Copula-based monitoring schemes for non-Gaussian multivariate processes

Copula-based monitoring schemes for non-Gaussian multivariate processes

Pavel Krupskii[a], Fouzi Harrou[a], Amanda S. Hering[b], Ying Sun[a]

[a]King Abdullah University of Science and Technology (KAUST)
Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Thuwal 23955-6900, Saudi Arabia
E-mail: pavel.krupskiy@kaust.edu.sa, fouzi.harrou@kaust.edu.sa, ying.sun@kaust.edu.sa
[b]Department of Statistical Science, Baylor University, Waco, Texas, USA
E-mail: mandy_hering@baylor.edu

Abstract

Multivariate statistical monitoring charts are efficient tools for assessing the quality of a process by identifying abnormalities. Most commonly used multivariate monitoring charts, such as the Hotelling T^2 rule, however, assume the availability of uncorrelated Gaussian observations. Unfortunately, very often, real data do not satisfy these assumptions, and thus limit the usefulness of these techniques in practice. Furthermore, in many real applications, changes can occur in the shape of the multivariate distribution of the process while its mean or variance remains the same. Conventional process monitoring charts, such as the T^2 chart, fail to detect such changes in the distribution. In this paper, we develop new copula-based multivariate monitoring techniques for possibly autocorrelated, non-Gaussian data that can detect changes in the shape of a multivariate distribution that are usually overlooked by conventional monitoring charts. Using synthetic data, we demonstrate the effectiveness of the developed charts over a conventional monitoring chart. Results indicate that the proposed charts are very promising because copula-based charts are, in practice, designed to monitor the entire distribution of a process instead of the individual components. The developed monitoring charts are validated through practical application on data from a decentralized wastewater treatment plant in Golden, CO, USA.

Keywords: Copula; Multivariate control chart; Non-Gaussian distribution; Shape of distribution; Skewness.

1. Introduction

Fault detection and diagnosis have a vital role in modern industrial processes to enhance productivity, efficiency, and safety, and to reduce expensive maintenance (Isermann, 2006). Early detection of anomalies is crucial not only to maintain proper process operation, but also for the sake of people's health. Therefore, to avoid catastrophic consequences, it is essential to have a process monitoring system. Monitoring is required to maintain both product quality and safe operation. Fault detection and diagnosis are two vital components of process monitoring, during which faults are first identified and then isolated to ensure that they are appropriately handled (Basseville and Nikiforov, 1993).

Over the last two decades, the growing demand for safety, reliability, and efficiency in complex systems has led to an increased research interest in the areas of fault detection (FD) and diagnosis. Various techniques for fault detection have been developed, which can be classified into two main categories: data-based and model-based techniques. Model-based FD techniques are performed by comparing the process-measured variables with information obtained from the process model (Harrou et al., 2014). In contrast to the model-based techniques, data-based techniques do not need a process model since they rely on the availability of historical process data to represent process performance (Russell et al., 2012; Gertler, 1998). Such methods often generate residuals to quantify the distance between the measured and the desired behaviors. These residuals remain small or close to zero in the absence of faults and become significantly large in the presence of faults (Yin et al., 2014).

Several data-based fault detection approaches have been developed in the literature to meet various requirements in practical use, and they can be grouped into two main classes: univariate and multivariate techniques (Qiu, 2013; Montgomery, 2005; Bissell, 1994). The univariate statistical monitoring techniques are used to monitor only one process variable. Such techniques include the Shewhart chart (Shewhart, 1930; Montgomery, 2005), the cumulative sum (CUSUM) control charts (Montgomery, 2005), and the exponentially weighted moving average (EWMA) charts (Lucas and Saccucci, 1990; Rabhu and Runger, 1997). Although these univariate control charts have been used in most industries, they are only appropriate under the assumption that each observed variable is independent of others. When looking at multivariate data, these methods will ignore the interaction between the correlated variables and therefore result in a misleading analysis. Unlike the the traditional univariate statistical process control, multivariate statistical analysis methods take the correlation between variables into account and study many variables simultaneously. Multivariate statistical monitoring approaches include the multivariate Shewhart chart (Hotelling, 1947), multivariate EWMA (Lowry et al., 1992), and multivariate CUSUM (Crosier, 1988). The two essential assumptions underlying the design of conventional process monitoring charts are that the process observations are independent and identically normally distributed (i.i.d.) (Kazor et al., 2016). However, data collected from modern industrial processes are often autocorrelated or non-normally distributed. The violation of these major assumptions seriously affects the monitoring performance of conventional charts. The main goals of this paper are therefore to extend the capability of multivariate process monitoring techniques to handle autocorrelated and/or non-Gaussian processes, to detect changes in the shape of the joint distribution, and to improve the applicability of the developed methods in practice.

