Distribution-Free Multivariate Process Control Based On Log-Linear Modeling

Peihua Qiu
School of Statistics
University of Minnesota

Abstract

This paper considers statistical process control (SPC) when the process measurement is multivariate. In the literature, most existing multivariate SPC procedures assume that the in-control distribution of the multivariate process measurement is known and it is a Gaussian distribution. In applications, however, the measurement distribution is usually unknown and it needs to be estimated from data. Furthermore, multivariate measurements often do not follow a Gaussian distribution (e.g., cases when some measurement components are discrete). We demonstrate that results from conventional multivariate SPC procedures are usually unreliable when the data are non-Gaussian. Existing statistical tools for describing multivariate non-Gaussian data, or, transforming the multivariate non-Gaussian data to multivariate Gaussian data are limited, making appropriate multivariate SPC difficult in such cases. In this paper, we suggest a methodology for estimating the in-control multivariate measurement distribution when a set of in-control data is available, which is based on log-linear modeling and which takes into account the association structure among the measurement components. Based on this estimated in-control distribution, a multivariate CUSUM procedure for detecting shifts in the location parameter vector of the measurement distribution is also suggested for Phase II SPC. This procedure does not depend on the Gaussian distribution assumption; thus, it is appropriate to use for most multivariate SPC problems.

Key Words: Discrete measurements; Log-linear modeling; Multivariate distribution; Non-Gaussian data; Nonparametric procedures; Statistical process control.

1 Introduction

This paper discusses statistical process control (SPC) when the process measurement is multivariate. In the literature, SPC is often divided into two phases. In Phase I, a set of process data is gathered and analyzed. Any unusual "patterns" in the data lead to adjustments and fine tuning of the process. Once all such assignable causes are accounted for, we are left with a clean set of data, gathered under stable operating conditions and illustrative of the actual process performance. This set is then used for estimating the in-control distribution of the measurement. In Phase II, the estimated in-control measurement distribution from a Phase I data is used, and the major goal of this phase is to detect changes in the measurement distribution after an unknown time point. Performance of a Phase II SPC procedure is often measured by the average run length (ARL), which is the average number of observations needed for the procedure to signal a change in the measurement distribution. The in-control ARL value of the procedure is often controled at some specific level. Then, the procedure performs better if its out-of-control ARL is shorter, when detecting a given change. See, e.g., Chen *et al.* (2005), Chen and Zhang (2004), Mason *et al.* (1997), Woodall (2000), Yeh *et al.* (2003), Yeh *et al.* (2004), and Yeh *et al.* (2006) for detailed discussion.

In the literature, most multivariate SPC procedures (e.g., Healy 1987; Crosier 1988; Hawkins 1991; Yeh and Lin 2002, Yeh *et al.* 2006) discuss Phase II SPC for Gaussian data only. That is, they assume that the in-control measurement distribution is known, and this distribution is multivariate Gaussian. In applications, however, the multivariate measurement distribution should be estimated from data, and it is sometimes non-Gaussian, even when the process is in-control. In such cases, existing statistical tools for describing multivariate non-Gaussian distributions, or transforming a non-Gaussian data to a Gaussian data, are limited (cf., e.g., Fang *et al.* 1990). It is therefore challenging to estimate the in-control multivariate measurement distribution from a Phase I in-control data, for Phase II SPC, because we even do not know how to describe it when it is multivariate non-Gaussian.

Most existing Phase II SPC procedures may not be appropriate to use when the data are non-Gaussian, because in such cases their actual false alarm rate is often different from the assumed false alarm rate, and this difference is large in some cases. Stoumbos and Sullivan (2002) performed an excellent study to investigate the robustness of multivariate EWMA procedures to the Gaussian distribution assumption, and found that such procedures were quite robust when the distribution is continuous but non-Gaussian and when the procedure parameter r is chosen small. However, how small r should be depends on the skewness of the actual measurement distribution, which is difficult to measure in applications when the measurement distribution is multivariate non-Gaussian. Also, when r is chosen small, the corresponding EWMA procedure would not be sensitive enough to relatively large shifts. These issues will be further discussed in Section 4 with numerical examples.

To overcome the difficulty caused by the Gaussian distribution assumption, Qiu and Hawkins (2001) suggest a nonparametric multivariate CUSUM procedure, based on the antiranks of the measurement components. This procedure can only detect shifts in a location parameter of the measurement distribution, which are not on or close to one specific direction in which the components of the shift are all the same. This limitation is later lifted by Qiu and Hawkins (2003). But both papers only discuss the case when the in-control measurement distribution is assumed known; they have not discussed how to estimate the in-control distribution from an in-control dataset.

Nonparametric control charts for univariate SPC have been discussed by several authors, including Albers and Kallenberg (2004), Bakir (2004, 2005, 2006), Chakraborti *et al.* (2004), and some others. See Chakraborti *et al.* (2001) for a review on this topic.

This paper tries to make two contributions to the SPC literature. One is that a method is suggested for estimating the in-control, p-variate, non-Gaussian measurement distribution from an in-control dataset, where $p > 1$ is an integer. By this method, each measurement component is first transformed to a binary variable, which is an indicator function of the event that the measurement component is larger than its in-control median. Then, a log-linear model is used for describing possible associations among the p binary variables, providing a log-linear estimator of the in-control joint distribution of the p binary variables. This idea can be generalized, without much theoretical difficulty, in the way that each measurement component is transformed to a categorical variable with $q \geq 2$ categories, and the log-linear modeling procedure is applied to the p categorical variables. The other contribution is that a multivariate CUSUM procedure is suggested for Phase II SPC, for detecting shifts in a location parameter of the measurement distribution, based on the estimated in-control measurement distribution by the suggested log-linear modeling approach. This procedure is distribution-free in the sense that all its properties depend on the distribution of the p binary variables only; thus, it is appropriate to use for most multivariate SPC problems.

The rest of the article is organized as follows. In next section, the log-linear modeling procedure for estimating the in-control measurement distribution is introduced in details. In Section 3, the Phase II multivariate CUSUM procedure is discussed. Then, some numerical examples are presented in Section 4, for evaluating the performance of these procedures. A real-data application is also discussed there. Finally, several remarks conclude the article in Section 5.

2 Estimation of the In-Control Distribution

At the end of Phase I SPC, suppose that all bugs are fixed and we are left with a clean data set $\{\mathbf{X}(i) = (X_1(i), X_2(i), \ldots, X_p(i))', i = 1, 2, \ldots, n_1\}$, where n_1 is a fixed sample size. This sequence of observations are assumed to be i.i.d. with a common cumulative distribution function $F(\mathbf{x})$, which is also the in-control distribution of the process measurements. In the statistical literature, there are very limited tools for describing $F(\mathbf{x})$ when it is non-Gaussian, or, for transforming $\mathbf{X}(i)$ such that the transformed measurement follows a Gaussian distribution. A major difficulty lies behind proper description of the association structure among the components of $\mathbf{X}(i)$. When the components of $\mathbf{X}(i)$ are all categorical, however, there are a number of methods for describing the distribution of $\mathbf{X}(i)$. The most popular one is the log-linear modeling approach (see e.g., Agresti 2002, Chapter 8). In this section, we describe how to use this approach for estimating the in-control distribution of $\mathbf{X}(i)$.

Assume that the in-control median of $X_j(i)$ is m_j, for $j = 1, 2, \ldots, p$, which can be estimated well from the in-control data $\{\mathbf{X}(i), i = 1, 2, \ldots, n_1\}$. We then define

$$Y_j(i) = I(X_j(i) > m_j), \quad \text{for } j = 1, 2, \ldots, p, \tag{1}$$

and $\mathbf{Y}(i) = (Y_1(i), Y_2(i), \ldots, Y_p(i))'$, where $I(a)$ is an indicator function which equals 1 if a is "true" and 0 otherwise. So, equation (1) transforms the p original measurement components to p binary variables. Note that, in equation (1), the in-control median m_j can be replaced by the more general in-control r-th quantile of $X_j(i)$, with r being any real number in $(0, 1)$. We consider using the median, which is the 0.5-th quantile, because the resulting joint distribution of $\mathbf{Y}(i)$ could be more efficiently estimated by the log-linear modeling approach discussed in this section, due to the fact that relatively less cell probabilities of the contingency table formed by the components of $\mathbf{Y}(i)$ would be small in such a case. See Chapter 8 of Agresti (2002) for related discussion.

Of course, we lose information by transforming $\mathbf{X}(i)$ to $\mathbf{Y}(i)$. But, it is not difficult to check that the distribution of $\mathbf{Y}(i)$ would be changed by any shift in the median vector $(m_1, m_2, \ldots, m_p)'$ of the process, as long as the in-control measurement distribution $F(\mathbf{x})$ has a positive probability to take values in any neighborhood of its in-control median vector. Therefore, if we are interested in detecting shifts in a location parameter vector (e.g., the median vector $(m_1, m_2, \ldots, m_p)'$), then $\mathbf{Y}(i)$ is appropriate to use. If we are also interested in detecting other changes in $F(\mathbf{x})$ (e.g., changes in the covariance matrix of $F(\mathbf{x})$), then $\mathbf{Y}(i)$ needs to be modified. In the latter case, one possible approach is to transform each component of $\mathbf{X}(i)$ to a categorical variable with more than two categories. This issue will not be discussed in details in this article to keep the paper within reasonable length, and it is left for our future research.

