009), Zou and Qiu (2009), and Capizzi and Masarotto (2011) proposed variable-selection-based multivariate control charts, which are developed based on the sparsity assumption, i.e., in a high-dimensionality process the probability that all variables shift simultaneously is rather low. Their methods focus on the case that all the variables are correlated and generally require much more computational effort than T_{sum} and $T_{\max}$. Another related direction to tackle the problem of global on-line monitoring is to apply false discovery rate (FDR) control to on-line detection, e.g., see Benjamini and Kling (1999), Grigg and Spiegelhalter (2008), and Li and Tsung (2009). The basic idea is to test multiple CUSUM statistics $S_k(t)$ in the way of controlling the FDR. As shown by Li and Tsung (2009), such a FDR-based method is usually favorable in the fault-isolation (diagnosis) problem, but its change detection ability is similar to that of $T_{\max}$. See Section 2 for more discussions.

In this paper, motivated by the connection between global monitoring and testing heterogenous mixtures, we suggest a computationally efficient detection scheme based on a powerful goodness-of-fit (GOF) test proposed by Zhang (2002). Numerical results show that the proposed method is able to balance the detection of various fractions of affected streams, and generally preforms more robust than existing methods no matter how large p is. The remainder of the paper is organized as follows. In Section 2, we describe the mathematical formulation of testing heterogenous mixtures and suggest a powerful test which lays the groundwork for our proposal in the sequential detection problem. We then introduce the proposed scalable procedure followed by its extension in Section 3. Finite-sample performance comparison for detecting changes in multiple data streams is presented in Section 4. Section 5 contains a high-dimensional sensor detection example to illustrate the application of our proposed chart. Finally several remarks in Section 6 conclude the paper. Some technical details are provided in appendices, which are available online as supplementary materials.

2 Testing Heterogenous Mixtures

It should be pointed out that the global monitoring problem considered here is a natural on-line generalization of multiple hypothesis testing problems. Specially, the procedure T_{sum} and $T_{\max}$ can be viewed as performing omnibus tests on $\{S_1(t), \ldots, S_p(t)\}$ at each time point t based on their sum and extreme values, respectively. In other words, $S_1(t), \ldots, S_p(t)$ are similarly distributed when the process is IC, while the distributions of those affected streams would differ much from those of unaffected streams after the change occurs. A careful examination of the multiple hypothesis testing problem would shed light on how to construct efficient on-line detection schemes.

2.1 Higher criticism statistic

A closely related off-line problem of multiple hypothesis testing is the so-called *detection of heterogenous mixtures* considered by Donoho and Jin (2004). The problem is defined as follows. Given n independent observations units $\mathbf{X} = (X_1, \ldots, X_n)$. For each $1 \le i \le n$, we suppose that X_i has probability ε_n of being a non-null effect and probability $1 - \varepsilon_n$ of being a null effect. We model the null effects as samples from $N(0, 1)$ and non-null effects from $N(\mu_n, 1)$ for $\mu_n > 0$. The normality imposed here is consistent with literature and suffices to illustrate the basic idea. In practice, numerous other settings for the deployment in what follows have also been considered (c.f., Donoho and Jin 2004). Here, ε_n can be regarded as the proportion of non-null effects. The goal is to test whether any signals are present, say if $\varepsilon_n = 0$, or equivalently to test the joint null hypotheses $H_0 : X_i \overset{\text{iid}}{\sim} N(0, 1), 1 \le i \le n$ versus $H_1 : X_i \overset{\text{iid}}{\sim} (1 - \varepsilon_n)N(0, 1) + \varepsilon_n N(\mu_n, 1), 1 \le i \le n$. Thus, our on-line testing problem is essentially related to this detection of heterogenous mixtures problem by considering $\{S_1(t), \ldots, S_p(t)\}$ as $\{X_1, \ldots, X_n\}$ at each time point t.

In this detection problem, a major task is to characterize the so-called detection boundary, which is a curve that partitions the parameter space into two regions: the detectable region and the undetectable region. The asymptotic framework adopted in the literature is setting $\varepsilon_n = n^{-\beta}$ for $0 < \beta < 1$ and letting μ_n grows to infinity or degenerates to zero at an appropriate rate according to the value of β. In light of this, some papers focus on the so-called *sparse regime*: $\beta \in (1/2, 1)$ and $\mu_n = \sqrt{2r \log n}$ with $r \in (0, 1)$. The likelihood ratio test (LRT) of H_0 versus H_1 has a "detection boundary" defined in terms of the parameters $\beta \in (1/2, 1)$ and $r \in (0, 1)$ which is described as follows. Set

$$\rho^*(\beta) = \begin{cases} \beta - 1/2, & \text{if} \quad 1/2 < \beta \le 3/4, \\ (1 - \sqrt{1 - \beta})^2, & \text{if} \quad 3/4 < \beta < 1. \end{cases}$$

Then for $r > \rho^*(\beta)$, the LRT (which makes use of knowledge of β and r) is size and power consistent against H_1 as $n \to \infty$. Donoho and Jin (2004) considered several different statistics, among which the principal contenders are Tukey's "higher criticism" statistic HC_n^* defined by

$$HC_n^* \equiv \max_{1 \le i \le n} HC_{n,i}, \quad HC_{n,i} = \frac{\sqrt{n}(i/n - p_{(i)})}{\sqrt{p_{(i)}(1 - p_{(i)})}},$$

where $p_i = 1 - \Phi(X_i) \equiv \bar{\Phi}(X_i)$, $\Phi(\cdot)$ is the cumulative distribution function (C.D.F.) of standard normal distribution and $p_{(1)} < p_{(2)} < \cdots < p_{(n)}$ are the order statistics of the p-values. One rejects the null hypothesis when HC_n^* is large. Donoho and Jin (2004) showed that the test of H_0 versus H_1 based on HC_n^* is also size and power consistent for $r > \rho^*(\beta)$. It dominates several other tests based on multiple comparison procedures such as the sample maximum and false discovery rate (FDR). Cai et al. (2011) further considered the *dense regime* which is calibrated as follows: $\mu_n = n^{-r}, 0 < r < 1/2$ for $0 < \beta \le 1/2$. In this case, Cai et al. (2011) found the detection boundary $r = \rho^*(\beta) = 1/2 - \beta, \ \ 0 < \beta \le 1/2$, and the detectable region now corresponds to $r < \rho^*(\beta)$. Cai et al. (2011) showed that the adaptivity of the test based on HC_n^* still holds, i.e., HC_n^* is size and power consistent if $r < \rho^*(\beta)$ against H_1 as $n \to \infty$, similar to the LRT. However, HC_n^* does not need any specific information of the parameter β and r.

2.2 GOF test statistic

By examining the closeness of HC_n^* to the well-known GOF test statistic in Anderson and Darling (1952), we may expect that some efficient GOF statistics would be also effective in testing heterogenous mixtures. Zhang (2002) introduced a parameterization approach to establish powerful GOF tests based on the following log-likelihood ratio statistic

$$g_u = 2n \left\{ F_n(u) \ln\left(\frac{F_n(u)}{\Phi(u)}\right) + [1 - F_n(u)] \ln\left(\frac{1 - F_n(u)}{1 - \Phi(u)}\right) \right\}, \tag{3}$$

where $F_n(u)$ is the empirical C.D.F. of the sample $\{X_1, \ldots, X_n\}$, i.e., $F_n(u) = n^{-1} \sum_{j=1}^n I_{\{X_j \le u\}}$ with the indicator function $I_{\{\cdot\}}$. Then, some powerful test statistics can be constructed in the form $\sup_u g_u w(u)$ or $\int g(u) dw(u)$.

One of the most powerful tests introduced by Zhang (2002) is to use $dw(u) = [\Phi(u)(1 - \Phi(u))]^{-1} d\Phi(u)$ with the integral form, which leads to

$$Z_C = \int [\Phi(u)(1 - \Phi(u))]^{-1} g_u d\Phi(u), \tag{4}$$

and large values of Z_C reject the null hypothesis. Note that the function $[\Phi(u)(1 - \Phi(u))]^{-1}$ attains its minimum at $\Phi(u) = 1/2$, that is when u is the median of the sample. Intuitively, the more extreme observations (far way from the median) corresponding to large values of weight, the more informative to indicate the violation of H_0 and the weights may be chosen accordingly larger. As shown by Zhang (2002), it is much more powerful than the classical GOF tests, such as the Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling tests. In what follows, we will focus on this type of test for testing heterogenous mixtures.

2.3 Comparing HC and GOF statistics

We now develop a one-sided version of the statistic Z_C which has the same detection boundary as the statistic HC_n^* for testing H_0 versus H_1. Note that Z_C in (4) is equal to

$$\sum_{i=1}^{n} \left\{ \log \left[\frac{[\Phi(X_{(i)})]^{-1} - 1}{(n - 1/2)/(i - 3/4) - 1} \right] \right\}^2 + C_n, \tag{5}$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics and C_n is a constant (see Appendix A.3 in the supplemental file). We suggest the following one-sided statistic

$$D_{\text{GOF}} = \sum_{i=1}^{n} \left\{ \log \left[\frac{[\Phi(X_{(i)})]^{-1} - 1}{(n - 1/2)/(i - 3/4) - 1} \right] \right\}^2 I_{\{\Phi(X_{(i)}) > (i-3/4)/n\}}, \tag{6}$$

which will be shown to work reasonably well through numerical studies. Note that only those observations with p-values smaller than their expected levels have contributions in the test statistic (considering $\mu_n > 0$ under our assumption). Here $(i - 3/4)/n$ is a common "continuity correction" to $F_n(X_{(i)})$.

It can be shown that the test using D_{GOF} has the same detection boundary as the statistic HC_n^* for $\beta \in (0, 1)$. Please refer to Theorem 1 and its proof given in the supplemental file..

