A Distribution-free Multivariate Change-Point Model for Statistical Process Control

Maoyuan Zhou [a,c], Xuemin Zi [b], Wei Geng [c], Zhonghua Li [c,*]

[a] *College of Science, Civil Aviation University of China, Tianjin 300300, P.R.China*

[b] *School of Science, Tianjin University of Technology and Education, Tianjin 300222, P.R.China*

[c] *LPMC and School of Mathematical Sciences, Nankai University, Tianjin 300071, P.R.China*

Abstract

This paper develops a new distribution-free multivariate procedure for statistical process control based on minimal spanning tree (MST), which integrates a multivariate two-sample goodness-of-fit (GOF) test based on MST and change-point model. Simulation results show that our proposed procedure is quite robust to non-normally distributed data, and moreover, it is efficient in detecting process shifts, especially moderate to large shifts, which is one of the main drawbacks of most distribution-free procedures in the literature. The proposed procedure is particularly useful in start-up situations. Comparison results and a real data example show that our proposed procedure has great potential for application.

Key words: Change-Point; Minimal Spanning Tree; Multivariate Goodness-of-Fit Test; Multivariate Statistical Process Control
1991 MSC: 62P30

1 Introduction

Multivariate statistical process control (MSPC) are particularly useful, when there is need to monitor several quality characteristics of a process simultaneously (Stoumbos et al. (2000)). It is usually assumed that there are m_0 independent and identically distributed (iid) historical observations, $\mathbf{x}_{-m_0+1}, \ldots, \mathbf{x}_0 \in$

* Corresponding author.
Email address: zli@nankai.edu.cn (Zhonghua Li).

$\mathbb{R}^p$, for some integer, $p \geq 1$, and the ith future observation, $\mathbf{x}_i$, is collected over time from the following multivariate change-point model

$$\mathbf{x}_i \sim \begin{cases} F_0(\mathbf{x}), \text{ for } & i = -m_0 + 1, \ldots, 0, 1, \ldots, \tau, \\ F_1(\mathbf{x}), \text{ for } & i = \tau + 1, \ldots, \end{cases} \tag{1}$$

where τ is the unknown change point, $F_0(\mathbf{x}) \neq F_1(\mathbf{x})$ are respectively the pre-change distribution function and the post-change distribution function. This change-point model is also employed in this paper.

Most MSPC methods are based on a fundamental assumption that the process data have multinormal distributions (see Zou and Tsung (2011) for more details), and some recent works based on this assumption include Alkahtani and Schaffer (2012) and Lee (2012). However, it is well recognized that in many applications, the underlying process distribution is unknown and not multinormal, so that the statistical properties of commonly used procedures, which were designed to perform best under the normal distribution, could potentially be (highly) affected (Montgomery (2005)). Distribution-free or robust procedures may be useful in such situations.

In the last several years, univariate nonparametric control charts have attracted much attention from researchers and a nice overview of this topic was presented by Chakraborti et al. (2001). See Zou and Tsung (2010); Hawkins and Deng (2010); Qiu and Li (2011a,b), and the references therein for some recent development. Some efforts have also been devoted to robust MSPC. Liu (1995) and Li et al. (2013a) proposed control schemes based on data-depth; Qiu and Hawkins (2001) and Qiu and Hawkins (2003) suggested a computationally trivial nonparametric multivariate CUSUM procedure based on the antiranks of the measurement components; Stoumbos and Sullivan (2002) recommended the classical MEWMA chart because it is robust in the sense that the in-control (IC) run length distribution for a continuous non-normal process is quite close to the distribution for a multinormal process with the same control limit if the weighting parameter, λ, is small; Qiu (2008) proposed a distribution-free multivariate CUSUM procedure based on log-linear modeling; Zou and Tsung (2011) developed a multivariate sign EWMA (MSEWMA) control chart for monitoring location parameters; Phaladiganon et al. (2011) proposed a bootstrap-based multivariate T^2 control chart that can efficiently monitor a process when the distribution of observed data is nonnormal or unknown; and Li et al. (2013b) developed a multivariate spatial-sign EWMA (SSEWMA) control scheme for monitoring shape parameters.