Next, we discuss the log-linear modeling in the case of $p = 3$, for simplicity of presentation. For cases of $p > 3$, the model can be discussed similarly, although the notation would be more complicated. Let $O_{j_1 j_2 j_3}$ be the observed cell count of the (j_1, j_2, j_3)-th cell of the 3-way contingency table of the in-control data, with the three binary variables Y_1, Y_2 and Y_3 defined in equation (1) as classifiers, for $j_1, j_2, j_3 = 0, 1$. Then, a saturated log-linear model is defined by

$$log(O_{j_1 j_2 j_3}) = \mu + \lambda_{j_1}^{Y_1} + \lambda_{j_2}^{Y_2} + \lambda_{j_3}^{Y_3} + \lambda_{j_1 j_2}^{Y_1 Y_2} + \lambda_{j_1 j_3}^{Y_1 Y_3} + \lambda_{j_2 j_3}^{Y_2 Y_3} + \lambda_{j_1 j_2 j_3}^{Y_1 Y_2 Y_3}, \text{ for } j_1, j_2, j_3 = 0, 1, \qquad (2)$$

where μ is a constant term, $\lambda_{j_1}^{Y_1}, \lambda_{j_2}^{Y_2}$ and $\lambda_{j_3}^{Y_3}$ are the main effects of Y_1, Y_2 and Y_3, respectively, $\lambda_{j_1 j_2}^{Y_1 Y_2}, \lambda_{j_1 j_3}^{Y_1 Y_3}$ and $\lambda_{j_2 j_3}^{Y_2 Y_3}$ are the 2-way interaction terms, and $\lambda_{j_1 j_2 j_3}^{Y_1 Y_2 Y_3}$ is the 3-way interaction term. To make all parameters estimable, they should satisfy some conditions. One set of such conditions is as follows:

$$\sum_{j_1} \lambda_{j_1}^{Y_1} = \sum_{j_2} \lambda_{j_2}^{Y_2} = \sum_{j_3} \lambda_{j_3}^{Y_3} = 0;$$
$$\sum_{j_1} \lambda_{j_1 j_2}^{Y_1 Y_2} = \sum_{j_2} \lambda_{j_1 j_2}^{Y_1 Y_2} = 0; \; \sum_{j_1} \lambda_{j_1 j_3}^{Y_1 Y_3} = \sum_{j_3} \lambda_{j_1 j_3}^{Y_1 Y_3} = 0; \; \sum_{j_2} \lambda_{j_2 j_3}^{Y_2 Y_3} = \sum_{j_3} \lambda_{j_2 j_3}^{Y_2 Y_3} = 0;$$
$$\sum_{j_1} \lambda_{j_1 j_2 j_3}^{Y_1 Y_2 Y_3} = \sum_{j_2} \lambda_{j_1 j_2 j_3}^{Y_1 Y_2 Y_3} = \sum_{j_3} \lambda_{j_1 j_2 j_3}^{Y_1 Y_2 Y_3} = 0.$$

The main effects $\lambda_{j_1}^{Y_1}, \lambda_{j_2}^{Y_2}$ and $\lambda_{j_3}^{Y_3}$ are related to the marginal distributions of Y_1, Y_2 and Y_3. Because medians are used in equation (1) for defining Ys, $\lambda_{j_1}^{Y_1}, \lambda_{j_2}^{Y_2}$ and $\lambda_{j_3}^{Y_3}$ should be all zero in (2). However, in log-linear modeling, we should also follow the hierarchy principle that all lower-order terms need to be included in the model if a higher-order interaction term is in the model. For this reason, the main effects are still included in model (2).

Remark 2.1 Besides the reason given at the end of the paragraph containing equation (1), another main reason to use medians in equation (1) for defining Ys is for the property that: the main effects of model (2) are all zero. Otherwise, means or other location parameters of the measurement components can also be used for defining Ys. So, in model (2), there are actually $2 * 2 * 2 - 3 = 5$ non-redundant parameters. For a more general log-linear model, see Sections 8.1 and 8.2 of Agresti (2002) for related discussion about the number of non-redundant model parameters.

Model (2) can be denoted by $(Y_1Y_2Y_3)$, which lists the highest-order terms in the model for each variable. If the 3-way interaction term is not in the model, then the conditional association between any pair of Y_1, Y_2 and Y_3 is the same at the two levels of the remaining variable. In other words, the partial association between any pair of the variables is homogeneous across the different levels of the third variable. This model is denoted by (Y_1Y_2, Y_1Y_3, Y_2Y_3). Similarly, we use (Y_1Y_2, Y_1Y_3) to denote the model with the two 2-way interaction terms $\lambda_{j_1j_2}^{Y_1Y_2}$ and $\lambda_{j_1j_3}^{Y_1Y_3}$ included, and with the other 2-way interaction term $\lambda_{j_2j_3}^{Y_2Y_3}$ along with the 3-way interaction term $\lambda_{j_1j_2j_3}^{Y_1Y_2Y_3}$ excluded. That model implies that Y_2 and Y_3 are independent conditional on Y_1, the partial association between Y_1 and Y_2 is homogeneous across the different levels of Y_3, and the partial association between Y_1 and Y_3 is also homogeneous across the different levels of Y_2. So, model (2) and its variants can describe all kinds of possible association among Y_1, Y_2 and Y_3, by including appropriate 2-way and 3-way interaction terms.

There are some standard procedures for testing the goodness-of-fit of a log-linear model. These procedures include the Pearson's χ^2 test and the likelihood ratio G^2 test (cf., Agresti 2002, Section 8.3, for introduction). When the sample size is reasonably large in the sense that most cell counts in the related contingency table are larger than or equal to 5, these two tests usually give same conclusions.

Model selection based on the likelihood ratio test and the hierarchy principle is standard in the literature. In all numerical examples of this paper, we use the backward elimination procedure and the conventional rule that only one term is considered to be deleted at each step. For example, to test whether or not the 3-way interaction term should be deleted from model (2), we can use the following likelihood ratio test statistic:

$$G^2(M_0|M_1) = -2\log\left(\frac{\ell_{M_0}}{\ell_{M_1}}\right),$$

where ℓ_{M_0} and ℓ_{M_1} denote the likelihood functions of the submodel (Y_1Y_2, Y_1Y_3, Y_2Y_3) (denoted as M_0) and the full model $(Y_1Y_2Y_3)$ (denoted as M_1), respectively. Then the observed value of the test statistic can be compared to the $\chi^2(1)$ critical value for making decisions. Note that this test is based on the asymptotic distribution of $G^2(M_0|M_1)$. So, we should use it with caution when the sample size is small, although it has been shown in the literature that it is still quite reliable with fairly sparse tables (cf., Haberman 1977). For specific expressions of the likelihood functions used in the above equation, see expressions (9.3) and (9.4) in Agresti (2002, Section 9.2).

After the final model is determined, the expected cell count $E_{j_1j_2j_3}$ of the (j_1, j_2, j_3)-th cell can be computed, for $j_1, j_2, j_3 = 0, 1$, and the joint distribution of Y_1, Y_2 and Y_3 can be estimated by $\{E_{j_1j_2j_3}/n_1, j_1, j_2, j_3 = 0, 1\}$.

It should be noticed that if the final model is the saturated model (2), then $E_{j_1j_2j_3} = O_{j_1j_2j_3}$ for all j_1, j_2 and j_3. In such a case, $E_{j_1j_2j_3}/n_1$ is the ordinary relative frequency of the (j_1, j_2, j_3)-th cell. When the components of $\mathbf{Y}$ have some association structure, the selected log-linear model (e.g., the model (Y_1Y_2, Y_1Y_3, Y_2Y_3)) would be simpler than the saturated model to reflect this structure. In such cases, the variability of $E_{j_1j_2j_3}/n_1$ computed from the selected log-linear model is smaller than the variability of the corresponding relative frequency $O_{j_1j_2j_3}/n_1$, because the selected log-linear model has less parameters than the saturated model. In other words, $E_{j_1j_2j_3}/n_1$ is "smoother" than $O_{j_1j_2j_3}/n_1$. In that sense, the log-linear modeling is a smoothing process, and the degree of smoothing depends on the association structure of the components of $\mathbf{Y}$. Since both the log-linear estimator $E_{j_1j_2j_3}/n_1$ and the relative frequency estimator $O_{j_1j_2j_3}/n_1$ are unbiased in cases when the selected model holds, the log-linear estimator $\{E_{j_1j_2j_3}/n_1, j_1, j_2, j_3 = 0, 1\}$ would provide a better estimator for the joint distribution of $\mathbf{Y}$ in such cases.

The log-linear modeling approach described above is easy to use, since almost all existing statistical software packages have functions to do this analysis. For instance, the function glm() in S-Plus or R can be used for this purpose.

3 A Distribution-Free Multivariate CUSUM for Phase II SPC

In this section, we propose a phase II, distribution-free, multivariate CUSUM for detecting *location* shifts in the original measurement distribution $F(\mathbf{x})$. Note that $F(\mathbf{x})$ has a shift in a location parameter vector (e.g., the mean vector) if and only if it has the same shift in another location parameter vector (e.g., the median vector). For this reason and for reasons given in Section 2 (cf, Remark 2.1), we can focus on detecting shifts in the median vector $(m_1, m_2, \ldots, m_p)'$ of $F(\mathbf{x})$. To this end, it has been explained in Section 2 that $\mathbf{Y}(i) = (Y_1(i), Y_2(i), \ldots, Y_p(i))'$ is appropriate to use, because any shift in $(m_1, m_2, \ldots, m_p)'$ would alter the in-control distribution of $\mathbf{Y}(i)$. Thus, such a shift can be detected by a procedure designed for detecting shifts in the joint distribution of $\mathbf{Y}(i)$.