To model non-Gaussian multivariate observations, the copula is a powerful tool that can be used to construct flexible multivariate distributions. Copulas have been used in a wide range of actuarial, financial, and environmental studies; see (Nelsen, 2006) and (Joe, 2014) for detailed overviews. A copula is a multivariate cumulative distribution function (cdf) with univariate uniform $U(0, 1)$ marginals; this function can be used to link univariate marginals to construct a joint multivariate cdf. Sklar (1959) showed that for continuous univariate distribution functions $F_1, \dots, F_m$ and a continuous m-dimensional distribution function F, there exists a unique copula function, C, such that $F(z_1, \dots, z_m) = C\{F_1(z_1), \dots, F_m(z_m)\}$ for any $z_1, \dots, z_m$.

Other approaches address fault detection as a classification problem and without relying on any distributional assumptions (Deng et al., 2012; Sukchotrat et al., 2009; He and Wang, 2007). Liu et al. (2004) proposed statistical charts based on data depth. Recently, Zhang et al. (2016) proposed to use a multivariate goodness-of-fit test to detect changes in the shape of a multivariate distribution. All of these procedures do not provide additional information regarding the changes in the shape of the distribution that could help to localize the fault.

Until recently, copulas have not been widely used to improve the performance of multivariate monitoring charts. Fatahi et al. (2011) proposed a copula-based bivariate control chart to monitor two correlated events that have zero inflated Poisson distributions. Kuvattana et al. (2015) compared the performance of the CUSUM control chart constructed using four different copula families when observations follow exponential distributions. Han and Liu (2013) analyzed the theoretical performance of a principal component analysis for dependent multivariate data generated from the Gaussian copula with non-Gaussian univariate marginals and proposed Copula Component Analysis. All of these methods, however, require some assumptions on the marginal distributions or the joint copula density.

Verdier (2013) proposed using a copula function to construct monitoring charts based on the estimated density levels of the observations and showed that the classical Hotelling T^2 rule can produce many false alarms for non-Gaussian data. In this paper, we show that the charts constructed using density levels may not detect changes in the shape of the multivariate distribution of the monitored process variables. Moreover, this method of constructing charts requires an estimation of the copula density, and if this density is misspecified, these charts can also produce many false alarms. We develop two new copula-based multivariate monitoring techniques that can be more sensitive to these changes and that do not require estimation of the joint distribution. Unlike many existing nonparametric methods, these charts provide additional information regarding the observed change in the distribution and can ultimately assist in diagnosing the cause of the fault.

The following section briefly reviews the multivariate Shewhart chart used for multivariate process monitoring. In Section 3, we show that the monitoring charts based on density levels are not very robust against density misspecification and that these charts may not detect changes in the shape of the density. We then introduce new multivariate monitoring methods that do not require modeling of the joint cdf and that are more sensitive to changes in the density. In Section 4, the performances of the proposed methods are illustrated in a simulation study and a real data application, and Section 5 concludes with a discussion.