The good performance of the works above is generally based on a large number of historical observations so as to have enough information on the unknown distribution $F_0(\cdot)$ in Eq. (1). In many applications, however, we have no many

historical observations. The number of IC historical observations used for calibrating the necessary parameters are often rather small. In such situations, there would be considerable uncertainty in the parameter estimation, which in turn would distort the IC run-length distribution. Mahmoud and Maravelakis (2010) showed through simulation that the performance of the MEWMA chart would be seriously affected if the vector of means and the covariance matrix are estimated based on a small number of Phase I samples. Self-starting methods, which can handle sequential monitoring and estimating simultaneously, were developed accordingly. See Hawkins and Olwell (1998) and Montgomery (2005) for a detailed review. Recently, Zou et al. (2012) developed a multivariate self-starting method based on spatial rank EWMA (SREWMA) for monitoring location parameters. It has distribution-free properties over a broad class of population models in the sense that the IC run-length distribution is (or is always very close to) the nominal one when the same control limit designed for a multinormal distribution is used. But their method leaves a tuning parameter λ to choose, and when λ is not small, say $\lambda \geq 0.05$, its IC ARL performance will be unsatisfactory. So their procedure may be only efficient to small or moderate shifts.

This paper develops a new distribution-free multivariate procedure for statistical process control by integrating a powerful two-sample multivariate number of runs test based on MST (Friedman and Rafsky (1979)) into the effective change-point model. Simulation studies show that the proposed method is superior to SREWMA scheme of Zou et al. (2012) in monitoring moderate to large process shifts. The reason why we compare our proposed procedure with SREWMA scheme of Zou et al. (2012) is that there is no other corresponding distribution-free and self-starting multivariate detecting scheme as far as we know. As our proposed procedure avoids the need for a lengthy data-gathering step before charting (although it is generally necessary and advisable to collect several warm-up samples) and it does not require knowledge of the underlying distribution, the proposed procedure is particularly useful in start-up or short-run situations.

The rest of this paper is organized as follows. The brief review, description, design of our proposed procedure are given in Section 2. The performance comparisons with SREWMA scheme of Zou et al. (2012) are discussed in Section 3. A real data example is considered in Section 4. And the conclusion and discussion of the proposed chart are given in Section 5.

2 Methodology

Our proposed methodology is described in four parts. In Sections 2.1 and 2.2, brief reviews of the multivariate number of runs test based on MST and the

SREWMA scheme of Zou et al. (2012) are presented, respectively. In Section 2.3, a distribution-free multivariate procedure based on MST is derived for Phase I. In Section 2.4, our method is extended to self-starting control scheme, which can be used for monitoring and estimating simultaneously.

2.1 A review of multivariate number of runs test

Consider samples of size m and n respectively from distributions F_x and F_y, both defined on $\mathbb{R}^p$. The null hypothesis H_0 to be tested is $F_x = F_y$. We are interested in general alternative hypothesis $F_x \neq F_y$.

In the two-sample problem, we consider the edge weighted graph consisting of the N ($N = m+n$) pooled sample data points as nodes, and edges linking all pairs. This "complete" graph has $N(N-1)/2$ edges. Take the weight associated with each edge to be Euclidean distance between the nodes (points) defining it. The MST of this graph is thus the subgraph of minimum total distance that provides a path between every two nodes.

Friedman and Rafsky (1979) suggested that the multivariate number of runs test as follows: (1) construct the MST of the pooled sample data points, (2) remove all edges for which the defining nodes originate from different samples, and (3) define the test statistic R as the number of disjoint subtrees. Rejection of H_0 is for a small number of subtrees.

Friedman and Rafsky (1979) also derived the expectation and variance of the statistic R as

$$E[R] = \frac{2mn}{N} + 1, \tag{2}$$

$$Var[R|C] = \frac{2mn}{N(N-1)} \left\{ \frac{2mn - N}{N} + \frac{C - N + 2}{(N-2)(N-3)} [N(N-1) - 4mn + 2] \right\}, \tag{3}$$

where C is the number of edge pairs that share a common node.

2.2 A review of SREWMA scheme

Zou et al. (2012) developed a new multivariate self-starting methodology for monitoring location parameters, which is based on adapting the multivariate spatial rank to on-line sequential monitoring. The weighted version of the rank-based test is used to formulate the charting statistic by incorporating

the EWMA control scheme. They used the following multivariate location change-point model:

$$\mathbf{x}_i \sim \begin{cases} \mu_{\mathbf{0}} + \mathbf{\Omega}\varepsilon_{\mathbf{i}}, \text{ for} & i = -m_0 + 1, \ldots, 0, 1, \ldots, \tau, \\ \mu_{\mathbf{1}} + \mathbf{\Omega}\varepsilon_{\mathbf{i}}, \text{ for} & i = \tau + 1, \ldots, \end{cases} \tag{4}$$