Assume that the in-control joint distribution of $\mathbf{Y}(i)$ is $\{f^{(0)}_{j_1,\ldots,j_p}, j_1,\ldots,j_p = 0,1\}$, which can be estimated by the log-linear modeling procedure discussed in Section 2. For instance, when $p = 3$, $f^{(0)}_{j_1,j_2,j_3}$ is estimated by $E_{j_1j_2j_3}/n_1$, where $E_{j_1j_2j_3}/n_1$ is obtained from the selected log-linear model. In the statistical literature, the Pearson's χ^2 test is well-known for testing whether or not the distribution of a random vector equals a given distribution. Let

$$g_{j_1,\ldots,j_p}(i) = I(Y_1(i) = j_1,\ldots,Y_p(i) = j_p),$$

where $j_1,\ldots,j_p = 0$ or 1. Then $\sum_{i=1}^n g_{j_1,\ldots,j_p}(i)$ is the observed count of the $(j_1,\ldots,j_p)$-th cell as of time point n, and $nf^{(0)}_{j_1,\ldots,j_p}$ is the corresponding expected cell count. The conventional Pearson's χ^2 statistic is defined by

$$\sum_{j_1,\ldots,j_p=0,1} \frac{\left(\sum_{i=1}^n g_{j_1,\ldots,j_p}(i) - nf^{(0)}_{j_1,\ldots,j_p}\right)^2}{nf^{(0)}_{j_1,\ldots,j_p}},$$

which measures the discrepancy between the observed and expected cell counts.

Then, a natural idea to detect shifts in a location parameter vector of the measurement distribution of a process is to compare the observed value of the above Pearson's χ^2 statistic with a threshold value. If the former is larger, then a shift is signaled. In most existing CUSUM procedures, however, an "allowance" constant k is often used for repeatedly restarting the CUSUM procedures when there is no evidence of shifts, so that the CUSUMs can react to an incoming shift promptly. The size of k often depends on the magnitude of a target shift. If the target shift is small, then k should be chosen small as well, and vice versa. It has been well demonstrated in the literature (e.g., Hawkins and Olwell 1998) that inclusion of this constant would improve the performance of the CUSUM procedures, especially when the target shift is small. By combining the Pearson's χ^2 test and the idea of using the "allowance" constant k, the following CUSUM procedure, which has a similar form to that of the procedure by Crosier (1988), is suggested for detecting possible shifts in a location parameter vector of the measurement distribution of a process:

- When $C_n \leq k$, let

$$\begin{cases} \mathbf{S}_n^{\text{obs}} &= \mathbf{0} \\ \mathbf{S}_n^{\text{exp}} &= \mathbf{0}, \end{cases} \tag{3}$$

- When $C_n > k$, let

$$\begin{cases} \mathbf{S}_n^{\text{obs}} &= (\mathbf{S}_{n-1}^{\text{obs}} + \mathbf{g}(n))(C_n - k)/C_n \\ \mathbf{S}_n^{\text{exp}} &= (\mathbf{S}_{n-1}^{\text{exp}} + \mathbf{f}^{(\mathbf{0})})(C_n - k)/C_n, \end{cases} \tag{4}$$

where

$$\begin{aligned}C_n &= [(\mathbf{S}_{n-1}^{\text{obs}} - \mathbf{S}_{n-1}^{\text{exp}}) + (\mathbf{g}(n) - \mathbf{f}^{(\mathbf{0})})]'[diag(\mathbf{S}_{n-1}^{\text{exp}} + \mathbf{f}^{(\mathbf{0})})]^{-1} \\ &\quad [(\mathbf{S}_{n-1}^{\text{obs}} - \mathbf{S}_{n-1}^{\text{exp}}) + (\mathbf{g}(n) - \mathbf{f}^{(\mathbf{0})})],\end{aligned}$$

$\mathbf{S}_0^{\text{obs}} = \mathbf{S}_0^{\text{exp}} = \mathbf{0}$, $\mathbf{g}(n)$ is a vector of all $g_{j_1,\ldots,j_p}(n)$ values, for $j_1,\ldots,j_p = 0,1$; $\mathbf{f}^{(\mathbf{0})}$ is a vector of all $f_{j_1,\ldots,j_p}^{(0)}$ values, for $j_1,\ldots,j_p = 0,1$; $k \geq 0$ is the "allowance" constant; $diag(\mathbf{a})$ denotes a diagonal matrix with its diagonal elements equal to the corresponding components of the vector $\mathbf{a}$; and the superscripts "obs" and "exp" denote observed and expected counts, respectively. Define

$$u_n = (\mathbf{S}_n^{\text{obs}} - \mathbf{S}_n^{\text{exp}})'[diag(\mathbf{S}_n^{\text{exp}})]^{-1}(\mathbf{S}_n^{\text{obs}} - \mathbf{S}_n^{\text{exp}}). \tag{5}$$

Then

$$u_n > h \tag{6}$$

signals a shift, where $h > 0$ is a control limit.

It can be checked that u_n equals the conventional Pearson's χ^2 statistic when $k = 0$, and $u_n = \max(0, C_n - k)$ when $k \neq 0$. The latter conclusion can be verified by some similar arguments to those in Appendix C of Qiu and Hawkins (2001). Therefore, the constant k is indeed used for repeatedly restarting the CUSUM when there is no evidence of shifts, such that the CUSUM can react to a real shift promptly.

The cusum procedure (3)-(6), together with the log-linear modeling procedure discussed in Section 2, can detect any shift in a location parameter vector of the multivariate measurement distribution, without assuming the in-control measurement distribution to be known and Gaussian. Its two parameters h and k can be easily determined by the following two algorithms.

For a given in-control ARL value ARL_0 and a given k, the value of h in procedure (3)-(6) can be searched in a range $[0, U_h]$ by the following algorithm, where U_h is an upper bound satisfying the condition that the in-control ARL of the procedure is larger than ARL_0 when $h = U_h$.

Algorithm I (searching for h when k is given)

1. In the i-th iteration, h is searched in the range $[L_h^{(i)}, U_h^{(i)}]$. When $i = 1$, $L_h^{(1)} = 0$ and $U_h^{(1)} = U_h$.

2. A series of random vectors from the multinomial distribution with probability parameters $\{f^{(0)}_{j_1,\ldots,j_p}, j_1,\ldots,j_p = 0,1\}$ are generated by a random number generator.

3. This series of random vectors are used in the place of $\mathbf{g}(n)$ in (4) and the run length distribution is obtained by running the procedure (3)-(6) with $h = h^{(i)} := (L_h^{(i)} + U_h^{(i)})/2$ a number of times (10,000 times in all numerical examples in Section 4). The in-control ARL value $ARL_0^{(i)}$ is then computed by averaging all the run lengths obtained.

4. If $|ARL_0^{(i)} - ARL_0| < \varepsilon_1$, where $\varepsilon_1 > 0$ is a pre-specified threshold value, then the algorithm stops, and the searched value of h is $h^{(i)}$. Otherwise, define

$$L_h^{(i+1)} = h^{(i)} \text{ and } U_h^{(i+1)} = U_h^{(i)}, \text{ if } ARL_0^{(i)} < ARL_0;$$
$$L_h^{(i+1)} = L_h^{(i)} \text{ and } U_h^{(i+1)} = h^{(i)}, \text{ if } ARL_0^{(i)} > ARL_0;$$
$$\text{and } h^{(i+1)} := (L_h^{(i+1)} + U_h^{(i+1)})/2.$$

5. If $|h^{(i+1)} - h^{(i)}| < \varepsilon_2$, where $\varepsilon_2 > 0$ is another pre-specified threshold value, then the algorithm stops, and the searched value of h is $h^{(i)}$. In such a case, a message should be printed, to remind the user of the actual in-control ARL value. If $|h^{(i+1)} - h^{(i)}| \geq \varepsilon_2$, then the algorithm executes the next iteration.

Based on our experience, the above algorithm usually stops at the fourth step. But occasionally it can happen that it stops at the fifth step, especially when ε_1 is chosen relatively small and ε_2 is chosen relatively large. In such cases, users are reminded by the algorithm that the assumed in-control ARL value is not reached within a specified range by procedure (3)-(6) using the searched value of h; its actual in-control ARL value is also printed.

If we have a target shift in the location parameter vector of the measurement distribution, then for a given in-control ARL value ARL_0, the optimal value of k of procedure (3)-(6) can be searched in a range $[0, U_k]$ by the following algorithm.

Algorithm II (searching for the optimal value of k for a target shift)

1. In the i-th iteration, k is searched in the range $[L_k^{(i)}, U_k^{(i)}]$. When $i = 1$, $L_k^{(1)} = 0$ and $U_k^{(1)} = U_k$. Divide $[L_k^{(i)}, U_k^{(i)}]$ into m equally spaced subintervals, where m is pre-specified (e.g.,

$m = 10$). Then k is searched among all the end points $\{k_j^{(i)} = L_k^{(i)} + (U_k^{(i)} - L_k^{(i)}) * j/m,\ j = 0, 1, \ldots, m\}$ of these subintervals.

2. When $k = k_j^{(i)}$, for any $j = 0, 1, \ldots, m$, search for the corresponding h value by the Algorithm I, such that the in-control ARL equals ARL_0.