2. Multivariate Control Charts

Consider a data matrix $\mathbf{X} \in R^{n\times m}$ with n measurements and m process variables, $\mathbf{X} = \left[\mathbf{x}_1^T, \dots, \mathbf{x}_n^T\right]^T$. Assume $\mathbf{x}_t$, $t = 1, 2, \dots, n$, follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. A multivariate Shewhart chart, also known as a T^2 chart or a χ^2 chart (Hotelling, 1947), for monitoring the process mean is based on the decision statistic:

$$T_t^2 = \left[(\mathbf{x}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_t - \boldsymbol{\mu})\right], \tag{1}$$

where $\mathbf{x}_t$ is a vector of m variables, $\boldsymbol{\mu}$ is a vector of in-control means of each variable, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\mathbf{X}$. The upper limit for this control chart is: $UCL = \chi^2_{1-\alpha,m}$. For new testing data, when the value of T_t^2 exceeds the threshold value, $\chi^2_{1-\alpha,m}$, a fault is declared (Hotelling, 1933). Multivariate Shewhart monitoring charts tend to fail to detect small and moderate faults in the mean vector (Qiu, 2013). In cases when the normality assumption is violated, the detection performance of the T^2 chart would be degraded, and to handle such cases, this study presents the development of copula-based control charts that do not require any distributional assumptions for their construction.

3. Copula-based Multivariate Control Charts

In this section, we briefly describe how multivariate control charts can be constructed using density level sets and discuss two major weaknesses of this approach: the copula density estimation and the sensitivity of this procedure to the distribution changes. We then introduce two new multivariate control charts that are more sensitive to changes in the shape of the joint distribution of the monitored process and that do not require estimation of the joint copula density.

3.1. Using density levels for multivariate monitoring charts

One way to construct a multivariate control chart is to use procedures based on a density level set. For a given m-dimensional probability density function, f_0, and a given constant, c, the density level set is defined as $\{\mathbf{z} \in \mathbb{R}^m : f_0(\mathbf{z}) > c\}$. Let $\mathbf{Z} = (Z_1, \ldots, Z_m)^T$ be a monitored m-dimensional process, $\mathbf{z} = (z_1, \ldots, z_m)^T$, and let f_0 be the density of $\mathbf{Z}$. If f_0 is known, one can use a training sample to estimate quantiles of $f_0(\mathbf{Z})$ and find thresholds c_L and c_U, such that $\Pr\{f_0(\mathbf{Z}) < c_L\} = \Pr\{f_0(\mathbf{Z}) > c_U\} = \alpha/2$, where α is a prespecified significance level. Under the assumption of multivariate normality of the vector $\mathbf{Z}$, this procedure is equivalent to using the T^2 chart from Hotelling (1947).

One problem with using such a procedure is that the true density, denoted f_0, is usually unknown and needs to be estimated from data. If one uses a misspecified density function $\tilde{f}_0$, the resulting control chart may not retain the nominal false alarm rate. To illustrate this idea, assume that c_L and c_U are selected such that $\Pr\{c_L < \tilde{f}_0(\tilde{\mathbf{Z}}) < c_U\} = 1 - \alpha$, where the vector $\tilde{\mathbf{Z}}$ has the joint density $\tilde{f}_0$. Suppose there exist positive constants k_1, k_2 such that

$$f_0(\mathbf{z}) = \begin{cases} k_1 \tilde{f}_0(\mathbf{z}), & \text{if } c_L \le \tilde{f}_0(\mathbf{z}) \le c_U, \\ k_2 \tilde{f}_0(\mathbf{z}), & \text{otherwise}, \end{cases} \tag{2}$$

where f_0 is a valid density function, that is

$$\int_{\mathbf{z}} f_0(\mathbf{z}) \mathrm{d}\mathbf{z} = k_1 \int_{\mathbf{z}: c_L \le \tilde{f}_0(\mathbf{z}) \le c_U} \tilde{f}_0(\mathbf{z}) \mathrm{d}\mathbf{z} + k_2 \int_{\mathbf{z}: \tilde{f}_0(\mathbf{z}) < c_L \text{ or } \tilde{f}_0(\mathbf{z}) \ge c_U} \tilde{f}_0(\mathbf{z}) \mathrm{d}\mathbf{z} = k_1(1-\alpha) + k_2 \alpha = 1.$$