The charting statistic is given by

$$Q_t = \frac{2-\lambda}{\lambda} \mathbf{w_t}^T \{cov[R_F(\mathbf{M}\mathbf{x_t})]\}_t^{-1} \mathbf{w_t}, \tag{5}$$

where $\mathbf{w_t} = (1-\lambda)\mathbf{w_{t-1}} + \lambda R_F(\mathbf{M}\mathbf{x_t})$, $\mathbf{M} = \mathbf{\Omega}^{-1}$, $R_F(\mathbf{x}) = E_y[U(\mathbf{x}-\mathbf{y})]$, $\mathbf{y}$ is distributed according to F, and U($\mathbf{x}$) is the spatial sign function. See more details of the derivation and application of the charting statistic in Zou et al. (2012).

2.3 Multivariate procedure for Phase I

We begin by considering the Phase I problem of detecting a change-point in a fixed-size sequence of observations. We denote the observations available by $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$, and the goal is to test whether they are with the same probability distribution. Note that we assume that no prior knowledge is available regarding this distribution, other than that it is continuous.

Using the terminologies of statistical hypothesis testing, the null hypothesis is that there is no change-point so that all the observations come from the same distribution, while the alternative hypothesis is that there exists a single change-point τ in the sequence which partitions the data into two sets, with $\mathbf{x}_1, \ldots, \mathbf{x}_\tau$ coming from the pre-change distribution F_0, and $\mathbf{x}_{\tau+1}, \ldots, \mathbf{x}_t$ coming from a different post-change distribution F_1, i.e.,

$$H_0 : \mathbf{x}_i \sim F_0, \text{ for } i = 1, \ldots, t,$$

$$H_1 : \mathbf{x}_1, \ldots, \mathbf{x}_\tau \sim F_0, \ \mathbf{x}_{\tau+1}, \ldots, \mathbf{x}_t \sim F_1.$$

If we know the change point is at sample k, we can test for the change-point immediately by partitioning the observations into two samples $S_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and $S_2 = \{\mathbf{x}_{k+1}, \ldots, \mathbf{x}_t\}$ of sizes $n_1 = k$ and $n_2 = t - k$ respectively, and then performing an appropriate two sample hypothesis test. Here, we consider the opposite number of the standardized two-sample multivariate number of runs

test, i.e.,

$$W_{k,t} = -\frac{R_{k,t} - E[R_{k,t}]}{(Var[R_{k,t}|C])^{\frac{1}{2}}}, \tag{6}$$

where $R_{k,t}$ denotes the multivariate number of runs test statistic for the two samples S_1 and S_2, $E[R_{k,t}]$ and $Var[R_{k,t}|C]$ can be computed through Eq. (2) and Eq. (3), respectively. We reject the null hypothesis that no change occurs at k if $W_{k,t} > h_{k,t}$ for some appropriately chosen value of $h_{k,t}$.

Note that we use the opposite number rather than the original number in Eq. (6). Because we prefer monitoring statistics getting larger if the process is out-of-control, and if there exist some shifts in the observations, the runs of these observations will become smaller so that $W_{k,t}$ will become larger.

The statistic in Eq. (6) can be directly integrated into the change-point model. Because we have no idea in advance where the change-point is located, we do not know which value of k to use for partitioning. We can perform this test by computing $W_{k,t}$ for each integer value $0 < k < t$, and this leads to the maximized test statistic:

$$W_t = \max_{0<k<t} W_{k,t}. \tag{7}$$

If $W_t > h_t$ for some suitably chosen threshold h_t, then the null hypothesis is rejected, and we conclude that a change occurred at some point in the data. If $W_t \leq h_t$, then we do not have enough information to reject the null hypothesis, and hence conclude that no change has occurred. The choice of the threshold h_t will be discussed in detail in the following subsection.

2.4 Multivariate procedure for sequentially monitoring

Having considered the problem of detecting changes in a fixed-size sample, we now turn our attention to the task of sequentially Phase II monitoring where new observations are being collected over time. Let $\mathbf{x}_t$ denote the tth observation where t is increasing over time.

Once a new observation $\mathbf{x}_t$ is collected, we regard $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ to be a fixed-size sample, and employ our proposed method based on Eq. (7) to test whether a change-point has occurred. The problem of sequentially monitoring is then reduced to performing a sequence of fixed-size tests, although the sample size is increasing over time.