3. For the target shift, compute the out-of-control distribution of $\mathbf{Y}(i)$, denoted by $\{f_{j_1,\ldots,j_p}^{(1)}, j_1, \ldots, j_p = 0, 1\}$, by the log-linear modeling procedure discussed in Section 2.

4. Generate a series of random vectors from the multinomial distribution with probability parameters $\{f_{j_1,\ldots,j_p}^{(1)}, j_1, \ldots, j_p = 0, 1\}$, and this series of random vectors are used in the place of $\mathbf{g}(n)$ in (4).

5. For each j, compute the out-of-control ARL, denoted as $ARL_{1,j}^{(i)}$, by running the procedure (3)-(6) with $k = k_j^{(i)}$ a number of times. Suppose that the minimizer of $\{ARL_{1,j}^{(i)}, j = 0, 1, \ldots, m\}$ is $J^{(i)}$.

6. If $(U_k^{(i)} - L_k^{(i)})/m < \varepsilon_3$, where $\varepsilon_3 > 0$ is a pre-specified threshold value, then the algorithm stops, and the searched value of k is $k_{J^{(i)}}^{(i)}$. Otherwise, let $L_k^{(i+1)} = \max(0, k_{J^{(i)}-1}^{(i)})$ and $U_k^{(i+1)} = \min(k_{J^{(i)}+1}^{(i)}, U_k)$; the algorithm executes the next iteration.

Remark 3.1 The "allowance" constant k should be chosen from the interval $[0, \max_{j_1,\ldots,j_p=0,1}(1 - f_{j_1,\ldots,j_p}^{(0)})/f_{j_1,\ldots,j_p}^{(0)}]$. Otherwise, the CUSUM procedure (3)-(6) will restart at each time point when the process is in-control, and consequently the specified in-control ARL property can not be achieved.

Based on Remark 3.1, the upper bound U_k can be chosen to be $\max_{j_1,\ldots,j_p=0,1}(1-f_{j_1,\ldots,j_p}^{(0)})/f_{j_1,\ldots,j_p}^{(0)}$. Selection of U_h should not be difficult. For a given in-control ARL value ARL_0 and a given k, we could try a large number (e.g., 50 or 100) for U_h, and then run procedure (3)-(6) to make sure that its in-control ARL is larger than ARL_0. Since both Algorithms I and II converge reasonably fast, accurate selection of U_k and U_h is not essential to their convergence speed, which makes the selection of U_k and U_h much easier. That is, the two upper bounds can be chosen relatively large, without sacrificing much convergence speed of the two algorithms.

The values of $\varepsilon_1, \varepsilon_2$ and ε_3 are related to the accuracy requirements for the solutions. For example, if we require that the actual in-control ARL value equals with high probability the assumed in-control ARL value up to the third digit after the decimal point, then we can choose $\varepsilon_1 = 0.5 \times 10^{-3}$.

4 Numerical Examples

In this section, we present some numerical examples regarding the numerical performance of the procedures introduced in the previous two sections. The examples are organized in three parts. Those related to the log-linear modeling procedure for estimating the in-control measurement distribution are discussed in Section 4.1. The performance of the CUSUM procedure (3)-(6) for Phase II SPC is studied in Section 4.2. In Section 4.3, we apply our method to a real dataset.

4.1 Performance of the Log-Linear Modeling Procedure

Before we can apply the Phase II procedure (3)-(6) to a specific problem, the in-control joint distribution $\{f^{(0)}_{j_1,\ldots,j_p}, j_1, \ldots, j_p = 0, 1\}$ of $\mathbf{Y}(i)$ needs to be estimated from an in-control dataset. To this end, a traditional method is to use relative frequencies (RF) computed from the in-control dataset for estimating $f^{(0)}$, which is equivalent to estimating the joint distribution of $\mathbf{Y}(i)$ based on the saturated model (2). As discussed in Section 2, the variability of the RF estimators is relatively large, mainly due to the fact that the RF method ignores the association structure of the measurement components completely, which is demonstrated by the following example.

Suppose that $p = 3$, and the measurement vector $\mathbf{X}(i) = (X_1(i), X_2(i), X_3(i))'$ has one of the following two in-control distributions.

<u>Case I</u>	<u>Case II</u>
$X_1 \sim \chi^2(1)$	$X_1 \sim N(0,1)$
$X_2 \sim \chi^2(1)$	$X_2 \sim \chi^2(3)$
$X_3 \sim \chi^2(1)$	$X_3 = X_1 + \xi,\ \xi \sim N(0,1)$
X_1, X_2 and X_3 are independent	X_1, X_2 and ξ are independent

It can be seen that the three measurement components are independent of each other in Case I; X_1 and X_2 are independent, X_2 and X_3 are independent, and X_1 and X_3 are associated in Case II.

Using the notation of the log-linear (LL) modeling introduced in Section 2, Case I can be described by the model (Y_1, Y_2, Y_3), and Case II can be described by the model (Y_2, Y_1Y_3).

Three sample sizes $n_1 = 100$, 1000 and 10000 are considered. By the RF method, the averaged estimate of the in-control distribution of $\mathbf{Y}(i)$ and its standard error based on 1000 replications are presented in Table 1. As a comparison, the corresponding results by the LL procedure are presented in the same table. In the LL modeling process, some rules outlined in Section 2 are followed, and the significance level is chosen to be 0.05.

From Table 1, it can be seen that estimators by the LL procedure are much more accurate than estimators by the RF method, in terms of the point estimators and the confidence intervals as well. Comparing Case I with Case II, the association among Y_1, Y_2 and Y_3 is stronger in Case II than in Case I (the measurement components are actually independent of each other in Case I). From Table 1, the benefit to use the LL procedure is a little bit smaller in Case II in terms of the standard errors, which is reasonable because relatively less association structure among the components of $\mathbf{Y}$ is ignored by the RF method in Case II than in Case I. Comparing the cases with different sample sizes, it can be seen that: (1) both methods provide more accurate estimators when n_1 is larger, (2) the LL procedure gives quite accurate estimators even when n_1 is relatively small, (3) the estimators by the RF method have relatively large variabilities when n_1 is small, and (4) consequently the improvement by using the LL procedure is more significant when n_1 is smaller.

4.2 Performance of the Distribution-Free Procedure

In this part, we first demonstrate with numerical examples that the conventional multivariate SPC procedures based on normal distribution assumption may not be appropriate to use in cases when the normal distribution assumption is violated. Two such existing procedures are considered here. One is the multivariate CUSUM procedure suggested by Crosier (1988) (cf. equations (4) and (5) in Crosier (1988)), and the other one is the multivariate EMWA control chart discussed by Stoumbos and Sullivan (2002) (cf., equations (2)–(4) in Stoumbos and Sullivan (2002)).

Suppose that the process measurement is three-dimensional and its assumed in-control distribution is $N(\mathbf{0}, I_3)$. In Crosier's procedure, the control limit h and the constant k are chosen to be

Table 1: This table presents the true cell probabilities, their averaged estimates, and the corresponding standard errors (in parentheses), based on 1000 replications, by the relative frequency (RF) method and the log-linear (LL) modeling method, respectively.

			$Y_3 = 0$			
			$Y_2 = 0$		$Y_2 = 1$	
			$Y_1 = 0$	$Y_1 = 1$	$Y_1 = 0$	$Y_1 = 1$
Case I	True		.1250	.1250	.1250	.1250
	RF	n_1=100	.1240 (.0008)	.1257 (.0008)	.1247 (.0008)	.1256 (.0008)
		n_1=1000	.1253 (.0003)	.1249 (.0003)	.1250 (.0002)	.1248 (.0002)
		n_1=10000	.1250 (.0001)	.1250 (.0001)	.1249 (.0001)	.1250 (.0001)
	LL	n_1=100	.1250 (.0000)	.1250 (.0000)	.1250 (.0000)	.1250 (.0000)
		n_1=1000	.1250 (.0000)	.1250 (.0000)	.1250 (.0000)	.1250 (.0000)
		n_1=10000	.1250 (.0000)	.1250 (.0000)	.1250 (.0001)	.1250 (.0000)
Case II	True		.1875	.0625	.1875	.0625
	RF	n_1=100	.1864 (.0008)	.0632 (.0006)	.1883 (.0008)	.0622 (.0006)
		n_1=1000	.1871 (.0003)	.0623 (.0002)	.1878 (.0003)	.0628 (.0002)
		n_1=10000	.1874 (.0001)	.0625 (.0001)	.1876 (.0001)	.0625 (.0001)
	LL	n_1=100	.1873 (.0003)	.0627 (.0003)	.1873 (.0003)	.0627 (.0003)
		n_1=1000	.1875 (.0001)	.0625 (.0001)	.1875 (.0001)	.0625 (.0001)
		n_1=10000	.1875 (.0000)	.0625 (.0000)	.1875 (.0000)	.0625 (.0000)