Similar to HC_n^*, the test statistic D_{GOF}, which does not require specific information on model parameters, can adapt to the unknown degrees of heterogeneity in both sparse and dense cases. In comparison, the conventional tests based on $\max_{1\leq i\leq n} X_i$ and $\sum_{i=1}^n X_i$ (denoted by $D_{\max}$ and D_{sum}) work only well for the sparse case $1/2 < \beta \leq 3/4$ and the dense case $0 < \beta \leq 1/2$, respectively. See Theorem 1.4 in Donoho and Jin (2004) and Theorem 8 in Cai et al. (2011) for detailed discussions. In addition, as revealed by Donoho and Jin (2004), the FDR-controlling procedure is not different from that of $D_{\max}$.

To illustrate the statistical performance of the HC and GOF statistics, some simulation results are presented in Figures 1 and 2. In these simulations, we fix $n = 10,000$ and use 2,500 replications for each scenario. Other simulation results which are not reported here indicate similar conclusions as long as n is not too small ($n \geq 100$). For the null hypothesis, we drew n samples from $N(0,1)$; for the alternative hypothesis, we first drew $n_a \equiv n\varepsilon_n$ samples from $N(\delta_i, 1), i = 1, \ldots, n_a$ and then drew $n - n_a$ samples from $N(0,1)$. For comparison, four tests are considered: D_{GOF}, $D_{\max}$, D_{sum}, and the test based on HC_n^* (denoted as D_{HC}). To make a fair comparison, we perform a size-corrected power comparison in the sense that the actual critical values are found through simulations so that all of the four tests have the exact size of 0.01 in each case.

In Figure 1, we focus on the effect of the signal strength, plotting the empirical power against the signal strength $\delta_i = \delta$ with various values of ε_n. For the sparse case $\varepsilon_n = 5\times 10^{-4}$, D_{GOF} has comparable performance to that of $D_{\max}$ and D_{HC} but D_{sum} does not work in this case. In contrast, for the very dense cases with $\varepsilon_n = 0.05$ or $\varepsilon_n = 0.5$, D_{sum} performs the best but $D_{\max}$ and D_{HC} do not work well in all those three cases. D_{GOF} is inferior to D_{sum} but still has reasonably good detection ability. In the moderately sparse cases such as $\varepsilon_n = 0.001$ or 0.01, D_{GOF} seems the best. It outperforms the others by a quite significant margin when $\varepsilon_n = 0.01$.

The effect of the level of sparsity (calibrated by ε_n) is further investigated in Figure 2. Here the empirical power is plotted against ε_n. Two patterns of allocation are considered for

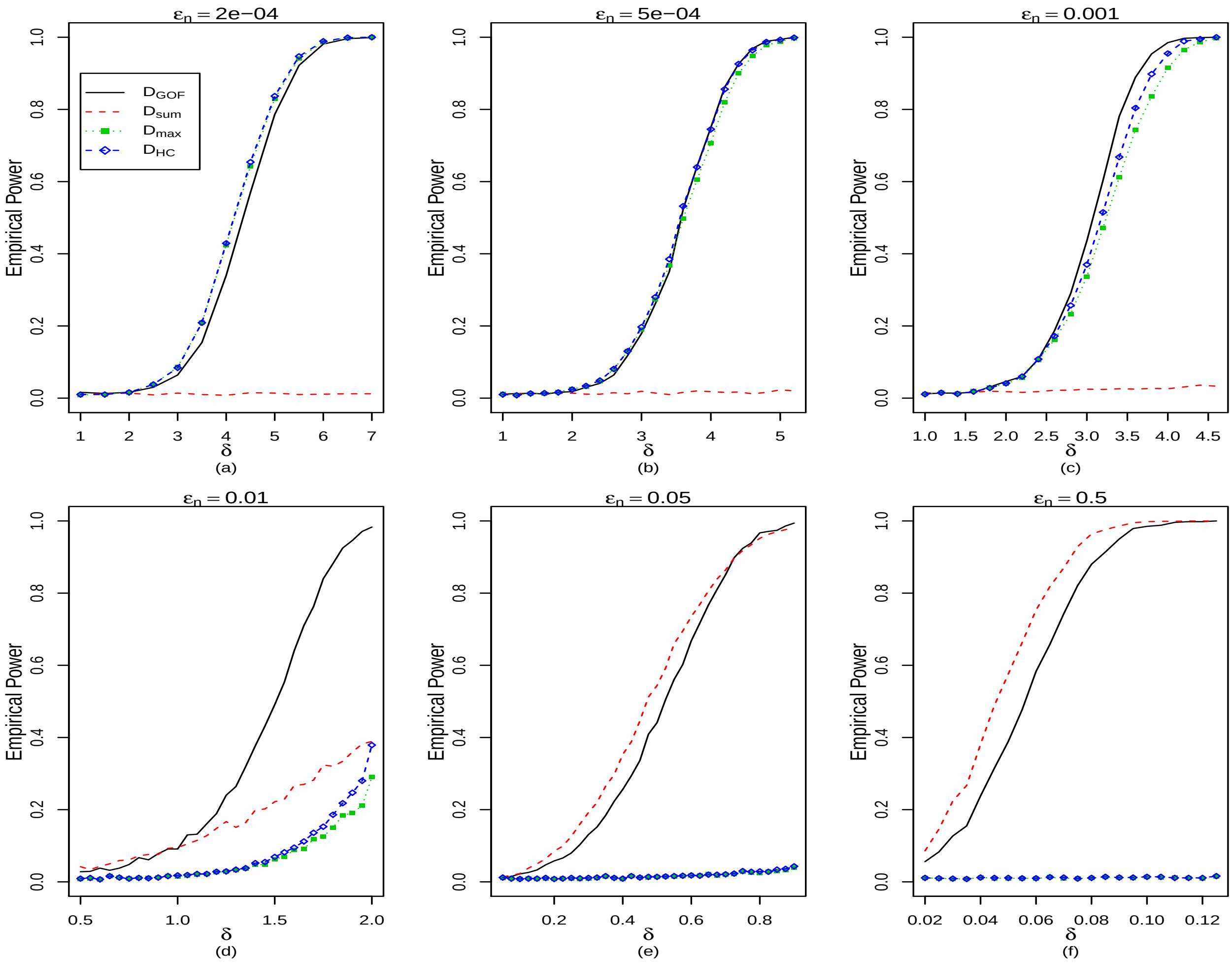

Figure 1: Size-corrected power comparison between D_{GOF}, $D_{\max}$, D_{sum} and D_{HC} under various ε_n. In each plot, the empirical power is plotted against μ_n (note the different scales). The legend in the first plot is applicable for all the others.

the nonzero expectation: the equal allocation where all the expectations of signal are equal to δ as in Figure 1; linearly increasing allocation $\delta_i = i\delta$ for $i = 1, \ldots, n_a$. Note that the linearly increasing allocation does not satisfy our original alternative hypothesis H_1 and is used for checking the robustness of tests against mis-specification. To make the power comparable among the configurations of H_1, we set $\eta = \sum_i \delta_i^2$ for each value of ε_n. Three values of η, 50, 100 and 200 are chosen. From this figure, we can observe that D_{GOF} offers a good balance protection against ε_n. This can be seen from the cases when D_{sum} works well but $D_{\max}$ does not or vice versa. In such cases, D_{GOF} may not be the best, but it is always close to the best.

In the cases when $0.003 \leq \varepsilon_n \leq 0.02$, D_{GOF} generally performs the best, and the margins are significant when η is large. In addition, the equal and increasing allocation scenarios present similar performance patterns. The above simulation study provides us a clear evidence that D_{GOF} is indeed powerful for testing heterogenous mixtures, and thus motivates us to develop corresponding scalable monitoring scheme in the next section.