For Phase II monitoring, a commonly used criterion is the average run length

(ARL). Suppose it is desired to have the IC ARL (denoted as ARL_0) equal to γ. This can be achieved if we choose the threshold h_t values such that the probability α of incurring a false alarm at the tth observation equals to $1/\gamma$, i.e.,

$$\begin{aligned} &P(W_1 > h_1) = \alpha \\ &P(W_t > h_t | W_{t-1} \leq h_{t-1}, \ldots, W_1 \leq h_1) = \alpha, t > 1. \end{aligned} \tag{8}$$

It is not trivial to find such sequence of h_t values which satisfy the requirement in Eq. (8). We adopt the Monte-Carlo simulation approach, which was also employed in Hawkins and Deng (2010). To be more specific, 200,000 realizations of the sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_{1000}\}$ were firstly generated. Note that Henze and Penrose (1999) proved that the multivariate number of runs test statistic $R_{m,n}$ is asymptotically distribution-free under H_0. And our simulation studies show that the h_t values are nearly the same under any continuous distributions. So these $\mathbf{x}_i$ values can be iid sampled from any continuous distribution. Then for each value of t, W_t in Eq. (7) is computed for each of the 200,000 realizations. The values for h_t corresponding to the desired ARL_0 can be obtained through the corresponding sample α percentiles. The values for h_t are essentially critical values for the test statistic W_t in Eq. (7), we, however, regard them as the dynamic control limits, as is done in Hawkins and Deng (2010).

Table 1 shows the values of h_t when $\text{ARL}_0 = 200$, $p = 5, m_0 = 10$ and $p = 10, m_0 = 20$, where m_0 is the number of the warm-up data. Note that these values appear to have converged by the 1000 observation, so if the stream contains more than 1000 observations, it will be acceptable to let $h_t = h_{1000}$ for $t > 1000$. The property of convergence for other commonly used ARL_0, or other dimensions, is still hold, so we do not report them in Table 1 to save space. The Fortran programmes for implementing our proposed scheme, including the procedures for finding h_t, are available from the authors upon request. Henceforth, our proposed procedure is referred to as the SMMST procedure for abbreviation, which is short for self-starting multivariate minimal spanning tree.

[Insert Table 1 about here]

3 Performance comparison

We present some simulation results in this section regarding the performance of our proposed SMMST procedure and compare it with the SREWMA procedure of Zou et al. (2012), which is reviewed in Section 2.2. Comparing the

SMMST procedure with alternative distribution-free methods turned out to be difficult due to the lack of an obvious comparable method. This is because most of the approaches in the literature were designed for the cases where sufficient historical observations are available so as to accurately estimate the IC distribution of the process or some IC parameters. See Zou and Tsung (2011) for some discussions and reviews. A Shewhart-type procedure is expected to detect large shifts more efficiently than change-point approaches. As we know, the multivariate version of Shewhart chart is not distribution-free, and we only aim to focus on distribution-free procedure in this paper. Thus, we consider the only exiting distribution-free and self-starting multivariate procedure SREWMA proposed by Zou et al. (2012), as we are aware of so far. All the results in this section are obtained from 10,000 replications unless indicated otherwise.

Following the robustness analysis of Zou and Tsung (2011) and Zou et al. (2012), we consider the following distributions: (i) multinormal; (ii) multivariate t with ζ degrees of freedom, denoted as $t_{p,\zeta}$; (iii) multivariate gamma with shape parameter ζ and scale parameter 1, denoted as $\text{Gam}_{p,\zeta}$. In addition, the following nonnormal distribution is involved in the comparison: (iv) in each observed vector, the first $[p/2]$ measurement components are i.i.d. t distributed with ζ_1 degrees of freedom and the other $p - [p/2]$ measurement components are i.i.d. chi-square distributed with ζ_2 degrees of freedom, where $[\cdot]$ is the function that rounds off a number to its nearest integer. The reason for considering (iv) is that unlike (i)-(iii), its marginal distributions are not all the same.

Note that the number and variety of covariance matrices and shift directions are too large to allow a comprehensive, all-encompassing comparison. Our goal is to show the effectiveness, robustness and sensitivity of the SMMST procedure, and thus we only choose certain representative models for illustration. Specifically, for the first three distribution cases, the covariance matrix $\sum_0 = (\sigma_{ij})_{p\times p}$ is chosen to be $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.5^{|i-j|}$, for $i, j = 1, 2, \ldots, p$. For brevity, a shift of size δ keeps the same in all components, i.e., $\mathbf{x}_i + \delta\mathbf{e}$ with $\mathbf{e} = (1, 1, \ldots, 1)^T$. Similar conclusions given below hold for other simulation settings with various shift types.