			$Y_3 = 1$			
			$Y_2 = 0$		$Y_2 = 1$	
			$Y_1 = 0$	$Y_1 = 1$	$Y_1 = 0$	$Y_1 = 1$
Case I	True		.1250	.1250	.1250	.1250
	RF	n_1=100	.1255 (.0008)	.1249 (.0008)	.1259 (.0008)	.1238 (.0008)
		n_1=1000	.1251 (.0003)	.1246 (.0002)	.1246 (.0003)	.1256 (.0003)
		n_1=10000	.1250 (.0001)	.1250 (.0001)	.1251 (.0001)	.1250 (.0001)
	LL	n_1=100	.1250 (.0000)	.1250 (.0000)	.1250 (.0000)	.1250 (.0000)
		n_1=1000	.1250 (.0000)	.1250 (.0000)	.1250 (.0000)	.1250 (.0000)
		n_1=10000	.1250 (.0000)	.1250 (.0000)	.1250 (.0000)	.1250 (.0000)
Case II	True		.0625	.1875	.0625	.1875
	RF	n_1=100	.0629 (.0006)	.1875 (.0008)	.0625 (.0006)	.1871 (.0008)
		n_1=1000	.0624 (.0002)	.1881 (.0003)	.0626 (.0002)	.1868 (.0003)
		n_1=10000	.0625 (.0001)	.1875 (.0001)	.0624 (.0001)	.1875 (.0001)
	LL	n_1=100	.0627 (.0003)	.1873 (.0003)	.0627 (.0003)	.1873 (.0003)
		n_1=1000	.0625 (.0001)	.1875 (.0001)	.0625 (.0001)	.1875 (.0001)
		n_1=10000	.0625 (.0000)	.1875 (.0000)	.0625 (.0000)	.1875 (.0000)

3.786 and 1.0, respectively, so that its in-control ARL equals 200 when the in-control measurement distribution is assumed to be $N(\mathbf{0}, I_3)$. In the multivariate EMWA procedure, we consider two values for the parameter r: $r = 0.05$ and $r = 0.2$. The first r value is in the recommended range by Stoumbos and Sullivan (2002), so that the EMWA procedure would be robust to the normal distribution assumption. The control limits for these two r values are chosen to be 9.603 and 11.956, respectively, so that the in-control ARL of the procedure equals 200 in both cases. In computing the in-control and out-of-control ARL values in this section, we assume that potential shifts would not start until the 100th phase II observation, for reasons explained in the second paragraph to follow. Now, suppose that the three measurement components are actually independent of each other, and the distribution of each component is a standardized version of the central χ^2 distribution with degrees of freedom m (i.e., each component is standardized to have mean 0 and standard deviation 1), then the true in-control ARL values of the related procedures are presented in Figure 1, when $m = 1, 5, 10, 20$ and 50. The long-dashed line in the top of this plot denotes the assumed in-control ARL value 200. It can be seen from the plot that: (i) when m is small, the true in-control ARL values of all three procedures are quite different from the assumed in-control ARL value, (ii) when m increases, their true in-control ARL values are closer to the assumed in-control ARL value because the true measurement distribution is closer to a normal distribution in such cases, and (iii) when r is smaller, the multivariate EMWA is more robust to the normal distribution assumption. Later, we will demonstrate in Figure 2 that, when r is chosen smaller in the conventional multivariate EMWA, its ability to detect large shifts is also weaker.

Next, we investigate the numerical performance of the CUSUM procedure (3)-(6) for detecting shifts in Phase II SPC. By using this procedure, the in-control distribution of $\mathbf{Y}$ is not assumed known, and it needs to be estimated from an in-control dataset. When $p = 3$ and the in-control distribution of $\mathbf{X}$ is the normalized version of the one specified in Case I above (i.e., each measurement component is normalized to have mean 0 and standard deviation 1), we randomly generate 100 such in-control datasets, each of which has sample size 100. Then, from each in-control dataset, the estimated in-control distribution of $\mathbf{Y}$ is computed, using both the LL and RF methods. Based on each estimated in-control distribution of $\mathbf{Y}$, the control limit value h in the procedure (3)-(6) is searched by the Algorithm I discussed in Section 3, when $k = 1.0$ and the in-control ARL is fixed at 200. In Algorithm I, the parameters are chosen to be $U_h = 30, \varepsilon_1 = .01$, and $\varepsilon_2 = 10^{-5}$; the search is based on 10000 replications. Then, for each of the LL and RF methods, the in-control

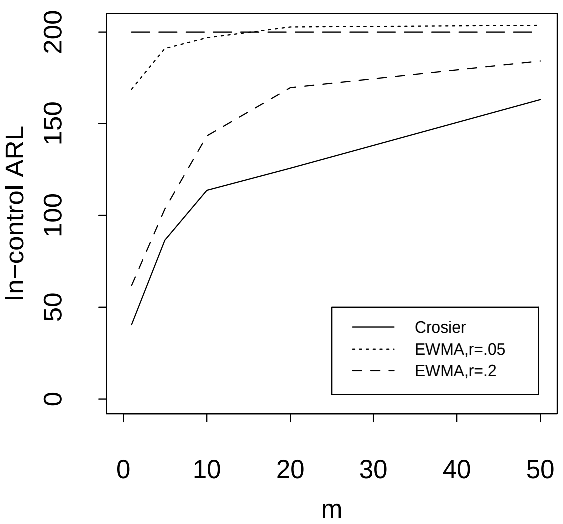

Figure 1: The solid, dotted, and short-dashed lines denote the true in-control ARL values of the Crosier's procedure, the multivariate EMWA procedure when $r = 0.05$, and the multivariate EMWA procedure when $r = 0.1$, respectively. In this example, the process measurement is three-dimensional, the measurement components are independent of each other and each component has an actual normalized χ^2 distribution with degrees of freedom m. The long-dashed line at the top denotes the assumed in-control ARL value 200.

dataset and the corresponding control limit value giving the median actual in-control ARL value is chosen for that method and used in all phase II examples described below.

Suppose that there is a shift of size $(a, 0, 0)$ in the median vector of the process measurements, the shift starts at the 100th phase II time point, and a changes its value between 0 and 0.4 with step 0.04. The specific starting time of the shift (i.e., the 100th time point) is used as an approximation to the "steady-state start", after which the distribution of the CUSUM statistic u_n approaches some "steady-state distribution" that does not depend on n (cf. e.g., Hawkins and Olwell 1998, Chapter 3). By the way, we also tried cases when the starting time equals 0; the results are similar and thus they are omitted here. Then, the out-of-control ARL of procedure (3)-(6) based on the LL estimate of the in-control distribution of $\mathbf{Y}$, which is refered to as the LL procedure below, is presented in Figure 2(a) by the solid curve. The corresponding results of procedure (3)-(6) based on the RF estimate of the in-control distribution of $\mathbf{Y}$, which is refered to as the RF procedure below, is presented in the same plot by the dotted curve.

In Section 1 and in the example of Figure 1 above, we already pointed out that the conventional CUSUM procedures based on the Gaussian distribution assumption, such as the one suggested by Crosier (1988), and the conventional EWMA procedures might not be appropriate to use in cases when the normal distribution assumption is violated. To further demonstrate this, the CUSUM procedure suggested by Crosier is applied to Case I in the following two ways. First, the procedure is used in a conventional manner that the in-control distribution of $\mathbf{X}$ is assumed to be $N(\mathbf{0}, I_3)$, as in Figure 1. Second, the procedure is used based on the assumption that the in-control distribution of $\mathbf{X}$ is $N(\mu, I_3)$, but μ should be estimated from an in-control data of size 100. The second way is the so-called "self-starting" version of the Crosier's procedure, in which the in-control parameters of the measurement distribution are estimated from an in-control data. For simplicity, we only estimate the in-control mean of the measurement distribution here, with the in-control covariance matrix assumed known. The two versions of the Crosier's procedure are denoted as CR and CRSS below, where CR denotes the conventional Crosier's procedure and CRSS denotes its self-starting version. ARL results of these two versions of the Crosier's procedure are presented in Figure 2(a) by the dashed and dot-dashed curves, respectively.

From Figure 2(a), it can be seen that among the four procedures LL, RF, CR, and CRSS, only the procedure LL has its actual in-control ARL equal to the assumed in-control ARL. The actual in-control ARL values of the remaining three procedures are far away from the assumed in-control ARL, which is 200. Therefore, they may *not* be appropriate to use in this case. For the procedure RF, this happens because the RF estimate of the in-control distribution of $\mathbf{Y}$ has much larger variability than the log-linear estimator, as explained at the end of Section 2 and demonstrated in Table 1. Consequently, the RF estimator is in general quite different from the true in-control distribution of $\mathbf{Y}$. Therefore, the RF procedure often detects a shift, even when the process is actually in-control. For the procedure CR, deviation of the actual measurement distribution from the assumed Gaussian distribution is a major cause of its small in-control ARL, because it treats the distributional deviation as a shift in the in-control measurement distribution. For the self-starting version of the CR procedure, its small in-control ARL can be explained by both reasons mentioned above.

Next, we would like to explain why the largest value of a is chosen to be 0.4 in this example. The normalized version of the $\chi^2(1)$ distribution, which is the distribution of the first measurement component, has a median of -0.3855, and its support has a lower bound of -0.7071. So, an upward shift of size 0.4 would make the first measurement component consistently larger than its in-control median. Consequently, the value of Y_1 is actually a constant 1, after the shift. Larger upward shifts in the first measurement component will not alter the distribution of $\mathbf{Y}$, and therefore will not change the out-of-control ARL of the procedure (3)-(6).

Figure 2(a) also shows that the out-of-control ARL of procedure LL is small, compared to procedures CR and CRSS, when the shift is reasonably large. Considering the fact that the in-control ARL of procedure LL is much larger than the in-control ARLs of procedures CR and CRSS, this is an endorsement of the former procedure. To further investigate this issue, next, we consider the larger shift $(a, .4, .4)$, where a changes its value between 0 and .4 with step .04. The corresponding out-of-control ARL values of the four procedures are presented in Figure 2(b). It can be seen that procedure LL is consistently better than procedures CR and CRSS.