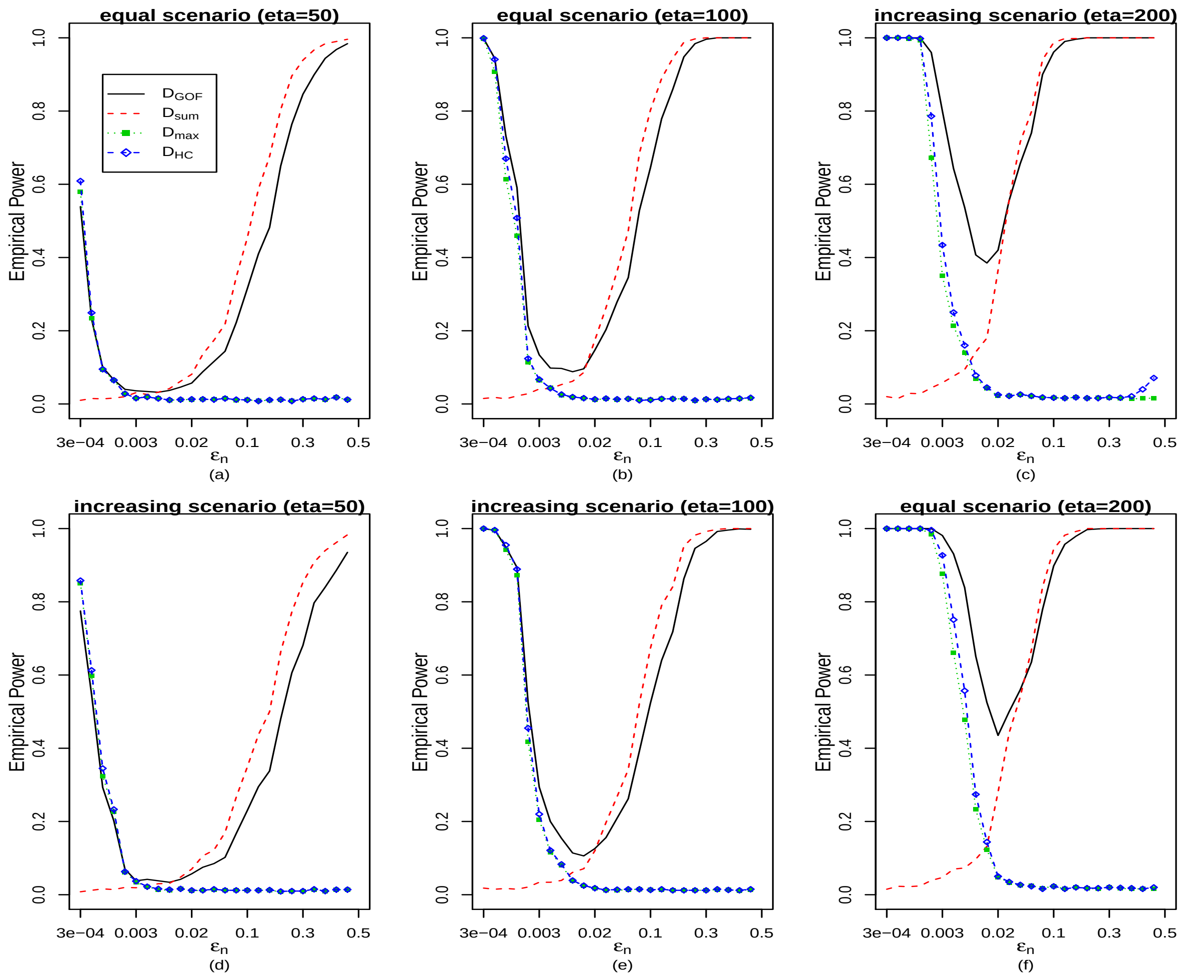

Figure 2: Size-corrected power comparison between D_{GOF}, $D_{\max}$, D_{sum} and D_{HC} under various signal strengths (η) for equal and linearly increasing scenarios. In each plot, the empirical power is plotted against ε_n.

3 Monitoring a Large Number of Data Streams

3.1 A new detection scheme

In light of the discussions in Section 2, the main idea of our proposal would be quite straightforward. Since D_{GOF} is powerful for testing heterogenous mixtures compared to D_{sum} and $D_{\max}$, it is natural to apply D_{GOF} to the local (individual) CUSUM statistics $S_k(t), k = 1, \ldots, p$, analogous to the construction of T_{sum} and $T_{\max}$. Replacing X_i by $S_k(t)$ in (6) yields

$$W_t = \sum_{i=1}^{p} \left\{ \log \left[\frac{U_{(i)}^{-1}(t) - 1}{(p - 1/2)/(i - 3/4) - 1} \right] \right\}^2 I_{\{U_{(i)}(t) > (i-3/4)/p\}}, \tag{7}$$

where $U_{(1)}(t) \le \cdots \le U_{(p)}(t)$ are the order statistics of $\{U_1(t), \ldots, U_p(t)\}$, $U_i(t) = H_t(S_i(t); \mu_i)$, and $H_t(\cdot; \mu)$ denotes the C.D.F. of $S_i(t)$ with specified parameter μ under in-control state. Here again we $(i - 3/4)/p$ as a continuity correction to the empirical C.D.F. value at $S_{(i)}(t)$. Accordingly, our proposed procedure is given by

$$T_{\text{new}} = \inf \{t : W_t \ge L\},$$

where $L > 0$ is a control limit chosen to achieve a specific value of IC average run length (ARL).

To apply the above monitoring scheme, $H_t(\cdot; \mu)$ must be obtained first. In practice, it is usually convenient to use the null steady-state distribution of the CUSUM statistic $H(\cdot; \mu)$, defined as the distribution of values obtained by running a CUSUM without threshold under the null state for an indefinite period of time. Based on some theoretical and empirical justifications, Grigg and Spiegelhalter (2008) developed an accurate approximation to the null steady-state distribution of the CUSUM statistic which is valid for CUSUMs applied to normal data. Their result leads to a closed-form formula summarized in Appendix A.2 of the supplemental file. The approximation greatly facilities the computation and implementation of T_{new}.

Note that T_{new} requires $O(p \log p)$ computations and $O(p)$ memory allocations at every time point t. Despite the requirement of more computational efforts than $T_{\max}$ and T_{sum}, T_{new} is still scalable and can be easily implemented for large p over long time periods. Our unreported simulation results show that T_{new} has comparable performances with the LASSO-based EWMA chart (LEWMA) proposed by Zou and Qiu (2009) when the between-streams correlations are weak, but the LEWMA is generally more efficient when the stream observations are strongly correlated. The major benefit of using T_{new} is its computation. Although the LEWMA chart can well handle the high-dimensional monitoring with sparsity features by efficiently utilizing the correlation information, it requires about $O(p^3)$ computations which are not trivial when p is very large. Moreover, when $p = 1$, T_{new} can be viewed as a monotonously increasing transformation of the classical local CUSUM procedure provided that $S_1(t)$ is sufficiently large, i.e., $H(S_1(t); \mu_1) > 1/2$. In other words, T_{new} would have similar detection ability to the single CUSUM for monitoring a single data stream.

Similar to D_{GOF}, T_{new} is robust and omnibus in the sense that it is not only sensitive to different combinations of affected data streams but can also detect changes in variance. The detection of variance changes is not the focus of this paper but deserves some further study in the future. Moreover, T_{new} is also applicable to the cases that different data streams incur changes at different change-points, i.e., τ_k's are different. To see this clearly, assume that p'_a streams have changed after the change-point τ_1 and the other $p_a - p'_a$ streams occur changes after τ_2 ($\tau_2 > \tau_1$). In such a situation, T_{new} would perform well for testing p'_a affected streams out of all p data streams within the period $[\tau_1 + 1, \tau_2]$. If a signal is not trigged before τ_2, T_{new} would issue an alarm with a larger probability after τ_2 since signal-to-noise ratio increases. Some simulation results (available from authors upon request) under scenarios when a common event triggers different onset time of changes at different data streams show that T_{new} still works well.

Another noteworthy aspect of T_{new} is that when different CUSUM sequences take different reference values (i.e., μ_i's are not the same), it is able to integrate all the individual

CUSUM statistics in a relatively "fair" way in the sense that $S_i(t)$'s are transformed to identically uniform distribution by $H_t(S_i(t); \mu_i)$ under in-control. In contrast, $T_{\max}$ and T_{sum} do not share this feature. The CUSUM statistics with different reference values would have different expectations and variances under null hypothesis and thus directly applying $T_{\max}$ or T_{sum} to the CUSUM sequences (especially when μ_i's differ considerably) would yield rather unbalanced detection ability in individual CUSUMs. Certainly, this issue could be overcome by employing the transformation $H_t(S_i(t); \mu_i)$ before $T_{\max}$ or T_{sum} is used.

3.2 Extensions and practical implementation

In this section, we discuss extensions of the proposed T_{new} to more general settings, such as, when the post-changes μ_k are unknown, and when data streams may be correlated or non-Gaussian.

When the post-changes cannot be completely specified

A major limitation with the procedure T_{new} as well as other existing multi-streams detection procedures, is that they assume the post-change distributions (or equivalently μ_k's under normality assumption) are completely prescribed, which is often not realistic in practice. When the assumed post-change μ_k deviates from the true one, those procedures suffer from performance degradation. Despite the observation we will see in Section 4 that T_{new} still has better overall performance in the linearly increasing scenario where the post-change distribution is mis-specified, more robust schemes to different change magnitudes would be desired.

This issue seems to be partially resolved by using Han and Tsung's (2006) reference-free-cumulative-score (RFC-Cuscore) method. Their main idea is to replace μ_k by the absolute value of the observations $|X_{kt}|$ at each point t, because $|X_{kt}|$ contains the real information on the magnitude and the pattern of mean change. To be more specific, the RFC-Cuscore chart is defined as

$$R_k(t) = \max\left\{0, R_k(t-1) + |X_{kt}|(X_{kt} - |X_{kt}|/2)\right\}. \tag{8}$$

This control scheme aims at tracing and detecting nonconstant, time-varying mean changes, which is different from ours. However, inspired by its good robustness and sensitivity features as shown by Han and Tsung (2006), it is natural to consider developing T_{new} with $S_k(t)$ replaced by $R_k(t)$ for the case where the post-changes μ_k cannot be specified accurately.

When between-streams correlations exist

Consider a model $\mathbf{X}_t = \boldsymbol{\mu} + \mathbf{Z}_t$, where the mean vector $\boldsymbol{\mu}$ is non-random and possibly sparse, and $\mathbf{Z}_t \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$ for some covariance matrix $\boldsymbol{\Sigma}$ (Zou and Qiu 2009). As revealed by Hall and Jin (2010), for testing in heterogenous mixture, the correlation structure in the noise is not necessarily a curse and could be a blessing. Suppose $\boldsymbol{\Sigma}$ can be specified before monitoring. A straightforward way is to transform the data $\mathbf{X}_t$ by $\mathbf{L}^{-1}\mathbf{X}_t$, where $\mathbf{L}\mathbf{L}^T$ is a Cholesky decomposition of $\boldsymbol{\Sigma}$. Then this becomes our problem with $\mathbf{L}^{-1}\mathbf{X}_t$ instead of $\mathbf{X}_t$ and thus T_{new} is still applicable. Note that after this transformation the sparsity property ($p_a << p$) in the original sequence $\mathbf{X}_t$ may not hold any more. Thus, another alternative is to apply T_{new} directly without any transformation. Of course, the dependence between streams would affect the performances of control schemes. We will use simulation to study the performance of different control schemes with or without transformations when the observations are weakly correlated. We found that the comparison conclusion made for independent cases generally holds for dependent cases and our proposed method is still able to balance the detection ability between the sparse and dense scenarios.