The SMMST and SREWMA procedures are compared in terms of the out-of-control average run length (OC ARL, denoted as ARL_1). Because similar conclusions hold for other cases, throughout this section, we only present the results when $\text{ARL}_0 = 200$ for illustration. A lower-dimensional case with $p = 5$ and a higher-dimensional case with $p = 10$ are involved for each distribution considered. We fix the number of warm up data $m_0 = 10$ and 20 for $p = 5$ and 10, respectively. This also satisfies the requirement for starting SREWMA ($m_0 \geq p + 2$), however, one has to keep in mind that there is no this kind of requirement for our SMMST procedure, although our SMMST procedure

will perform better if the number of warm up data m_0 is larger. For the ARL_1 comparison, we consider the steady-state ARL (SSARL). To evaluate the SSARL behavior of each chart, any series in which a signal occurs before the $(\tau + 1)$th observation is discarded.

We first consider the multinormal distribution. The simulation results for the SREWMA with $\lambda = 0.05$ and our SMMST procedure are presented in Table 2. Apart from the parameters above, the performance of self-starting procedures depends on the choice of τ . We consider $\tau = 40$ and 90. From this table, we observe that the SREWMA chart performs better than our SMMST procedure with small shifts as we would expect, since the tuning parameter used here $(\lambda = 0.05)$ is a small one, which is sensitive to small shifts. But the SMMST procedure performs better with moderate and large shifts, say $\delta \geq 1.5$.

[Insert Table 2 about here]

As we can see from Table 2, the SMMST procedure is better, when $\delta \geq 1.5$. Note that for $\delta \geq 1.5$, the modulus of shifts is $\delta\sqrt{p} \geq 1.5\sqrt{p}$. From the perspective of modulus, we can only conclude that, the SMMST procedure is better for moderate to large shifts. Most charting techniques in the literature, however, do not care about the detection of large shifts or have quite little power in detecting large shifts, because it is believed that one can notice the large shifts even with personal experiences instead of a chart, when large shifts occur. But we do not think it is the same case in high dimensional monitoring. Because in this circumstance, one can only check the process with experiences dimension by dimension. Therefore, although the modulus of shifts $\delta\sqrt{p}$ may be large with a not large δ, but a high dimension p, (say, $\delta = 1.5, p = 10$), one can not notice it immediately as in one dimension scenario. It is natural to adjust some tuning parameters to make control chart sensitive to moderate to large shifts, for example, take large λ in the SREWMA chart of Zou et al. (2012). But the SREWMA chart undergos a poor performance of ARL_0 with large λ, as is pointed out in Zou et al. (2012). From this perspective of view, our SMMST procedure is particularly useful with moderate to large modulus of shifts.

Next, the multivariate t distribution and the multivariate gamma distribution are considered. Tables 3 and 4 give ARL values with multivariate t observations with five degrees of freedom ($t_{p,5}$) and with multivariate gamma observations with three degrees of freedom ($\text{Gam}_{p,3}$), respectively. In these tables, we fix $m_0 + \tau = 100$ for each comparison scenario for simplicity. For the SREWMA chart, the value of λ is chosen to be 0.025, because the bigger the λ value is, the worse the performance of ARL_0 is. For example, Zou et al. (2012) showed that even when $\lambda = 0.05$ with $t_{10,3}$, the ARL_0 value is 177, which is far from the nominal value 200, hence not very satisfactory.

[Insert Tables 3-4 about here]

The results in Tables 3-4 are similar. The SREWMA chart is efficient in detecting the small shifts than the SMMST procedure. The SMMST is more efficient in detecting moderate to large shifts, such as $\delta \geq 1.5$. This is as expected, because the SREWMA, which is essentially based on ranks rather than distances, shares a similar drawback as those rank-based charts for univariate processes. That is, even though the shift is quite large, the ranks of the observations may not be able to grow larger. Moreover, here we use a small tuning parameter ($\lambda = 0.025$), which is very sensitive to small shifts. So it is not surprising that the SREWMA chart here is better than our SMMST procedure in small shifts.

Table 5 shows the ARL values of the SMMST and SREWMA procedures in monitoring a shift of the multivariate mixed-components observations. Again, the SMMST procedure is more efficient in detecting moderate to large shifts, such as $\delta \geq 1.5$. This demonstrates that the SMMST procedure is more sensitive to moderate to large process shifts in non-normal observations, even for a distribution with different marginal distributions.