The corresponding results of the EWMA procedures considered in Figure 1 are presented in Figure 2(c)–(d) by the dotted and dashed curves, respectively, for cases when $r = .05$ and $r = .2$. For convenience of comparison, the out-of-control ARL values of procedure LL are presented in this plot again by the solid curve. It can be seen that: (i) when r is chosen smaller (i.e., $r = 0.05$ in this example), the EWMA procedure is more robust to the normal distribution assumption, as observed in Figure 1, because its actual in-control ARL value is closer to the nominal value 200, (ii) in such cases, its ability to detect possible shifts is also weaker, compared to EWMA procedures when r is chosen larger, and (iii) procedure LL performs better when the process is in-control and when the process becomes out-of-control with a quite large shift (e.g., shifts $(a, 0, 0)$ when $a > 0.34$ in plot (c) and shifts $(a, .4, .4)$ for all a values in plot (d)).

Next, we compare procedure LL with the antirank-based CUSUM procedure suggested by Qiu and Hawkins (QH, 2000), which does not depend on the Gaussian distribution assumption. We still consider the shift $(a, 0, 0)$, with a changing its value between 0 and 0.4 with step 0.04. Since this shift is upward, the CUSUM procedure based on the third antirank is prefered here, among all CUSUM procedures based on a single antirank. To use this procedure, the in-control distribution of the third antirank should be specified. Qiu and Hawkins (2000) assume that this in-control distribution is known. In applications, it needs to be estimated from an in-control dataset. When the true in-control distribution of the third antirank is assumed known, the control limit value of the procedure is searched to be 4.300 such that its in-control ARL equals 200 when $k = 1$. Results under this condition are labelled by QHTR, where TR denotes "true" (i.e., the true in-control distribution is assumed known). If the in-control distribution is estimated by the relative frequencies computed from the same in-control dataset used above in estimating the in-control distribution of $\mathbf{Y}$, then

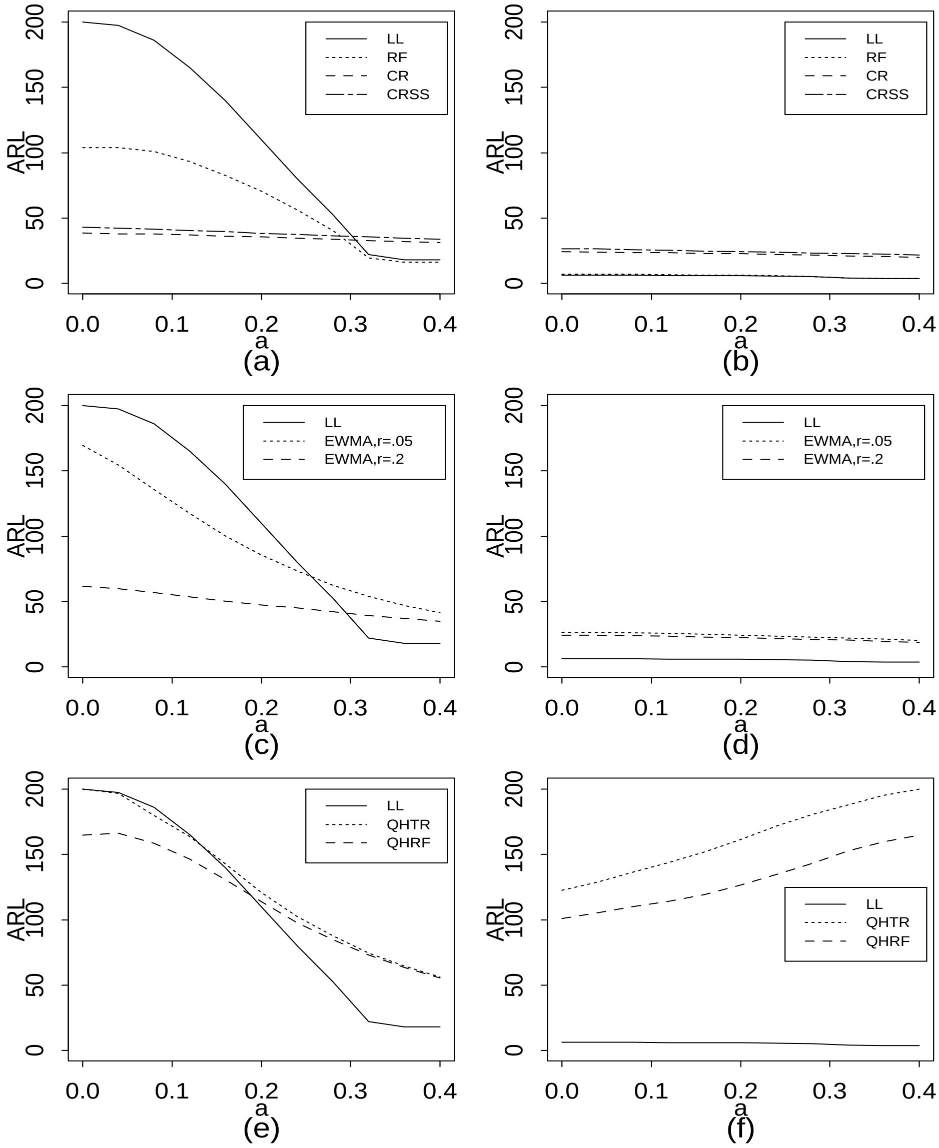

Figure 2: (a) This plot presents ARL values of procedures LL (solid curve), RF (dotted curve), CR (dashed curve), and CRSS (dot-dashed curve). The shift is assumed to be $(a, 0, 0)$ in the median vector of the process measurements starting at the 100th time point. (b) Corresponding results of plot (a) when the shift is $(a, .4, .4)$. (c) ARL values of procedures LL (solid curve), EWMA with $r = .05$ (dotted curve), and EWMA with $r = .2$ (dashed curve), when there is a shift $(a, 0, 0)$ in the median vector of the process measurements starting at the 100th time point. (d) Corresponding results of plot (c) when the shift is $(a, .4, .4)$. (e) ARL values of procedures LL (solid curve), QHTR (dotted curve), and QHRF (dashed curve), when there is a shift $(a, 0, 0)$ in the median vector of the process measurements starting at the 100th time point. (f) Corresponding results of plot (e) when the shift is $(a, .4, .4)$.

the control limit value of the QH procedure is searched to be 4.430, to reach the in-control ARL value 200 when $k = 1$. Results in this case are labelled by QHRF, where RF denotes "relative frequencies". In these two cases, the ARL values of the QH procedure are presented in Figure 2(e) by the dotted and dashed curves, respectively. For convenience in comparison, the ARL values of procedure LL are also presented in this plot by the solid curve. It can be seen that, for the self-starting version of the QH procedure (i.e., the procedure QHRF), its actual in-control ARL is also different from the assumed in-control ARL. But the difference is not so large, compared to procedures RF, EWMA when $r = 0.2$, CR, and CRSS discussed above. Procedure QHTR is indeed appropriate to use, since its actual in-control ARL is about the same as the assumed in-control ARL. But, in most cases, the current procedure LL outperforms procedure QHTR, and the difference is quite large when the shift gets large.

The corresponding results of Figure 2(e) when the shift is $(a, .4, .4)$ are shown in Figure 2(f), where a changes its value between 0 and .4 with step .04, as before. This shift is closer to the "equal-shift" direction, in which all components of the shift are the same, when a is closer to .4. It is expected that performance of the QH procedure would get worse when a is closer to .4, because the out-of-control distribution of the third antirank is closer to its in-control distribution in such cases, which is demonstrated in Figure 2(f). From the plot, it can be seen that the current procedure LL performs much better than both versions of the QH procedure in this case.

The optimal values of the two parameters h and k of the procedure (3)-(6) can be searched easily by the algorithms described in Section 3. Next, we consider the following shifts in the median vector of $\mathbf{X}(i)$ occurred at the time point "start": (-1,0,0), (-2,0,0), (-2,-1,0), (-2,-2,0), (-2,-2,-1) and (-2,-2,-2). These shifts are ordered from the smallest to largest in magnitudes. The corresponding out-of-control joint distributions of $\mathbf{Y}(i)$, denoted as $\{f^{(1)}_{j_1j_2j_3}, j_1, j_2, j_3 = 0, 1\}$, are displayed in Table 2. The Euclidean distance between the out-of-control and the in-control (the latter can be found from Table 1) distributions of $\mathbf{Y}(i)$ is shown in the last column labeled by Q.