If $k_2 = M$, $0 < M < 1/\alpha$, and $k_1 = (1 - M\alpha)/(1-\alpha)$, then the probability of a false alarm is

$$\Pr\{\tilde{f}_0(\mathbf{Z}) < c_L \text{ or } \tilde{f}_0(\mathbf{Z}) > c_U\} = \int_{\mathbf{z}: \tilde{f}_0(\mathbf{z}) < c_L \text{ or } \tilde{f}_0(\mathbf{z}) \ge c_U} f_0(\mathbf{z}) \mathrm{d}\mathbf{z} = k_2 \int_{\mathbf{z}: \tilde{f}_0(\mathbf{z}) < c_L \text{ or } \tilde{f}_0(\mathbf{z}) \ge c_U} \tilde{f}_0(\mathbf{z}) \mathrm{d}\mathbf{z} = M\alpha,$$

and this value can be very large if $M > 1$. The reason for a large false alarm probability is that $\tilde{f}_0$ is inflated by a constant within a given range and by another constant outside of that range. In other words, the misspecified density $\tilde{f}_0$ has a different shape, and the monitoring chart based on a density level set cannot detect this change in the shape of the distribution.

3.2. Estimation of density levels using copulas

Verdier (2013) proposed using copulas to model the true multivariate density f_0. Let c_0 be the copula corresponding to the density f_0, and $F_i(z_i)$ and $f_i(z_i)$ are the marginal cdf and density of the i-th component of the vector $\mathbf{Z}$, respectively. For any vector $\mathbf{z}$, we have $f_0(\mathbf{z}) = c_0(F_1(z_1), \ldots, F_m(z_m)) \prod_{i=1}^m f_i(z_i)$. To estimate the true density, one therefore needs to estimate marginal distributions F_i, $i = 1, \ldots, m$, and the copula density c_0. Marginal distributions can be estimated by many methods, either parametrically or nonparametrically. If the Z_i's are autocorrelated, then one can use the autoregressive-moving-average model below to remove the serial autocorrelation first:

$$Z_{i,t} = \mu_i + \epsilon_{i,t} + \sum_{k=1}^{k_i} \alpha_{k,i} Z_{i,t-k} + \sum_{l=1}^{l_i} \beta_{l,i} \epsilon_{i,t-l}, \quad i = 1, \ldots, m,$$

where $Z_{i,t}$ is the i-th component of the multivariate process $\mathbf{Z}$ observed at time t; see Box et al. (1994) for details. Other variables, such as a seasonal cycle or a trend, can be added if needed. The vectors of i.i.d. residuals $\epsilon_t = (\epsilon_{1,t}, \ldots, \epsilon_{m,t})^T$ can then be used to estimate the joint copula density c_0. If the observed data are not autocorrelated, one can simply use the original data $(Z_{1,t}, \ldots, Z_{m,t})^T$ to estimate c_0. For the remainder of this paper, we assume that the observations of our process are independent as was done in Verdier (2013).

Modeling the copula c_0 is usually the most challenging part of using copula-based control charts, especially if the dimension m is high. If the training set is not large, it is very difficult to estimate the joint copula density, especially using nonparametric methods. One can use some parametric copula families in which dependence can be modeled using only one or two dependence parameters. Examples with one dependence parameter, θ, include the Frank copula (Frank, 1979) with the cdf

$$C_{\text{Frank}}(u_1, \ldots, u_m; \theta) = -\frac{1}{\theta} \log \left[1 + \frac{\prod_{i=1}^m \{\exp(-\theta u_i) - 1\}}{\{\exp(-\theta) - 1\}^{m-1}} \right], \quad \theta \neq 0,$$

the Gumbel copula (Gumbel, 1960) with the cdf

$$C_{\text{Gumbel}}(u_1, \ldots, u_m; \theta) = \exp \left[- \left\{ \sum_{i=1}^m (-\log u_i)^\theta \right\}^{1/\theta} \right], \quad \theta \geq 1,$$

and the Clayton copula (Clayton, 1978) with the cdf

$$C_{\text{Clayton}}(u_1, \ldots, u_m; \theta) = \left(m - 1 + \sum_{i=1}^m u_i^{-\theta} \right)^{-1/\theta}, \quad \theta > 0.$$

However, these copula families are not very flexible when $m > 2$ as they assume that all lower-dimensional marginals are the same. Moreover, many goodness-of-fit tests, such as the ones proposed by Genest et al.