[Insert Table 5 about here]

4 A real data example

In this section, we illustrate our SMMST procedure using a real data example from a white wine production process from May 2004 to February 2007. The data contains totally 4898 observations, and is publicly available in the "Wine Quality Data Set" of the UCI Machine Learning Repository and can be downloaded from http://archive.ics.uci.edu/ml/datasets/Wine+Quality. The data were recorded by a computerized system, which automatically manages the process of wine sample testing from producer requests to laboratory and sensory analysis. For each of these observations, there are eleven continuous measurements (based on physicochemical tests) including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol (denoted by $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{11}$, respectively). Another categorical variable, quality, indicating the wine quality between 0 (very bad) and 10 (very excellent), is also provided based on sensory analysis. The goal of this data analysis is mainly to model and monitor wine quality based on physicochemical tests. Interested readers are referred to Cortez et al. (2009) for more detail about this example and data set.

As pointed out by Cortez et al. (2009), it is desirable to setup an on-line

detection system to monitor the production process of *Vinho Verde* wine to guarantee its quality. Under the MSPC context of sequentially monitoring the wine production process, we assume that the standard quality level is 7 (LV7; as also suggested by Cortez et al. (2009)). As shown in Zou et al. (2012) and Li et al. (2013b), the sample correlation matrix of this data contains several large entries, which demonstrates that the variables have considerable interrelationships and consequently a multivariate control chart is likely to be more appropriate than a univariate control chart. Zou et al. (2012) and Li et al. (2013b) also showed that the multivariate normality assumption is not valid and thus we could expect that the SMMST chart would be more robust and powerful than normal-based approaches for this data set.

It is also worth noting that although it is usually easy to collect observations from physicochemical tests, obtaining sufficiently large IC reference sample in this process is difficult: sensory tests rely mainly on human experts, and thus are rather time-consuming and expensive. Once a new technology in wine making is used or some improvements are made in the production process, one usually wants to monitor the process at the start-up stages in which only a small reference sample (through sensory tests) would be available. Therefore, our SMMST chart, which is a self-starting control chart, would be more desirable in this situation.

Next, we assume that we have $m_0 = 20$ historical observations from LV7 and initially monitored another 20 observations from LV7 and then collected the LV6 observations sequentially. We construct the SMMST procedure to monitor the wine quality. The ARL_0 is fixed at 500. Figure 1 shows the resulting SMMST procedure statistics along with its threshold h_t values (the dashed curve).

[Insert Figure 1 about here]

From Figure 1, it can be seen that the SMMST procedure exceeds its threshold h_t from around the 24th observation (the 4th LV6 observation). This excursion suggests that a marked change has occurred. And its delay to detect this change is 4 observations. In comparison, in the SREWMA chart of Zou et al. (2012) with $m_0 = 30$ historical observations from LV7, the delay is 25 observations. So, the SMMST is a reasonable alternative for non-multinormal processes if we take its efficiency and robustness into account.

5 Conclusions

We developed a new distribution-free multivariate change-point model for statistical process control based on minimal spanning tree. It integrates a

two-sample multivariate goodness-of-fit (GOF) test based on MST into the change-point model. It is robust to non-normally distributed data, and efficient in detecting multivariate process shifts, especially moderate to large shifts. As it avoids the need for a lengthy data-gathering step and it does not require knowledge of the underlying distribution, our proposed procedure is particularly useful in start-up or short-run situations.

It is worth pointing out here that apart from quick detecting abnormal changes, isolating the shifted components or factors that are responsible for the change is also a fundamental task of MSPC. For example, in the application from the above section, it would be interesting and helpful to determine which physicochemical factors are responsible for the change of quality. The problem of making the LASSO-based post-signal diagnostic method proposed by Zou and Qiu (2009) suitable to be used after our proposed SMMST procedure triggers a signal warrants further research.

Acknowledgements

The authors are grateful to the editor and the anonymous referee for their valuable comments that have vastly improved this paper. This paper is supported by the NNSF of China Grants 11001138, 11071128, 11131002, 11101306, 11201246 and the RFDP of China Grant 20110031110002, and the Fundamental Research Funds for the Central Universities 65012231.

References

Alkahtani, S., Schaffer, J., 2012. A double multivariate exponentially weighted moving average (dMEWMA) control chart for a process location monitoring. Commun. Stat.-Simul. C. 41 (2), 238-252.

Chakraborti, S., Van der Laan, P., Bakir, S.T., 2001. Nonparametric control charts: an overview and some results. J. Qual. Tech. 33, 304-315.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. Decision Support Syst. 47, 547-553.