From the construction of procedure (3)-(6), the optimal values of h and k depend on the in-control ARL, and the in-control and out-of-control distributions of $\mathbf{Y}(i)$ only. In that sense, procedure (3)-(6) is distribution-free. For each shift, the optimal values of h and k are searched by Algorithms I and II discussed in Section 3 with the related parameters chosen to be: $U_h = 30$, $U_k = \max_{j_1,\ldots,j_p=0,1}(1 - f^{(0)}_{j_1,\ldots,j_p})/f^{(0)}_{j_1,\ldots,j_p} = 7$, $\varepsilon_1 = .01$, $\varepsilon_2 = 10^{-5}$ and $\varepsilon_3 = .001$. The searched

Table 2: The shifts in the median vector of $\mathbf{X}(i)$ and the corresponding out-of-control joint distribution of $\mathbf{Y}(i)$ are presented in this table. $Q = \sqrt{\sum_{j_1j_2j_3=0,1}(f^{(1)}_{j_1j_2j_3} - f^{(0)}_{j_1j_2j_3})^2}$ denotes the Euclidean distance between $f^{(1)}$ and $f^{(0)}$.

shifts in median	$f^{(1)}_{000}$	$f^{(1)}_{100}$	$f^{(1)}_{010}$	$f^{(1)}_{110}$	$f^{(1)}_{001}$	$f^{(1)}_{101}$	$f^{(1)}_{011}$	$f^{(1)}_{111}$	Q
(-1,0,0)	.2072	.0429	.2070	.0429	.2071	.0428	.2072	.0429	.2323
(-2,0,0)	.2325	.0175	.2325	.0175	.2325	.0174	.2326	.0175	.3041
(-2,-1,0)	.3852	.0290	.0797	.0060	.3854	.0289	.0797	.0060	.4317
(-2,-2,0)	.4325	.0326	.0325	.0025	.4325	.0325	.0325	.0024	.5033
(-2,-2,-1)	.7167	.0539	.0540	.0041	.1483	.0111	.0111	.0008	.6456
(-2,-2,-2)	.8045	.0605	.0605	.0046	.0605	.0045	.0045	.0003	.7303

Table 3: The optimal values of h and k, the corresponding out-of-control ARLs, and their standard errors (in parentheses) for various shifts occured at two different starting times.

	start=0			start=100		
shifts in mean	h	k	arl (se)	h	k	arl (se)
(-1,0,0)	9.1268	.004	6.6309 (.0597)	9.6364	.121	24.9056 (.2619)
(-2,0,0)	9.1268	.004	4.7357 (.0310)	11.2941	.351	16.8808 (.1160)
(-2,-1,0)	9.1268	.004	3.6631 (.0205)	11.7540	.546	10.4746 (.0674)
(-2,-2,0)	9.1268	.004	3.1380 (.0145)	11.9642	.867	8.4220 (.0512)
(-2,-2,-1)	9.1878	.003	2.7704 (.0117)	11.9577	.967	6.2420 (.0376)
(-2,-2,-2)	9.1878	.003	2.5063 (.0091)	11.5997	1.458	5.2300 (.0319)

results for two different starting times are displayed in Table 3. One starting time is start=0 and the other one is start=100. As demonstrated by several authors (e.g., Qiu and Hawkins 2001), h and k values should be chosen differently for shifts occured at the initial time point (start=0) and shifts occured at the "steady-state start". We also performed simulations in the case when start=200 and found that results in that case are similar to results when start=100. Therefore, the case when start=100 might be a good approximation to the "steady-state start" already, and the results when start=200 are omitted here. By checking the results in Table 3, it can be seen that the value of k should be chosen small for shifts occured at the initial time point. Its value and the corresponding value of h do not depend on the magnitude of the shift much in such cases. For shifts occured at the "steady-state start", k should be chosen larger for larger shifts, which is true for most CUSUM procedures. Although the mathematical relationship between Q and the optimal value of k is still unknown, the computer algorithms in Section 3 can search for the optimal value of k easily for any target shift.

4.3 Application to the Aluminum Smelter Data

In this part we illustrate the method discussed in the previous sections with a dataset from an aluminum smelter. The measurement vector includes three components which denote the content of SiO_2, MgO and Al_2O_3 (labeled as x_1, x_2 and x_3 below) in the charge. All these measurements are relevant to the operation of the smelter. Stability of the alumina level is desirable. The silica and magnesium oxide levels are affected by the raw materials and are potential covariates to be taken into account in a fully fledged multivariate scheme. There are 189 vectors in the dataset. Like many other phase II SPC procedures, our procedure assumes that observations at different time points are independent of each other. However, for this dataset, we found that the observations at different time points are actually correlated. In statistics, there is a specific research area, i.e., the time series analysis, for modeling this type of correlation over time, i.e., the autocorrelation (cf., e.g., Brockwell and Davis 2002). To properly apply the proposed procedure to this dataset, we first remove the autocorrelation by the following estimated autoregression models:

$$\left\{ \begin{array}{lcl} x_1(i) - 0.63 & = & 0.07(x_1(i-1) - 0.63) + 0.12(x_1(i-2) - 0.63) + 0.28(x_1(i-3) - 0.63) + \epsilon_1(i) \\ x_2(i) - 12.97 & = & 0.55(x_2(i-1) - 12.97) + \epsilon_2(i) \\ x_3(i) - 57.86 & = & 0.32(x_3(i-1) - 57.86) + \epsilon_3(i), \end{array} \right. \tag{7}$$

where $\epsilon_1(i), \epsilon_2(i)$, and $\epsilon_3(i)$, for $3 \leq i \leq 189$, are residuals that are independent of each other at different time points. After the autocorrelation is removed, the data (i.e., $\epsilon_1(i), \epsilon_2(i)$, and $\epsilon_3(i)$, for $3 \leq i \leq 189$) are shown in Figure 3(a)-(c). The corresponding density curves are shown in Figure 3(d)-(f). It can be checked that the median vector of the original data $(x_1(i), x_2(i), x_3(i))$ has a shift if and only if the median vector of the residuals $(\epsilon_1(i), \epsilon_2(i), \epsilon_3(i))$ has a shift. See Lu and Reynolds (1999) for more discussions about the relationship between the original measurements and their residuals.

As an illustration of the log-linear modeling approach, the first 95 vectors of the residuals are used as an in-control dataset. Both the χ^2 and the Kolmogorov-Smirnov goodness-of-fit tests conclude that ϵ_1 and ϵ_3 in this dataset are not Normally distributed (χ^2 test: $\chi^2 = 141.3263$ and p-value=0 for ϵ_1, $\chi^2 = 65.2421$ and p-value=0 for ϵ_3; Kolmogorov-Smirnov test: p-value=0 for ϵ_1, p-value=0.022 for ϵ_3). So the joint distribution of $\mathbf{X}(i) = (\epsilon_1(i), \epsilon_2(i), \epsilon_3(i))$ can not be Normal because a joint Normal distribution implies that all marginal distributions are Normal. From

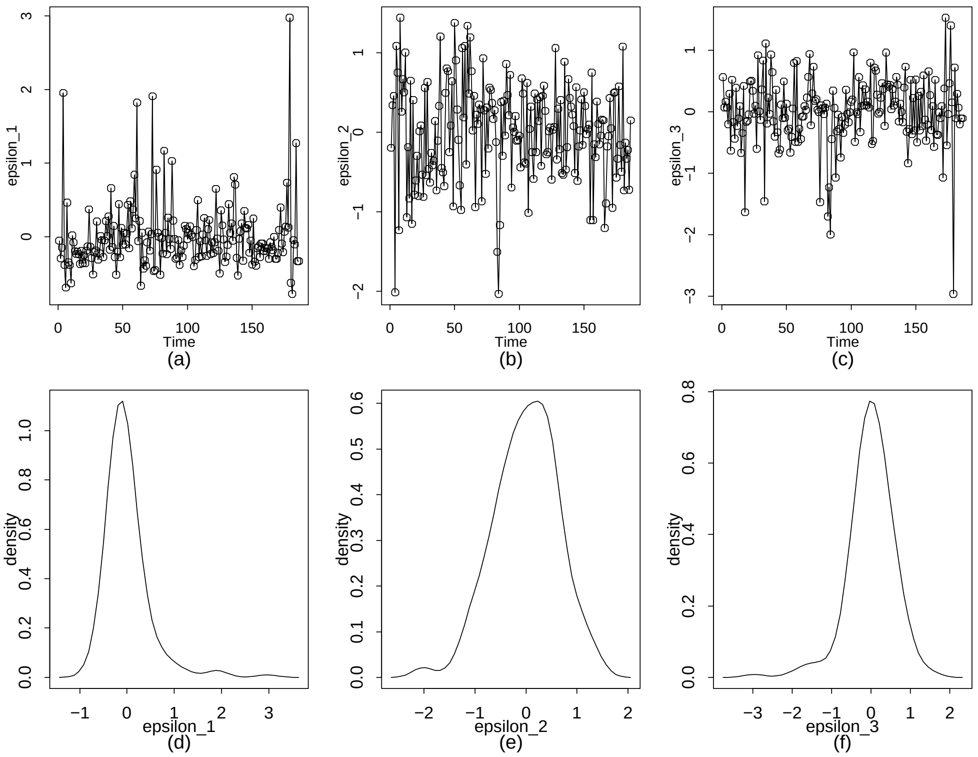

Figure 3: (a)-(c) The data after autocorrelation is removed from each measurement component (cf., equation (7)). (d)-(f) The corresponding density curves.

Figures 3(a) and 3(d), we can see that there are a number of large observations in ϵ_1 which make its density curve skewed to right. Similarly, the density curve of ϵ_3 seems skewed to the left. We tried log, square-root, and several other commonly used transformations for each of ϵ_1 and ϵ_3, and neither of these transformations can make the transformed variables nearly Normally distributed. A possible explanation of this phenomenon is that, excluding some relatively large observations of ϵ_1, distribution of its remaining observations is quite symmetric. The log, square-root, and similar transformations can pull down those relatively large observations; but they will make the distribution of the remaining observations less symmetric as well. A similar explanation can be made for ϵ_3. By the way, elementary statistical theory tells us that, even in cases when we can transform the individual variables to be nearly Normally distributed, it is not guaranteed that the joint distribution of the transformed individual variables would be nearly Normal. So, it may be difficult to transform this data to a multivariate Gaussian data.