When the data streams are non-Gaussian

For each $k = 1, \ldots, p$, the probability density functions of the observations in the kth data stream before and after the change are f_k and g_k, respectively, where f_k and g_k are given. Then, $S_k(t)$ in (1) is naturally replaced by

$$S'_k(t) = \max\left\{0, S'_k(t-1) + \log\frac{g_k(X_{kt})}{f_k(X_{kt})}\right\}. \tag{9}$$

In this case, how to calculate the value of individual $H'_t(S'_k(t); f, g)$ poses challenges, where $H'_t(\cdot; f, g)$ denotes the C.D.F. of $S'_k(t)$. For exponential family data, appropriate approximations can be developed by mimicking Grigg and Spiegelhalter's (2008) procedure. Alternatively, it has been widely accepted that the ARL of the CUSUM chart can be calculated by the Markov chain method as shown in Brook and Evans (1972). Likewise, the Markov chain method can also be used to approximate the steady-state distribution of the CUSUM statistics under general distribution assumptions. We refer to Li and Tsung (2009) for a recent development. It should be stressed that although the Markov-chain based approximation requires more computational effort than the analytical one such as Grigg and Spiegelhalter's (2008), the total on-line computational task is trivial because it is still of order $O(p \log p)$. The only challenge is to store N different $m \times m$ transition matrices, where m is the number of transition states (usually about 50 to 100) and N is the number of different steady-state distributions needed to evaluate.

When two-sided changes are of interest

In practice, we are often concerned with both positive and negative shifts. As a convention, the following two one-sided CUSUM sequences can be used (e.g., see Chapter 4.2 of Qiu 2014)

$$S_k^U(t) = \max\left\{0, S_k^U(t-1) + \mu_k^U(X_{kt} - \mu_k^U/2)\right\},$$
$$S_k^L(t) = \min\left\{0, S_k^L(t-1) + \mu_k^L(X_{kt} + \mu_k^L/2)\right\},$$

where μ_k^U and μ_k^L are the prescribed change magnitudes for the upper- and lower-sided CUSUMs, respectively. Accordingly, $S_k^{UL}(t) = \max\{S_k^U(t), -S_k^L(t)\}$ can be employed to replace $S_k(t)$ in the definition of T_{new}. The distribution of $S_k^{UL}(t)$ can be numerically obtained by the Markov-chain based approximation (Woodall 1984).

Another easier and more direct way is to apply the EWMA-type control chart, i.e.,

$$Z_k(t) = (1 - \lambda) Z_k(t-1) + \lambda X_{kt},$$

which is naturally a two-sided detection scheme (Lucas and Saccucci 1990). Here λ is a pre-chosen smoothing parameter. Clearly, the marginal distribution of $Z_k(t)$ is easy to be determined in many cases (e.g., in the considered Gaussian case it is just normally distributed). Replacing $S_k(t)$ with $Z_k(t)$ in W_t enables us to monitor two-sided changes in a simple and effective fashion.

When the IC parameters cannot be completely specified

The proposed procedure as well as other existing multi-streams detection procedures, assume the IC distributions (or equivalently μ_{0k}'s and σ_{0k}'s under normality assumption) are completely prescribed. A sufficiently large reference sample (say $\{\mathbf{X}_{-m+1}, \ldots, \mathbf{X}_0\}$) is required to obtain reliable estimates of those parameters to alleviate the parameter estimation effect (Jensen et al. 2006). However, this may be costly in practice; it may not be feasible to wait for the accumulation of sufficiently large calibration samples because users usually want to monitor the process at the start-up stages. Generally, there are two ways to address this issue to certain degree. One method is to adjust the control limit of a chart to make its IC ARL equal to the nominal one given the historical sample size (e.g., see Jones 2002). Another one is the self-starting methods that handle sequential monitoring by simultaneously updating parameter estimates and checking for OC conditions. Specially, we may use the following transformations Q_{kt} to replace X_{kt} (Quesenberry 1991)

$$Q_{kt} = \Phi^{-1}\left(G_{m+t-2}\left\{ \left(\frac{m+t-1}{m+t} \right)^{1/2} \left(\frac{X_{kt} - \bar{X}_{k,t-1}}{\widehat{\sigma}_{k,t-1}} \right) \right\} \right), \tag{10}$$

where $\bar{X}_{kt} = (m+t)^{-1} \sum_{i=-m+1}^{t} X_{ki}$, $\widehat{\sigma}^2_{kt} = (m+t-1)^{-1} \sum_{i=-m+1}^{t} (X_{ki} - \bar{X}_{kt})^2$, and $G_\nu(\cdot)$ denotes the student t distribution function with ν degrees of freedom. It can be easily seen that under IC, Q_{kt}'s are i.i.d. standard normal variables as if the IC parameters were known. We will see in Section 4 that these two methods have comparable performances and our proposed method still performs reasonably well compared to the existing works.

On the use of nonparametric statistics

In a high-dimensional environment, the distributions of some data streams are likely to be nonnormal, so that the statistical properties of commonly used charts, which were designed to perform best under the normal distribution, could potentially be affected. Nonparametric or robust charts may be useful in such situations (c.f., Chapters 8 and 9 of Qiu 2014). For example, we may transform the observation X_{kt} to its nonparametric counterpart

$$R_{kt}^{*} = \frac{R_{kt} - (m+t+1)/2}{\sqrt{(m+t+1)(m+t-1)/12}}, \quad t = 1, 2, \ldots,$$

where $R_{kt} = \sum_{i=-m+1}^{t} I(X_{ki} \leq X_{kt})$. Note that $R_{kt}^{*}, R_{k,t+1}^{*}, \ldots$ are independent and discretely uniform variables. Again, the Markov chain method can be used to approximate the distribution of CUSUM or EWMA charting statistics with R_{kt}^{*}.

4 Simulation Study

We present some simulation results in this section regarding the performance of the proposed T_{new} and compare it with other procedures in the literature. All results in this section are obtained from 10,000 replications. The Fortran codes for implementing the proposed procedure are available in the supplemental material. Because a similar conclusion holds for other cases, throughout this section, we only present the results when IC ARL (ARL_0) is $1,000$ or $10,000$ for illustration purposes. Since the zero-state ($\tau = 0$) and steady-state ARL (SSARL) comparison results are similar, only the SSARLs are provided. To evaluate the SSARL behavior of each chart, any series in which a signal occurs before the $(\tau+1)$-th observation is discarded (c.f., Hawkins and Olwell 1998). Here we consider $\tau = 25$ for illustration. For comparison, besides $T_{\max}$ and T_{sum}, the method based on the higher criticism test HC_n^{*}, denoted as T_{hc}, is included as well. The construction of T_{hc} is similar to T_{new} by replacing the D_{GOF} statistic with HC_n^{*} at each time point t.

4.1 I.I.D. cases with known IC parameters

For the in-control state, we drew samples from $N(0, 1)$; for the out-of-control state, at each time point, we first drew p_a samples from $N(\delta_i, 1), i = 1, \ldots, n_a$ and then drew $p - p_a$ samples from $N(0, 1)$. As an illustration, two scenarios are considered: (I) equal allocation, i.e., $\delta_i = \mu$ where μ is the target value specified before monitoring; (II) linearly increasing allocation $\delta_i = i\delta$ for $i = 1, \ldots, p_a$. Note that the linearly increasing scenario corresponds to the situation that the post-change distributions are mis-specified. To make the ARL performance comparable among the configurations of various p_a, we set $\sum_{i=1}^{p_a} \delta_i^2 = p_a\mu^2$ for each value of p_a.

The simulation results with $p = 100$ and $\mu = 0.5$ are summarized in Table 1. Besides the SSARLs, the corresponding standard deviations of the run lengths (SDRL) are also included in this table to give a broader picture of the run-length distribution. Table 1 shows that in both scenarios, our proposed scheme performs slightly worse than T_{max} when p_a is very small (less than 3) but outperforms T_{sum} by quite a large margin in these situations. For moderately sparse case $5 \leq p_a \leq 10$, T_{new} is almost uniformly superior to the other three schemes. For $p_a \geq 20$, T_{new} and T_{sum} have comparable performance and their advantage over T_{max} and T_{hc} are remarkable. This is consistent with our intuition from the results given in Section 2. As a side note, T_{hc} seems more preferable to T_{max} since they have similar performance when p_a is small but T_{hc} works better when p_a is large.

Next a higher dimension, $p = 1000$, is considered. In this case, we choose a smaller μ, i.e., $\mu = 0.2$. The simulation results are summarized in Figure 3, which shows the SSARL curves (in the log scale) of all the four schemes in the top two and bottom two panels for equal and increasing scenarios, respectively, when $\mathrm{ARL}_0 = 1000$ (the left panels) or $10,000$ (the right panels). These figures show similar evidence to that of Table 1. The advantage of T_{new} is clear: it is either the best or close to the best in all the cases. These findings confirm our earlier statement that T_{sum} is effective no matter what the sparsity is in the data streams and it offers a balanced protection against various OC conditions.

Table 1: ARL comparison between four schemes when $p = 100$ and $\mu = 0.5$. Standard deviations of run lengths are in parentheses. The standard error of ARL can be approximated by SDRL/100.