Friedman,J.H., Rafsky, L.C., 1979. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. Ann. Stat. 7, 697-717.

Hawkins, D.M., Deng, Q., 2010. A nonparametric change point control chart. J. Qual. Tech. 42, 165-173.

Hawkins, D.M., Olwell, D.H., 1998. *Cumulative sum charts and charting for quality improvement*. New York: SpringerVerlag.

Henze, N., Penrose, M.D., 1999. On the multivariate runs test. Ann. Stat. 27, 290-298.

Lee, M.H., 2012. The design of the multivariate synthetic exponentially weighted moving average control chart. Commun. Stat.-Simul. C. 41 (10), 1785-1793.

Li, Z., Dai, Y., Wang, Z., 2013a. Multivariate change point control chart based on data depth for phase I analysis. To appear in Commun. Stat.-Simul. C, DOI:10.1080/03610918.2012.735319.

Li, Z., Zou, C., Wang, Z., Huwang, L., 2013b. A multivariate sign chart for monitoring process shape parameters. J. Qual. Tech. 45 (2), 149-165.

Liu,R., 1995. Control charts for multivariate processes. J. Amer. Stat. Assoc. 90, 1380-1388.

Mahmoud, M.A., Maravelakis, P.E., 2010. The performance of the MEWMA control chart when parameters are estimated. Commun. Stat.-Simul. C. 39 (9), 1803-1817.

Montgomery, D.C. (2005). *Introduction to statistical quality control*, 5th ed. John Wiley & Sons: New York.

Phaladiganon, P., Kim, S.B., Chen, V.C.P., Baek, J.-G., Park, S.-K., 2011. Bootstrap-based T^2 multivariate control charts. Commun. Stat.-Simul. C. 40 (5), 645-662.

Qiu, P., 2008. Distribution-free multivariate process control based on log-linear modeling. IIE Trans. 40, 664-677.

Qiu, P., Hawkins, D.M., 2001. A rank-based multivariate CUSUM procedure. Technometrics. 43, 120-132.

Qiu, P., Hawkins, D.M., 2003. A nonparametric multivariate CUSUM procedure for detecting shifts in all directions. J. R. Stat. Soc. Ser D. 52, 151-164.

Qiu, P., Li, Z., 2011a. On nonparametric statistical process control of univariate processes. Technometrics. 53, 390-405.

Qiu, P., Li, Z., 2011b. Distribution-free monitoring of univariate processes. Statist. Probab. Lett. 81 (12), 1833-1840.

Stoumbos, Z.G., Reynolds, M.R., Ryan, T.P., Woodall, W.H., 2000. The state of statistical process control as we proceed into the 21st century. J. Amer. Stat. Assoc. 95, 992-998.

Stoumbos, Z.G., Sullivan, J.H., 2002. Robustness to non-normality of the multivariate EWMA control chart. J. Qual. Tech. 34, 260-276.

Zou, C., Qiu, P., 2009. Multivariate statistical process control using LASSO. J. Amer. Stat. Assoc. 104, 1586-1596.

Zou, C., Tsung, F., 2010. Likelihood ratio based distribution-free EWMA schemes. J. Qual. Tech. 42, 174-196.

Zou, C., Tsung, F., 2011. A multivariate sign EWMA control chart. Technometrics. 53, 84-97.

Zou, C., Wang, Z., Tsung, F., 2012. A spatial rank-based multivariate EWMA control chart. Nav. Res. Logistic. 59, 91-110.

Table 1
Values of the threshold sequence h_t corresponding to $p = 5, 10$ and $\mathrm{ARL}_0 = 200$.

t	p=5	p=10	t	p=5	p=10
1	2.835	3.034	60	2.832	2.693
2	2.710	2.987	70	2.812	2.676
3	2.717	2.883	80	2.792	2.664
4	2.822	2.889	90	2.796	2.653
5	2.785	2.820	100	2.777	2.657
6	2.829	2.850	200	2.724	2.552
7	2.828	2.864	300	2.642	2.489
8	2.849	2.791	400	2.620	2.447
9	2.784	2.844	500	2.619	2.436
10	2.822	2.822	600	2.544	2.428
20	2.892	2.792	700	2.491	2.417
30	2.875	2.757	800	2.453	2.405
40	2.868	2.724	900	2.438	2.399
50	2.830	2.710	1000	2.431	2.394

Table 2
ARL values with multinormal distributions.