From the in-control dataset, the selected log-linear model is (Y_1, Y_2Y_3), the estimated in-control distribution $\{f^{(0)}_{j_1,j_2,j_3}, j_1, j_2, j_3 = 0, 1\}$ of $\mathbf{Y}(i)$ by this selected model is calculated to be (0.1053,0.1474,0.1158,0.1368,0.1895, 0.0632,0.0947,0.1474). By the Algorithm I presented in Section 3, the control limit value is searched to be 10.793, for $k = 0.1$ and the in-control ARL=200. Then procedure (3)-(6) is used for detecting shifts in the remaining part of the data. The values of the CUSUM criterion u_n are presented in Figure 4. The control limit value is indicated in the same plot by the dotted line. It can be seen that there is a convincing evidence for a shift occured right at the begining of the second half of the data based on procedure (3)-(6).

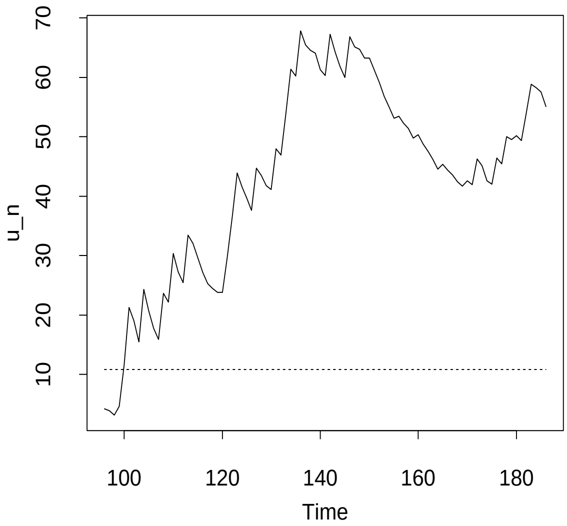

Figure 4: The cusum criterion u_n of procedure (3)-(6). The dotted line indicates the control limit value of the procedure such that its in-control ARL equals 200.

5 Some Concluding Remarks

We have presented a procedure to describe associations among measurement components of a process when the measurement distribution is non-Gaussian. This procedure is based on log-linear modeling. It is shown that the log-linear estimator of the in-control distribution of $\mathbf{Y}$ is often better than the estimator by the conventional relative frequency method. A CUSUM procedure for detecting shifts in a location parameter vector of the measurement distribution is also suggested for Phase II SPC. This procedure is distribution-free in the sense that all its properties depend on the distribution of $\mathbf{Y}$ only. It does not require the Gaussian distribution assumption on the measurement distribution as most existing procedures did. It is shown that the performance of this distribution-free procedure is improved by using the log-linear estimator of the in-control distribution of $\mathbf{Y}$, compared to its performance based on the conventional relative frequency estimator.

In applications, when the observed multivariate data are non-Gaussian, the first approach that we can think of is to transform the multivariate non-Gaussian data to a multivariate Gaussian data. However, this task itself is notoriously difficult for statisticians, because a p-dimensional Gaussian distribution implies that all its lower-dimensional marginal distributions should be Gaussian, and it is usually not good enough to just transform individual variables to be nearly Normally distributed (cf., e.g., Eaton 1983). So far, in the literature, statistical tools to transform multivariate non-Gaussian data to multivariate Gaussian data are very limited. When the Gaussian distribution assumption is violated, it has been demonstrated in the paper that conventional multivariate CUSUM procedures may not be reliable (cf., Figure 1). The proposed procedure provides a reasonable tool for SPC in such cases.

When the observed multivariate data are non-Gaussian and when a set of in-control data is available, another possible approach for phase II SPC is to adjust the control limit of a conventional multivariate CUSUM (e.g., the one by Crosier (1988)) using a numerical algorithm, such that the nominal ARL_0 value is achieved in certain accuracy. However, when the amount of in-control observations is limited, which is always the case in applications, it is still an open problem how to design such a numerical algorithm for adjusting the control limit. A major challenge lies behind the fact that such numerical algorithms (e.g., the ones described in Section 2) usually require a great amount of in-control observations in order to obtain a reasonably large amount of estimates of the in-control run length. To this end, we may consider using the bootstrap resampling technique. But, as pointed out by Hall *et al.* (1989), the bootstrap estimate of a tail probability of a distribution is often unreliable if we resample directly from the observed data. Note that the control limit in the current problem is related to a right-tail probability of the in-control distribution of the CUSUM statistic. Therefore, it may require much future research to accomplish this idea.

Regarding our proposed method, there are still many issues that need to be addressed properly in our future research. For instance, theoretically speaking, the proposed log-linear modeling procedure can handle cases with any number of measurement components. However, when this number is large, log-linear modeling would become challenging because the corresponding contingency table would become sparse in the sense that it will have many empty cells. There are some discussions in the literature about analysis of sparse contingency tables (cf., e.g., Agresti 2002, Section 9.8). It is still unknown to us how the proposed log-linear modeling procedure would perform in such cases for describing the in-control multivariate measurement distribution. At this moment, besides shifts in a location parameter vector, there is not a robust companion procedure for detecting shifts in dispersion or other aspects of the multivariate measurement distribution for phase II SPC. No robust companion procedures exist for detecting shifts, outliers, or other problems, for phase I SPC either.

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

References

Agresti, A. (2002), *Categorical Data Analysis (2nd edition)*, John Wiley & Sons: New York.

Albers, W., and Kallenberg, W.C.M. (2004), "Empirical nonparametric control charts: estimation effects and corrections," *Journal of.Applied Statistics*, **31**, 345–360.

Bakir, S.T. (2004), "A distribution-free Shewhart quality control chart based on signed-ranks," *Quality Engineering*, **16**, 611–621.

Bakir, S.T. (2005), "A quality control chart for work performance appraisal," *Quality Engineering*, **17**, 429–434.

Bakir, S.T. (2006), "Distribution-free quality control charts based on signed-rank-like statistics," *Communications in Statistics – Theory and Methods*, **35**, 743–757.

Brockwell, P.J., and Davis, R.A. (2002), *Introduction to Time Series and Forecasting (2nd. ed.),* Springer-Verlang: New York.

Chakraborti, S., van der Laan, P. and Bakir, S.T. (2001), "Nonparametric control charts: an overview and some results," *Journal of Quality Technology*, **33**, 304–315.

Chakraborti, S., van der Laan, P. and van de Wiel, M. (2004), "A class of distribution-free control charts," *Journal of the Royal Statistical Society, Series C*, **53**, 443–462.

Chen, G., Cheng, S.W., and Xie, H. (2005), "A new multivariate control chart for monitoring both location and dispersion," *Communications in Statistics – Simulation and Computation*, **34**, 203–217.

Chen, G., and Zhang, L. (2004), "EWMA charts for monitoring the mean censored Weibull lifetimes," *Journal of Quality Technology*, **36**, 321–328.

Crosier, R.B. (1988), "Multivariate generalizations of cumulative sum quality-control schemes," *Technometrics*, **30**, 291-303.

Eaton, M.L. (1983), *Multivariate Statistics,* John Wiley & Sons: New York.

Fang, K.T., Kotz, S., and Ng., K.W. (1990), *Symmetric Multivariate and Related Distributions*, New York: Chapman and Hall.

Haberman, S.J. (1977), "Log-linear models and frequency tables with small expected cell counts," *The Annals of Statistics*, **5**, 1148–1169.

Hall, P., DiCiccio, T.J., and Romano, J.P. (1989), "On smoothing and the bootstrap," *The Annals of Statistics*, **17**, 692–704.

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*, New York: Springer-Verlag.

Healy, J.D. (1987), "A note on multivariate CUSUM procedures," *Technometrics,* **29**, 409-412.

Lu, C.W., and Reynolds, M.R., Jr. (1999), "Control charts for monitoring the mean and variance of autocorrelated processes," *Journal of Quality Technology*, **31**, 259-274.

Mason, R.L., Champ, C.W., Tracy, N.D., Wierda, S.J., and Young, J.C. (1997), "Assessment of multivariate process control techniques," *Journal of Quality Technology*, **29**, 140-143.

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," *JRSS-D (The Statistician)*, **52**, 151-164.

Stoumbos, Z.G., and Sullivan, J.H. (2002), "Robustness to non-normality of the multivariate EWMA control chart," *Journal of Quality Technology*, **34**, 260–276.

Woodall, W.H. (2000), "Controversies and contradictions in statistical process control," *Journal of Quality Technology*, **32**, 341-350.

Yeh, A.B., and Lin, D.K.-J. (2002), "A new variables control chart for simultaneously monitoring multivariate process mean and variability," *International Journal of Reliability, Quality and Safety Engineering,* **9**, 41–59.

Yeh, A.B., Lin, D.K.-J., Zhou, H., and Venkataramani, C. (2003), "A multivariate exponentially weighted moving average control chart for monitoring process variability," *Journal of Applied Statistics*, **30**, 507–536.

Yeh, A.B., Huwang, L., and Wu, Y.-F. (2004), "A likelihood ratio based EWMA control chart for monitoring variability of multivariate normal processes," *IIE Transactions on Quality and Reliability Engineering*, **36**, 865–879.

Yeh, A.B., Lin, D.K.-J., Zhou, H., and McGrath, R.N. (2006), "Multivariate control charts for monitoring the covariance matrix: a review," *Quality Technology and Quantitative Management,* **3**, 415–436.