ARL_0	p_a	Scenario (I)				Scenario (II)			
		T_{new}	$T_{\max}$	T_{sum}	T_{hc}	T_{new}	$T_{\max}$	T_{sum}	T_{hc}
	1	71.4 (31.4)	62.2 (28.8)	122 (55.8)	62.9 (29.2)	71.6 (31.5)	62.5 (29.5)	121 (55.8)	62.7 (29.1)
	3	40.5 (13.5)	40.7 (14.2)	55.5 (21.3)	40.0 (13.5)	37.5 (12.1)	34.9 (11.9)	57.9 (22.4)	34.8 (11.6)
	5	30.5 (9.53)	34.9 (10.8)	37.3 (13.3)	33.4 (9.84)	28.7 (8.35)	29.2 (8.83)	39.3 (13.9)	28.5 (8.34)
	8	23.1 (6.75)	30.8 (8.82)	25.2 (8.73)	28.5 (7.47)	22.2 (6.03)	25.3 (6.85)	26.8 (8.97)	24.3 (6.30)
1000	10	19.9 (5.74)	29.0 (7.93)	20.8 (7.04)	26.2 (6.39)	19.6 (5.23)	23.9 (6.18)	22.4 (7.28)	22.5 (5.41)
	20	11.6 (3.40)	24.6 (6.14)	11.2 (3.42)	20.2 (4.12)	12.3 (3.28)	20.2 (4.75)	12.3 (3.74)	17.8 (3.59)
	50	4.65 (1.40)	20.2 (4.79)	4.71 (1.31)	12.4 (2.02)	5.40 (1.56)	16.7 (3.58)	5.23 (1.44)	12.3 (1.96)
	80	2.74 (0.78)	18.4 (4.31)	3.04 (0.79)	8.49 (1.29)	3.22 (0.91)	15.3 (3.26)	3.39 (0.89)	9.07 (1.34)
	100	2.16 (0.59)	17.7 (4.11)	2.51 (0.65)	6.93 (1.03)	2.53 (0.71)	14.7 (3.09)	2.79 (0.71)	7.55 (1.10)
	1	89.0 (35.7)	82.4 (34.5)	189 (68.2)	82.2 (34.3)	89.1 (36.2)	82.8 (34.0)	189 (69.2)	82.9 (34.7)
	3	50.6 (15.2)	56.4 (17.5)	78.2 (24.9)	55.2 (16.7)	45.9 (13.3)	46.2 (14.1)	82.1 (26.2)	46.1 (13.8)
	5	38.3 (10.1)	48.9 (13.0)	50.5 (15.5)	47.2 (12.3)	35.1 (8.67)	38.9 (10.2)	53.7 (16.1)	38.0 (9.84)
	8	29.2 (7.06)	43.7 (10.8)	33.5 (9.70)	41.5 (9.55)	27.4 (6.17)	34.1 (8.17)	36.0 (10.1)	33.3 (7.57)
10000	10	25.3 (5.97)	41.7 (9.84)	27.5 (7.75)	39.0 (8.45)	24.2 (5.28)	32.2 (7.41)	29.6 (8.20)	31.1 (6.71)
	20	15.3 (3.53)	36.2 (7.63)	14.4 (3.83)	32.2 (5.70)	15.6 (3.30)	27.8 (5.58)	15.7 (3.98)	25.9 (4.60)
	50	6.28 (1.53)	30.6 (5.81)	5.92 (1.41)	24.4 (3.11)	7.18 (1.65)	23.8 (4.19)	6.60 (1.55)	20.5 (2.75)
	80	3.66 (0.87)	28.4 (5.23)	3.78 (0.85)	20.2 (2.12)	4.29 (1.02)	22.1 (3.74)	4.24 (0.95)	17.8 (1.99)
	100	2.80 (0.66)	27.2 (4.97)	3.09 (0.68)	18.1 (1.68)	3.33 (0.79)	21.3 (3.55)	3.43 (0.76)	16.5 (1.70)

We also conduct a simulation study in a low-dimension case, $p = 30$, to check whether T_{new} is still effective. The SSARL curves (in the log scale) of all the four schemes with $\mu = 0.5$ are plotted in Figure 4. It is a little surprising to see that T_{new} works reasonably well in most cases when $p = 30$. We conducted some other simulations with different values of p, ARL_0, and μ to check whether the above conclusions would change in other cases. These simulation results, not reported here but available from the authors, show that the proposed scheme works well for other cases as well in terms of its OC ARL. Its relatively superior performance comparing to T_{sum} and $T_{\max}$ still holds for other settings.

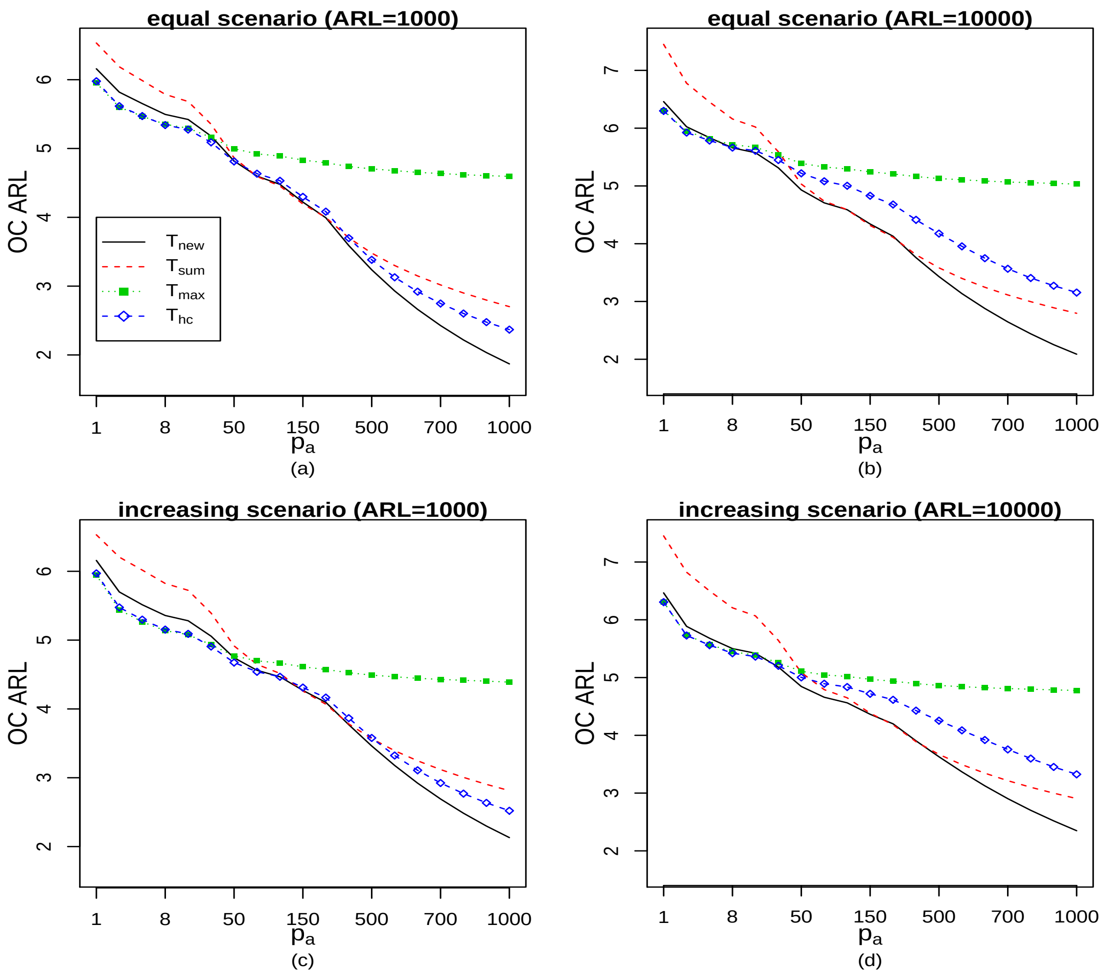

Figure 3: SSARL curves (in the log scale) of T_{new}, T_{sum}, $T_{\max}$ and T_{hc} for $p = 1000$ and $\mu = 0.2$ when $\text{ARL}_0 = 1,000$ (the left two panels) or $10,000$ (the right two panels). The legend in the first plot is applicable for all the others.

4.2 Cases when between-streams correlations exist

In this subsection, we study the performance of T_{new} when data streams are correlated. To this end, we generate multivariate normal distributions with the covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ is chosen to be $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.5$, for $i,j = 1,\ldots,p$. The marginal distributions of streams and the other settings are identical to those in Table 1. We considered the schemes T_{new}, T_{sum} and $T_{\max}$ with or without Cholesky transformation. Since the performances depend on the shift positions, two shift cases are considered: (i) The

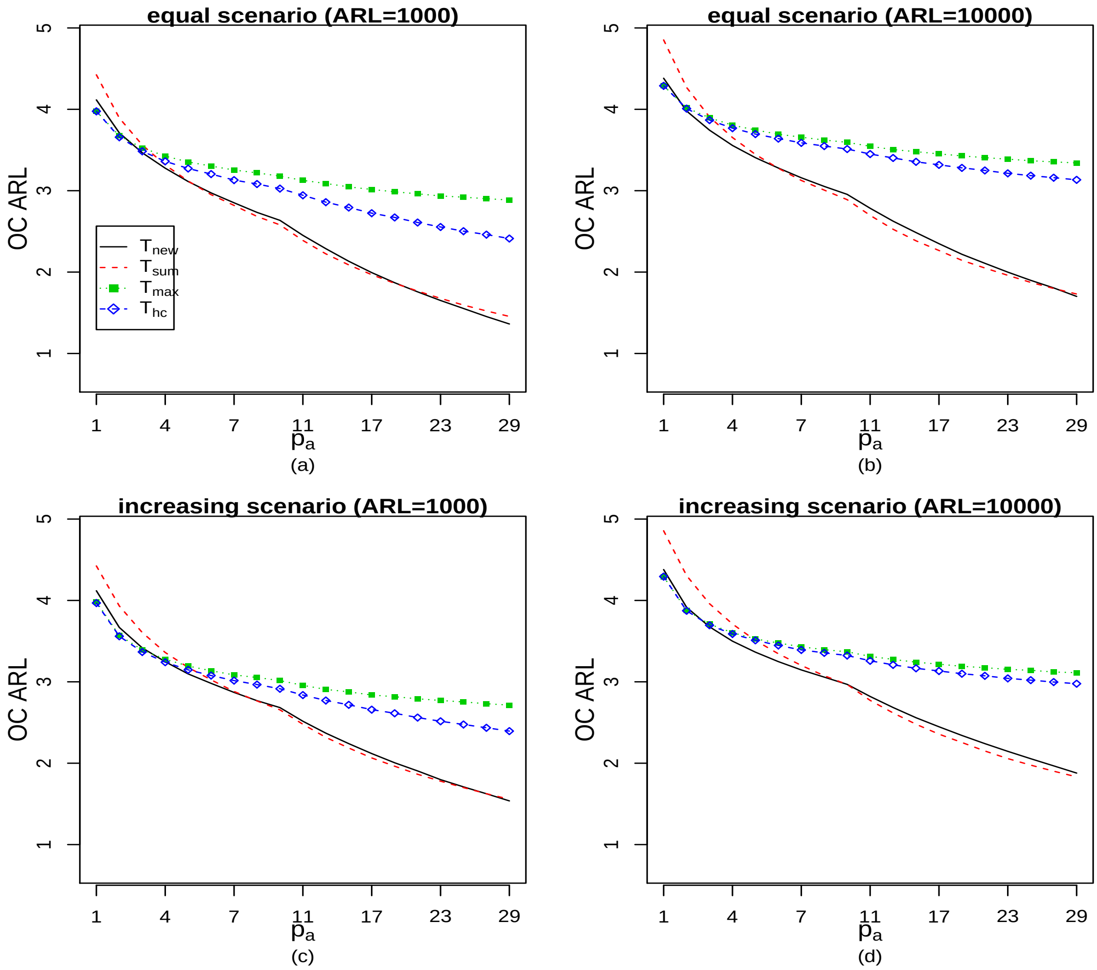

Figure 4: SSARL curves (in the log scale) of T_{new}, T_{sum}, $T_{\max}$ and T_{hc} for $p = 30$ and $\mu = 0.5$ when $\text{ARL}_0 = 1,000$ (the left two panels) or $10,000$ (the right two panels). The legend in the first plot is applicable for all the others.