(2009) or Kojadinovic and Yan (2011), may be not very sensitive if the training set is small. Also, these tests only provide p-values and do not give information about which copula is most suitable for the data. One therefore cannot rely on these tests to select a suitable copula family to model the joint density f_0.

If one uses a wrong copula density $\tilde{c}_0$ when the true one is c_0, then the joint density f_0 will be misspecified even if the univariate marginal cdfs $F_1, \ldots, F_m$ are modeled correctly. This implies that one needs to estimate both the marginal distributions and the joint copula accurately in order for multivariate control charts that use density level sets to construct the rejection region.

3.3. Sensitivity of the procedure based on density levels

Even if the density f_0 is estimated correctly, the resulting multivariate chart may be not sensitive with respect to changes in the shape of the density f_0. While detecting changes in the mean or variance of the individual components Z_i of the vector $\mathbf{Z}$ is important, one also needs procedures that can efficiently detect changes in the shape of f_0 as these can hinder detection of anomalies in the individual components. We now illustrate these ideas with the following example. Let the true density for the training set be f_0. We assume that f_0 is a symmetric density, i.e., $f_0(-\mathbf{z}) = f_0(\mathbf{z})$ for all vectors $\mathbf{z}$. Typical examples include the multivariate normal or Student's t distributions. Let the true density for the testing sample be f_0^* where

$$f_0^*(\mathbf{z}) = \begin{cases} k_1 f_0(\mathbf{z}), & \text{if} \sum_{i=1}^d \beta_i z_i \geq 0, \\ k_2 f_0(\mathbf{z}), & \text{otherwise,} \end{cases} \tag{3}$$

where $\beta_1, \ldots, \beta_d$ are any constants, and positive constants k_1 and k_2 are selected such that f_0^* is a valid density function:

$$\begin{aligned} \int_{\mathbf{z}} f_0^*(\mathbf{z}) \mathrm{d}\mathbf{z} &= k_1 \int_{\mathbf{z}: \sum_{i=1}^d \beta_i z_i \geq 0} f_0(\mathbf{z}) \mathrm{d}\mathbf{z} + k_2 \int_{\mathbf{z}: \sum_{i=1}^d \beta_i z_i < 0} f_0(\mathbf{z}) \mathrm{d}\mathbf{z} \\ &= 0.5 k_1 \int_{\mathbf{z}} f_0(\mathbf{z}) \mathrm{d}\mathbf{z} + 0.5 k_2 \int_{\mathbf{z}} f_0(\mathbf{z}) \mathrm{d}\mathbf{z} = 0.5(k_1 + k_2) = 1. \end{aligned}$$

For a vector $\mathbf{Z}$ with the joint density f_0, we select c_L and c_U such that $\Pr\{f_0(\mathbf{Z}) < c_L\} = \Pr\{f_0(\mathbf{Z}) > c_U\} = \alpha/2$, as before. If $\mathbf{Z}^*$ is a sample from the testing set with the density f_0^*, then the probability of