		SMMST		SREWMA	
	δ	τ =40	τ =90	τ =40	τ=90
$p = 5$	0	200	203	197	198
	1	14.2	12.8	11.4	9.93
	1.5	7.35	6.46	7.69	6.95
	2	4.97	4.95	6.39	5.77
	3	3.59	3.35	5.42	4.89
	4	3.38	3.29	5.10	4.59
$p = 10$	0	199	202	196	197
	1	11.0	10.1	9.57	8.77
	1.5	6.26	6.13	6.80	6.27
	2	4.27	4.13	5.73	5.25
	3	3.85	3.81	4.89	4.44
	4	3.61	3.49	4.60	4.17

Table 3
ARL values with multivariate $t_{p,5}$ distributions.

	δ	SMMST	SREWMA
$p = 5$	0	197	202
	1	19.8	13.2
	1.5	9.13	9.28
	2	7.43	7.87
	3	4.20	6.35
	4	3.52	5.84
$p = 10$	0	203	196
	1	12.8	11.7
	1.5	7.72	8.41
	2	5.84	7.01
	3	4.21	5.80
	4	3.90	5.35

Table 4
ARL values with multivariate $\text{Gam}_{p,3}$ distributions.

	δ	SMMST	SREWMA
$p = 5$	0	201	203
	1	14.3	10.7
	1.5	7.64	8.03
	2	5.01	6.89
	3	3.45	6.01
	4	3.06	5.69
$p = 10$	0	203	201
	1	11.5	9.73
	1.5	5.97	7.32
	2	4.13	6.34
	3	3.83	5.52
	4	3.32	5.23

Table 5
ARL values with mixed-components multivariate distributions.

	δ	SMMST	SREWMA
$p = 5$	0	200	200
	1	19.8	12.4
	1.5	8.91	9.04
	2	7.06	7.37
	3	4.50	6.18
	4	3.48	5.74
$p = 10$	0	197	198
	1	18.2	10.0
	1.5	7.36	7.48
	2	5.78	6.34
	3	3.90	5.47
	4	3.74	5.13

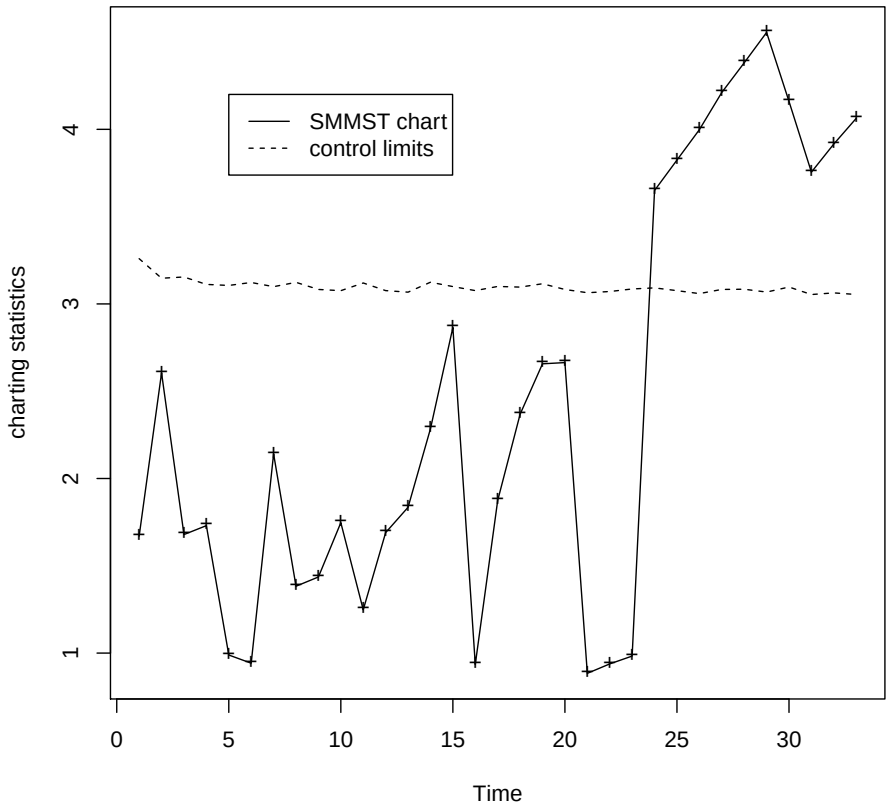

Fig. 1. The SMMST procedure for monitoring the white wine production process, along with the dashed curve indicating its critical values.