first p_a streams occur changes; (ii) the shifted p_a stream indices are randomly drawn from $\{1, \ldots, p\}$ without replacement. Table 2 presents ARL comparison of of T_{new}, T_{sum} and $T_{\max}$ when $\text{ARL}_0 = 1,000$. In this table, T^C_{new}, T^C_{sum}, and $T^C_{\max}$ denote the schemes with Cholesky transformation.

The results for both Scenarios (I) and (II) are similar: under the case (i), the charts applied to original observations perform better than those with transformations. T_{new} provides a compromise between $T_{\max}$ and T_{sum}; Although it is not always the best of the three charts we note that it is always either the best or come close to the best, including in cases where

Table 2: ARL comparison of of T_{new}, T_{sum} and $T_{\max}$ when the streams are correlated when $p = 100$, $\mu = 0.5$ and $\text{ARL}_0 = 1,000$. T^C_{new}, T^C_{sum}, and $T^C_{\max}$ denote the schemes with Cholesky transformation

Shift case	p_a	Scenario (I)						Scenario (II)					
		$T_{\max}$	T_{sum}	T_{new}	$T^C_{\max}$	T^C_{sum}	T^C_{new}	$T_{\max}$	T_{sum}	T_{new}	$T^C_{\max}$	T^C_{sum}	T^C_{new}
	1	62.7	163	88.5	61.6	124	71.8	61.9	159	88.4	62.0	125	72.4
	3	45.5	78.6	53.2	57.1	87.6	59.8	36.6	80.6	48.2	49.4	74.8	52.6
	5	38.8	54.0	41.4	53.0	67.0	51.1	30.6	55.6	37.4	46.6	59.0	45.4
	8	33.8	37.1	31.0	49.8	47.5	42.3	27.4	39.4	28.9	43.0	45.3	38.0
(i)	10	32.0	30.2	27.2	47.4	40.1	37.1	25.3	32.9	25.9	41.2	38.5	34.4
	20	26.3	16.5	16.5	41.8	22.3	22.7	21.3	18.1	16.7	35.9	22.8	23.0
	50	21.3	7.01	7.07	33.6	8.99	9.30	17.5	7.80	8.00	28.8	9.95	10.3
	80	18.9	4.50	4.16	30.5	5.63	5.35	16.0	4.98	4.90	26.3	6.27	6.20
	100	18.1	3.61	3.18	29.3	4.44	4.02	15.2	4.07	3.77	25.0	4.94	4.75
	1	62.7	162	88.4	48.7	104	57.4	62.1	163	90.0	48.7	105	58.1
	3	41.1	76.8	51.8	34.1	49.8	34.2	35.2	81.4	46.9	28.8	52.5	31.6
	5	34.8	52.4	40.1	29.8	34.6	26.5	29.0	56.3	36.7	24.4	36.6	24.5
	8	31.1	36.7	30.6	26.5	24.3	20.6	25.3	38.3	28.7	21.5	25.8	19.5
(ii)	10	29.4	29.9	26.5	25.2	20.9	18.4	23.9	32.3	25.3	20.6	21.9	17.6
	20	24.8	16.4	16.4	22.1	12.5	12.5	20.1	17.7	16.6	17.8	13.1	12.2
	50	20.8	7.02	7.10	20.0	6.64	7.05	16.9	7.69	7.92	15.8	6.82	7.21
	80	19.1	4.44	4.21	21.2	4.98	4.95	15.4	4.92	4.91	15.6	5.07	5.25
	100	18.0	3.63	3.19	28.8	4.41	4.05	14.9	4.09	3.76	16.2	4.39	4.47

there is a significant difference in performance between the best and the worse. By comparing the results with those in Table 1 with no correlations between streams, the positive dependence has adversely effects on detection abilities of all the three charts. Under the case (ii), it appears that the control charts with transformations are more efficient than the charts without transformations when p_a is not too large. This can be understood that the sparsity properties generally remain to hold after the transformation in this case. From this table, we can confirm our earlier statement that T_{new} is still effective when the correlations exist because it offers protection against various OC conditions compared to the competitors. We can also conclude that the performances of the control charts with and without transformations would depend on the covariance structure and shift directions. In this paper, we make no attempt to further analyze this problem, but rather think that this issue certainly warrants future research.

4.3 Cases when the number of historical samples is not large

In all the foregoing examples, it is assumed that the IC parameters are known or, equivalently, that they are estimated from a sufficiently large reference dataset. Finally, we study the performance of T_{new} when this assumption is violated. Only the case $\text{ARL}_0 = 1,000$ is considered and all the other settings are the same as those in Table 1. Table 3 shows the IC ARLs and SDRLs of T_{new}, T_{sum}, and $T_{\max}$ when the IC parameters μ_{0k}'s and σ_{0k}'s are computed from an IC dataset with various historical sample sizes, m. In each replication, a sample of size m is firstly generated and the IC parameters are estimated from this sample. Then, an independent sequence of multivariate observations is generated and all the three charts are used to obtain the corresponding run lengths. From this table, it can be seen that (i) when the sample size of the IC dataset is relatively small, the actual IC ARLs and SDRLs of the three charts are both quite far away from the nominal level of 1,000, (ii) when the sample size of the IC dataset increases, such biases decrease, and (iii) the biases in ARL_0 of the three charts are similar. For $p = 100$, it seems that at least more than 4,000 historical observations are required to make the IC performances of all the detection schemes close to the nominal value. Therefore, we next study the performances of T_{new} if one applies the two modifications suggested in Section 3.2 for small m.

For simplicity, only Scenario (I) is considered. Table 4 shows OC ARL comparison of of T_{new}, T_{sum} and $T_{\max}$ with various values of m when $p = 100$, $\mu = 0.5$ and $\text{ARL}_0 = 1,000$. Both the two modifications for small m, i.e., using adjusted control limits and self-starting statistics (10), are considered for each detection scheme. We observe that T_{new} still has the ability to offer a good balance protection against p_a regardless of whatever m is and which

Table 3: IC ARL and SDRL values of T_{new}, T_{sum} and $T_{\max}$ for $p = 100$ and $\mu = 0.5$ with various Phase I sample sizes m and nominal $\text{ARL}_0 = 1,000$. Numbers in parentheses are SDRLs.

m	$T_{\max}$	T_{sum}	T_{new}
250	268 (225)	279 (287)	235 (208)
500	455 (415)	494 (521)	448 (439)
1000	643 (595)	707 (724)	664 (648)
2000	784 (742)	850 (843)	825 (804)
4000	869 (807)	937 (924)	935 (918)
8000	912 (862)	995 (969)	980 (963)

modification is used. When m is small, such as 250, the charts using (10) perform better for the cases that the overall signal is strong (say p_a is large), while the charts with adjusted control limits have a certain advantage for the weak signals. When m is as large as 1,000, the performances of two modifications are similar. Considering its convenience and robustness in various circumstances, our empirical results indicate that T_{new} should be a reasonable alternative for monitoring high-dimensional data streams.

5 An Example of Quality Control in Semiconductor Manufacturing

In this section, we demonstrate the proposed methodology by applying it to a real dateset from a semiconductor manufacturing process which is under consistent surveillance via the monitoring of signals/variables collected from sensors at many measurement points. For each observation, there are originally $p = 591$ continuous measurements (from sensors). The goal of the data analysis is mainly to model and monitor production quality based on those sensor measurements.