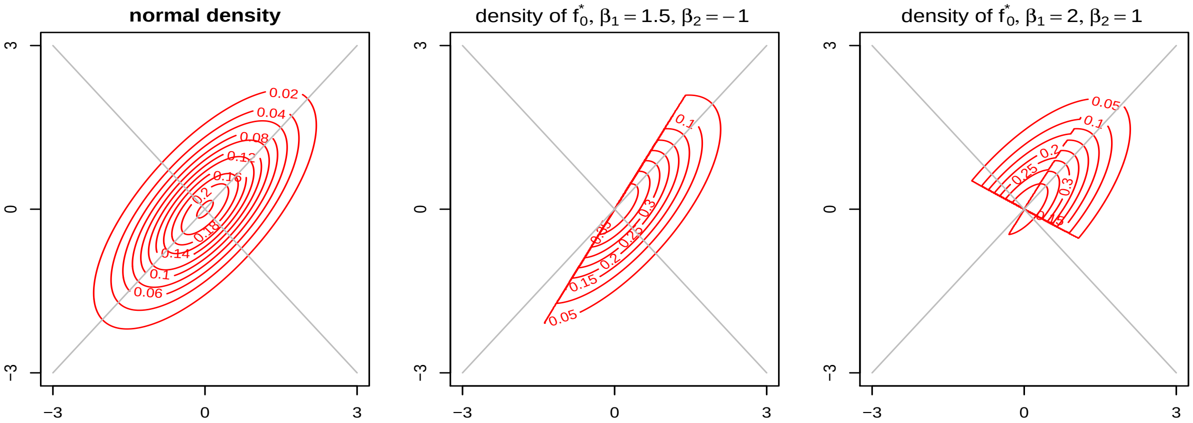

Figure 1: Contour plots of the normal density with correlation 0.7, f_0 (left), and the new density, f_0^*, with $\beta_1 = 1.5$, $\beta_2 = -1$, $k_1 = 1.75$, $k_2 = 0.25$ (middle), and $\beta_1 = 1$, $\beta_2 = 2$, $k_1 = 1.75$, $k_2 = 0.25$ (right).

detection using f_0 is α:

$$\begin{aligned}
&\Pr\{f_0(\mathbf{Z}^*) < c_L \text{ or } f_0(\mathbf{Z}^*) > c_U\} = \int_{\mathbf{z}: f_0(\mathbf{z}) < c_L \text{ or } f_0(\mathbf{z}) > c_U} f_0^*(\mathbf{z}) \mathrm{d}\mathbf{z} \\
&= k_1 \int_{\mathbf{z}: \sum_{i=1}^m \beta_i z_i \geq 0 \text{ and } f_0(\mathbf{z}) < c_L \text{ or } f_0(\mathbf{z}) > c_U} f_0(\mathbf{z}) \mathrm{d}\mathbf{z} + k_2 \int_{\mathbf{z}: \sum_{i=1}^m \beta_i z_i < 0 \text{ and } f_0(\mathbf{z}) < c_L \text{ or } f_0(\mathbf{z}) > c_U} f_0(\mathbf{z}) \mathrm{d}\mathbf{z} \\
&= 0.5(k_1 + k_2) \int_{\mathbf{z}: f_0(\mathbf{z}) < c_L \text{ or } f_0(\mathbf{z}) > c_U} f_0(\mathbf{z}) \mathrm{d}\mathbf{z} = \Pr\{f_0(\mathbf{Z}) < c_L \text{ or } f_0(\mathbf{Z}) > c_U\} = \alpha. \qquad (4)
\end{aligned}$$

Note that the density f_0^* is no longer symmetric when $k_1 \neq k_2$ (so that there is a change in the shape of the density), and in the general case all components of the vector $\mathbf{Z}^*$ will have means and variances different from those of the corresponding components of the vector $\mathbf{Z}$.

Figure 1 shows contour plots for the bivariate normal density, f_0, with correlation 0.7 and for the new density, f^*, defined in (3), using two sets of parameters: $\beta_1 = 1.5$, $\beta_2 = -1$, $k_1 = 1.75$, $k_2 = 0.25$ and $\beta_1 = 1$, $\beta_2 = 2$, $k_1 = 1.75$, $k_2 = 0.25$. The new density f_0^* has a different shape. The Spearman's correlation for the original density, f_0, is 0.70, and for the new density, f_0^*, they are 0.77 and 0.49 for the two sets of parameters, respectively. The mean vector for the original density is $\mathrm{E}[\mathbf{Z}] = (0, 0)^T$, and for the new densities, they are $\mathrm{E}[\mathbf{Z}^*] = (0.03, 0.51)^T$ and $\mathrm{E}[\mathbf{Z}^*] = (0.51, 0.57)^T$, respectively. However, the multivariate control chart still has a probability of detection equal to α as it follows from (4), and therefore it cannot detect these changes in the joint distribution.