The data set contains a total of $n = 1,567$ vector observations, and is publicly available in the UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/SECOM). The data were collected from July 2008 to October 2008 by a computerized system which

Table 4: ARL comparison of of T_{new}, T_{sum} and $T_{\max}$ with various values of Phase I sample size m for Scenario (I) when $p = 100$, $\mu = 0.5$ and $\text{ARL}_0 = 1,000$.

m	p_a	Adjusted control limits			Self-starting schemes using (10)		
		$T_{\max}$	T_{sum}	T_{new}	$T_{\max}$	T_{sum}	T_{new}
250	1	84.8 (45.9)	136 (70.7)	90.7 (46.8)	117 (234)	254 (411)	141 (267)
	3	54.4 (18.9)	62.7 (25.1)	49.2 (16.8)	48.1 (23.4)	66.9 (35.1)	47.8 (21.4)
	5	46.5 (13.9)	42.1 (15.2)	37.1 (10.9)	39.8 (14.6)	42.1 (17.8)	33.8 (12.3)
	8	41.0 (11.0)	29.1 (9.56)	28.2 (7.46)	33.8 (10.9)	27.3 (10.4)	24.8 (8.01)
	10	38.9 (10.0)	24.0 (7.66)	24.5 (6.26)	31.7 (9.66)	22.2 (7.95)	21.2 (6.60)
	20	33.2 (7.74)	13.0 (3.79)	15.1 (3.59)	26.5 (7.17)	11.5 (3.74)	12.1 (3.62)
	50	27.6 (5.92)	5.46 (1.43)	6.48 (1.57)	21.3 (5.34)	4.76 (1.34)	4.70 (1.43)
	80	25.4 (5.27)	3.51 (0.85)	3.84 (0.92)	19.3 (4.70)	3.07 (0.81)	2.76 (0.80)
	100	24.3 (5.09)	2.87 (0.69)	2.98 (0.70)	18.5 (4.49)	2.52 (0.66)	2.17 (0.60)
500	1	72.7 (38.6)	126 (62.9)	78.6 (37.8)	75.0 (63.8)	155 (148)	89.1 (74.6)
	3	46.9 (16.1)	58.1 (23.1)	43.6 (14.8)	44.5 (17.6)	60.9 (26.3)	43.9 (16.8)
	5	40.3 (12.1)	39.2 (14.3)	33.1 (10.1)	37.2 (12.5)	39.1 (15.3)	32.4 (10.7)
	8	35.6 (9.79)	26.5 (9.09)	25.0 (6.91)	32.4 (9.98)	26.1 (9.31)	23.8 (7.38)
	10	33.7 (9.02)	22.2 (7.23)	21.6 (5.89)	30.4 (8.88)	21.5 (7.44)	20.5 (6.11)
	20	28.6 (6.99)	11.9 (3.58)	13.0 (3.49)	25.7 (6.77)	11.4 (3.59)	11.8 (3.55)
	50	23.7 (5.32)	5.00 (1.35)	5.38 (1.48)	20.9 (5.09)	4.75 (1.31)	4.70 (1.44)
	80	21.6 (4.75)	3.23 (0.83)	3.15 (0.84)	18.9 (4.53)	3.05 (0.81)	2.77 (0.80)
	100	20.8 (4.53)	2.66 (0.67)	2.46 (0.64)	18.1 (4.34)	2.52 (0.65)	2.17 (0.59)
1,000	1	67.3 (33.2)	124 (59.3)	74.4 (34.7)	67.4 (36.7)	134 (71.4)	78.3 (40.2)
	3	43.7 (15.1)	56.5 (22.3)	42.0 (14.4)	42.5 (15.8)	57.8 (23.7)	42.2 (15.0)
	5	37.5 (11.3)	38.0 (13.7)	31.6 (9.63)	36.2 (11.7)	38.2 (14.2)	31.5 (10.1)
	8	33.2 (9.33)	25.9 (8.93)	23.9 (6.76)	31.6 (9.30)	25.7 (9.00)	23.4 (7.08)
	10	31.3 (8.44)	21.4 (7.11)	20.6 (5.86)	29.8 (8.57)	21.2 (7.20)	20.1 (5.96)
	20	26.6 (6.59)	11.5 (3.54)	12.2 (3.43)	25.3 (6.41)	11.2 (3.50)	11.8 (3.45)
	50	22.0 (5.03)	4.84 (1.34)	4.96 (1.44)	20.5 (4.95)	4.71 (1.31)	4.66 (1.40)
	80	20.0 (4.54)	3.14 (0.82)	2.94 (0.82)	18.6 (4.42)	3.07 (0.81)	2.75 (0.80)
	100	19.2 (4.31)	2.58 (0.66)	2.30 (0.61)	17.9 (4.23)	2.52 (0.65)	2.16 (0.60)

automatically manages the process. Among them, 104 observations are classified as inferior (non-conforming) based on physical testing and experience of the engineers and the remaining observations (1,463) are conforming. When the process incurs a change, we are not sure if all sensor variables are relevant to the change. It is often the case that the measured signals contain a combination of useful information, irrelevant information as well as noise, and is often the case that useful information is buried in the latter two. It is interesting to consider some scalable monitoring methods which are computationally efficient without the knowledge about the number of affected sensors. Thus, we apply the proposed T_{new} to this dataset.

Similar to any real life situation, this dataset contains null values varying in intensity depending on the individuals features. Since the fraction of missing values is trivial in this dataset, we simply use mean imputation (to replace each missing value with the mean of the observed values for that variable). In addition, 117 constant features (streams) are removed from the analysis and totally $p = 474$ data streams are monitored simultaneously.

For illustration, we use all the observations belonging to the conforming group (totally 1463 observations) as the historical sample and the others for testing. First of all, we conduct Shapiro-Wilk goodness-of-fit tests for normality and conclude that at least 164 streams are not normally distributed (the p-values are smaller than 0.01). In order to make our assumption approximately valid, we perform an inverse transformation to the 104 "on-line" observations, say $\Phi^{-1}(\hat{F}_{nk}(X_{kt})), t = 1, \ldots, 104$, where $\hat{F}_{nk}$ is the empirical distribution function based on the 1463 historical observations of the kth streams. It should be noticed that this transformation does not imply the joint normality. When the joint normality assumption is invalid, the nominal IC ARL may not be achieved. Some nonparametric procedures (e.g., Qiu and Hawkins 2001; 2003) would be more appealing when the joint distribution is far away from normal.

Furthermore, a calibration sample of size 1463 may be smaller than ideal to determine fully the in-control parameters by considering the dimension is as high as 474. Hence, to alleviate the estimated parameters problem, we also employ the transformation in (10) for each sensor observation. In addition, we calculated the sample correlation matrix based on the historical sample and found there were totally 4710 off-diagonal entries (among all the 224,202 entries) whose absolute values are larger than 0.3. This number is very similar to that in an autoregressive correlation matrix $\mathbf{\Sigma} = (\rho^{|i-j|})_{p\times p}$ with $\rho = 0.8$. Thus, we may conclude that the between-streams correlations exist in this example but certain weak dependence structure may hold. Accordingly the detection schemes discussed in Section 4 may still be effective.

We artificially assume that we monitor the observations categorized as the nonconforming level sequentially. We suppose the positive location shift is of greatest interest and construct T_{new}, $T_{\max}$ and T_{sum} to monitor production quality deterioration. We choose $\mu_k = 0.5$ for all the data streams and in-control average run length as 2,000. Accordingly, the control limits for T_{new}, $T_{\max}$ and T_{sum} are 36.58, 22.19 and 415.16, respectively.

Figure 5 shows the resulting chart T_{new} (solid curve connecting the dots) along with a solid horizontal line. The charting statistics are divided by the control limits so that we can plot all the three charts in one figure. The corresponding T_{sum} (dashed curve connecting diamonds) and $T_{\max}$ (dotted curve connecting circles) are also presented in the figure. We considered all 104 nonconforming observations, but only plot the first 20 of them in Figure 5 as all the methods signal by that point. From the plot, it can be seen that the T_{new} chart exceeds its control limit around the 14th observation and it remains above the control limit. This excursion suggests that a marked change has occurred which concurs with our setting that the nonconforming observations are sequentially monitored. In comparison, the $T_{\max}$ chart does not give a signal until the 16th observation and the T_{sum} statistics remain below the control limit until the 19th observation. This result may reflect that only a small number of data streams are affected since T_{sum} is outperformed by $T_{\max}$.

In monitoring complex systems, apart from quick detection of abnormal changes of system performance and key parameters, accurate fault diagnosis of responsible factors has become increasingly critical in a variety of applications. This will increase process throughput, decrease time to learning, and reduce per unit production costs. Some post-signal fault isolation schemes can be used, e.g., see Zou et al. (2011) and the references therein.

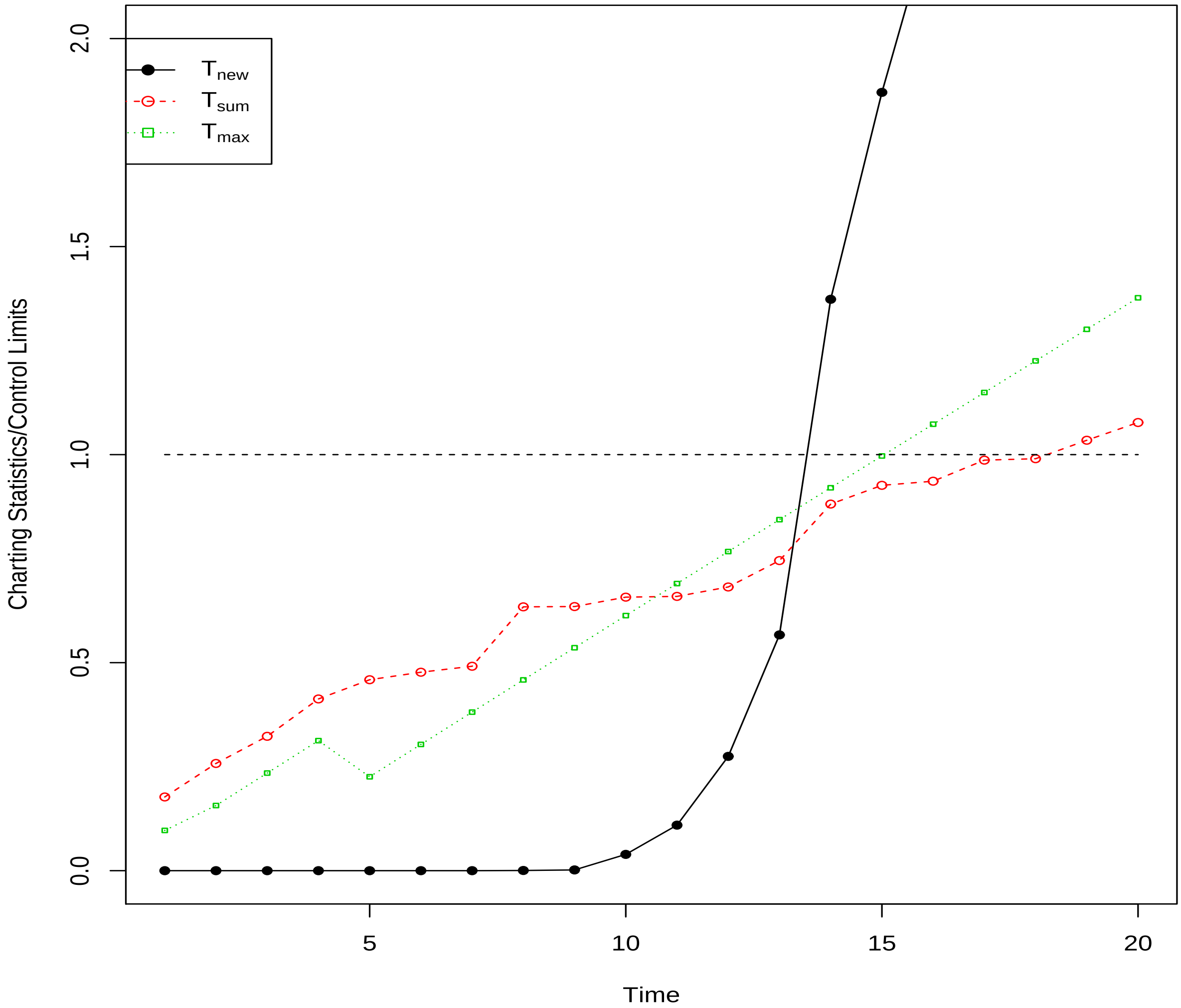

Figure 5: The T_{new}, $T_{\max}$ and T_{sum} control charts for monitoring the semiconductor manufacturing process. The values on the Y-axis are the ratios of the charting statistics to the control limits.

6 Concluding Remarks

In this paper, we propose a new detection scheme, T_{new}, for monitoring high-dimensional data streams. This procedure is derived based on a powerful GOF test and naturally integrates the information from multiple data streams with the CUSUM scheme. With updating formulations, the proposed scheme is fast to compute with a similar computational effort to existing schemes. Compared with existing methods, it is more robust in the sense that it is able to balance the detection of various fractions of affected streams. In many cases, the improvement is quite remarkable.

This paper focuses on Phase II monitoring only and assumes that all historical observations used for estimating the IC parameters are i.i.d. following a given distribution. However, there is no such assurance in many applications. Hence, it requires more research to extend our method to Phase I analysis, in which detection of outliers or change-points in a historical dataset and estimation of the baseline parameters would be of great interest. In the last several years, univariate and multivariate nonparametric control charts have attracted much attention from researchers. The need for robust multivariate SPC has been noted in a number of articles and some effort has been devoted to this problem, see, e.g., Qiu and Hawkins (2001; 2003), Qiu (2008), Zou and Tsung (2011), Qiu and Li (2011), Woodall and Montgomery (2014) and references therein. Extension of the proposed method to nonparametric settings also warrants future study. In this work, we assume that the observations $\{X_{k1}, \ldots, X_{kt}\}$ are independent. In practice, they are often correlated. It is necessary to investigate the performance of T_{new} under some time series models (Qiu and Xiang 2014). Moreover, Mei (2010) and Tartakovsky et al. (2006) established some asymptotical property of T_{sum} and $T_{\max}$, respectively. It is of interest to make some asymptotic analysis of the proposed method and to asymptotically compare the three methods.

Supplementary material It contains the technical details and codes for implementing the proposed method.

Acknowledgement

The authors thank the editor, associate editor, and three anonymous referees for their many helpful comments that have resulted in significant improvements in the article. The authors would like to thank Prof. William H. Woodall for valuable comments and suggestions. This research was supported by the NNSF of China Grants 11131002, 11101306, 11371202, 11271205, 71202087, the RFDP of China Grant 20110031110002, the Foundation for the Author of National Excellent Doctoral Dissertation of PR China 201232, New Century Excellent Talents in University and PAPD of Jiangsu Higher Education Institutions.

References:

Anderson, T. W., and Darling, D. A. (1952), "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes," *Annals of Mathematical Statistics*, 23, 193–212.

Benjamini, Y., and Kling, Y. (1999), "A Look at Statistical Process Control Through the p-Values," Technical report, RP-SOR-99-08, Tel Aviv University, Israel.

Brook, D., and Evans, D. A. (1972), "An Approach to the Probability Distribution of CUSUM Run Length," *Biometrika*, 59, 539–549.

Cai, T., Jeng, J., and Jin, J. (2011), "Optimal Detection of Heterogeneous and Heteroscedastic Mixtures", *Journal of the Royal Statistical Society, Series B*, 73,

Capizzi, G., and Masarotto, G. (2011), "A Least Angle Regression Control Chart for Multidimensional Data," *Technometrics*, 53, 285–296.

Croisier, R. B. (1988), "Multivariate Generalizations of Cumulative Sum Quality-Control Schemes," *Technometrics*, 30, 243–251.

Donoho, D., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *Annals of Statistics*, 32, 962–994.

Guerriero, M., Willett, P., and Glaz, J. (2009), "Distributed Target Detection in Sensor Networks Using Scan Statistics," *IEEE Transactions on Signal Processing*, 57, 2629–2639.

Grigg, O. A., and Spiegelhalter, D. J. (2008), "An Empirical Approximation to the Null Unbounded Steady-State Distribution of the Cumulative Sum Statistic," *Technometrics*, 50, 501–511.

Hall, P., and Jin, J. (2010), "Innovated Higher Criticism for Detecting Sparse Signals in Correlated Noise," *Annals of Statistics*, 38, 1686–1732.

Han, D., and Tsung, F. (2006), "A reference-free cuscore chart for dynamic mean change detection and a unified framework for charting performance comparison," *Journal of the American Statistical Association*, 101, 368–386.

Hawkins, D. M., (1991), "Multivariate Quality Control Based on Regression-Adjusted Variables," *Technometrics*, 33, 61–75.

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

Jensen, W. A., Jones, L. A., Champ, C. W., and Woodall, W. H. (2006), "Effects of Parameter Estimation on Control Chart Properties: A Literature Review," *Journal of Quality Technology*, 38, 349–364.

Jiang, W. Au, S.T. and Tsui, K.-L. (2007), "A Statistical Process Control Approach for Customer Activity Monitoring," *IIE Transactions*, 39, 235–249

Jones, L. A. (2002), "The Statistical Design of EWMA Control Charts with Estimated Parameters," *Journal of Quality Technology*, 34, 277–288.

Li, Y., and Tsung, F. (2009), "False Discovery Rate-Adjusted Charting Schemes for Multistage Process Monitoring and Fault Identification," *Technometrics*, 51, 186–205.

Lucas, J. M., and Saccucci, M. S. (1990), "Exponentially Weighted Moving Average Control Scheme Properties and Enhancements," *Technometrics*, 32, 1–29.

Mei, Y. (2010), "Efficient Scalable Schemes for Monitoring a Large Number of Data Streams," *Biometrika*, 97, 419–433.

Moustakides, G. V. (1986), "Optimal Stopping Times for Detecting Changes in Distributions," *Annals of Statistics*, 14, 1379-1387.

Qiu, P. (2008), "Distribution-Free Multivariate Process Control Based on Log-Linear Modeling," *IIE Transactions*, 40, 664–677.

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

Qiu, P., and Hawkins, D. M. (2001), "A Rank-Based Multivariate CUSUM Procedure," *Technometrics*, 43, 120–132.

Qiu, P., and Hawkins, D. M. (2003), "A Nonparametric Multivariate CUSUM Procedure for Detecting Shifts in All Directions," *Journal of the Royal Statistical Society, Ser. D*, 52, 151–164.

Qiu, P., and Li, Z. (2011), "On Nonparametric Statistical Process Control of Univariate Processes," *Technometrics*, 53, 390–405.

Qiu, P., and Xiang, D. (2014), "Univariate Dynamic Screening System: An Approach For Identifying Individuals With Irregular Longitudinal Behavior," *Technometrics*, 56, 248–260.

Quesenberry, C. P. (1991), "SPC Q Charts for Start-Up Processes and Short or Long Runs", *Journal of Quality Technology*, 23, 213–224.

Tartakovsky, A. G., Rozovskii, B. L., Blazek, R. B., and Kim, H. (2006), "Detection of Intrusions in Information Systems by Sequential Change-Point Methods (with Discussion)," *Statistical Methodology*, 3, 252–340.

Tartakovsky, A. G., and Veeravalli, V. V. (2008), Asymptotically Optimal Quickest Change Detection in Distributed Sensor," *Sequential Analysis*, 27, 441–475.

Tsung, F., Zhou, Z.H., and Jiang, W. (2007), "Applying Manufacturing Batch Techniques to Fraud Detection with Incomplete Customer Information," *IIE Transactions*, 39, 671–680.

Veeravalli, V. V. (2001), "Decentralized Quickest Change Detection," *IEEE Transaction Information Theory*, 47, 1657–1665.

Wang, K., and Jiang, W. (2009) "High-Dimensional Process Monitoring and Fault Isolation via Variable Selection," *Journal of Quality Technology*, 41, 247–258.

Woodall, W. H. (1984), "On the Markov Chain Approach to the Two-Sided CUSUM Procedure," *Technometrics*, 26, 41–46.

Woodall, W. H., and Montgomery, D. C. (2014), "Some Current Directions in the Theory and Application of Statistical Process Monitoring," *Journal of Quality Technology*, 46, 79–94.

Zhang, J. (2002), "Powerful Goodness-of-Fit Tests Based on Likelihood Ratio". *Journal of the Royal Statistical Society, Series B*, 64, 281–294.

Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. (2010), "Detecting Simultaneous Change-Points in Multiple Sequences," *Biometrika*, 97, 631–645.

Zou, C., Jiang, W., and Tsung, F. (2011) "A LASSO-Based SPC Diagnostic Framework for Multivariate Statistical Process Control," *Technometrics*, 53, 297–309.

Zou, C., and Qiu, P. (2009), "Multivariate Statistical Process Control Using LASSO," *Journal of the American Statistical Association*, 104, 1586–1596.

Zou, C., and Tsung, F. (2011), "Multivariate Sign EWMA Control Chart," *Technometrics*, 53, 84–97.