More generally, let S_1 and S_2 be two disjoint subsets of $\mathbb{R}^m$ such that $S_1 \cup S_2 = \mathbb{R}^m$ and if $\mathbf{z} \in S_1$, then

$-\mathbf{z} \in S_2$, or vice versa. In other words, S_2 is the reflection of S_1 with respect to the origin. Define

$$f_0^*(\mathbf{z}) = \begin{cases} f_0(\mathbf{z}) - h(\mathbf{z}), & \text{if } \mathbf{z} \in S_1, \\ f_0(\mathbf{z}) + h(\mathbf{z}), & \text{otherwise,} \end{cases} \tag{5}$$

where $h(\mathbf{z})$ is a function such that $|h(\mathbf{z})| \leq f_0(\mathbf{z})$. Here, $h(\mathbf{z})$ represents the change in the shape of the density $f_0(\mathbf{z})$.

Similar to (3), one can check that (5) is a valid density function and that the probability of detection using f_0 is α. Many different asymmetric distributions can be obtained using this construction. For example, if $h(\mathbf{z})$ is positive in the joint lower tail, then the corresponding distribution is skewed with stronger dependence in the lower tail. As a result, the multivariate chart based on the copula density fails to detect these changes in the density $f_0(\mathbf{z})$.

In the next section, we propose two new monitoring procedures that can more efficiently detect such changes in the shape of the joint density f_0 and that do not require the estimation of the copula c_0.

3.4. The proposed multivariate control charts based on skewness and overall dependence measures for multivariate data

Assume that $\{(Z_{1,t}, \ldots, Z_{m,t})\}_{t=1}^{T_0}$ is the learning sample for the process $\mathbf{Z}$ with the density f_0, and $\{(Z_{1,t}, \ldots, Z_{m,t})\}_{t=T_0+1}^{T_1}$ is the testing sample for the process $\mathbf{Z}^*$ with possibly different density f_0^*. For $i = 1, \ldots, m$, define uniform ranks as follows:

$$U_{i,t} = \frac{\texttt{rank}(Z_{i,t}) - 0.5}{T_0}, \quad t = 1, \ldots, T_0.$$

Let c_0 be the copula corresponding to the density f_0, and let the density of a vector $(U_1, \ldots, U_m)^T$ be c_0. Define

$$\zeta_k = \mathrm{E}\left(\frac{U_1 + \ldots + U_m}{m} - \frac{1}{2}\right)^k,$$

and the empirical estimate of ζ_k is

$$\widehat{\zeta}_k = \frac{1}{T_0} \sum_{t=1}^{T_0} \left(\frac{U_{1,t} + \ldots + U_{m,t}}{m} - \frac{1}{2}\right)^k. \tag{6}$$

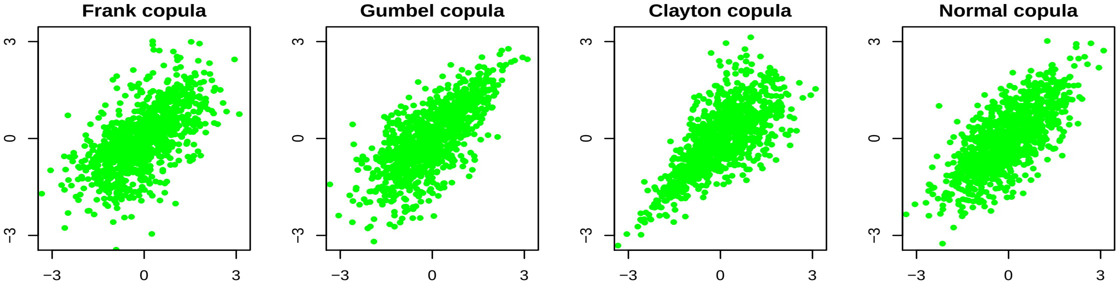

Figure 2: Scatter plots of f_0 with the following copula densities (left to right): Frank, Gumbel, Clayton, and normal, with standard normal marginal cdfs. Copula parameters are selected such that the Spearman's correlation equals 0.7 for all 4 copulas.

The measure ζ_k depends on the copula c_0:

$$\zeta_k = \int_{[0,1]^m} \left(\frac{u_1 + \ldots + u_m}{m} - \frac{1}{2} \right)^k c_0(u_1, \ldots, u_m) \mathrm{d}u_1 \ldots \mathrm{d}u_m.$$

With an even k, ζ_k can be used to detect changes in the strength of overall dependence, and with an odd k, this measure can be used to detect skewness (reflection asymmetry). A copula density c_0 is called reflection symmetric if $c_0(\mathrm{u}) = c_0(1 - \mathrm{u})$ for any $\mathrm{u} = (u_1, \ldots, u_m)^T \in [0,1]^m$. For a reflection symmetric copula density and an odd k, $\zeta_k = 0$. Otherwise, the measure ζ_k can be positive or negative, depending on the direction of reflection asymmetry; see Krupskii (2016) for details. If uniform marginal cdfs $F_1, \ldots, F_m$ are all symmetric (for example, normal), and the copula density c_0 is reflection symmetric, then $f_0(-\mathbf{z}) = f_0(\mathbf{z})$ for any $\mathbf{z} = (z_1, \ldots, z_m)^T \in \mathbb{R}^m$. Figure 2 shows the scatter plots of the bivariate density f_0 ($m = 2$) with standard normal $N(0, 1)$ marginal cdfs and three copulas (Frank, Gumbel, and Clayton) from Section 3.2. We also include a scatter plot of the bivariate normal density with standard normal marginals for comparison.

We see that the shapes of these distributions are quite different. The Frank and normal copulas are symmetric, while the Gumbel and Clayton are reflection asymmetric copulas, skewed to the left and to the right, respectively. We therefore use ζ_{2k+1} as a measure of skewness (reflection asymmetry). The estimate $\widehat{\zeta}_{2k+1}$, obtained using non-parametric ranks, $U_{i,t}$, $t = 1, \ldots, T_0$, $i = 1, \ldots, m$, is asymptotically normal with some variance σ^2_{2k+1} that can be estimated using non-parametric methods such as the jackknife; see Shao and Wu (1989). Let $\widehat{\sigma}^2_{2k+1}$ be an estimate of σ^2_{2k+1}. The 95% confidence interval for ζ_{2k+1} can then be constructed as follows:

$$\left(\widehat{\zeta}_{2k+1} - 1.96 \frac{\widehat{\sigma}_{2k+1}}{\sqrt{T_0}}, \widehat{\zeta}_{2k+1} + 1.96 \frac{\widehat{\sigma}_{2k+1}}{\sqrt{T_0}} \right).$$

We can use ζ_2 as a measure of overall dependence. This measure can be seen as a multivariate version of the Spearman's correlation coefficient. It can be shown that

$$
\begin{aligned}
\zeta_2 &= \frac{1}{m^2}\left\{\sum_{i=1}^{m} \mathrm{E}\left(U_i - \frac{1}{2}\right)^2 + 2\sum_{i_1>i_2} \mathrm{E}\left(U_{i_1} - \frac{1}{2}\right)\left(U_{i_2} - \frac{1}{2}\right)\right\} \\
&= \frac{1}{m}\mathrm{Var}(U_1) + 2\sum_{i_1>i_2} \mathrm{Cov}(U_{i_1}, U_{i_2}) \\
&= \frac{1}{12m} + \frac{1}{6m^2}\sum_{i_1>i_2} \rho_{i_1,i_2},
\end{aligned}
$$

where $\rho_{i_1,i_2} = \mathrm{cor}(U_{i_1}, U_{i_2})$ is the Spearman's correlation for pair (i_1, i_2). Larger Spearman's correlations between different pairs imply larger values for ζ_2. Similar to ζ_{2k+1}, one can construct the 95% confidence interval for ζ_2 as follows:

$$
\left(\widehat{\zeta}_2 - 1.96\frac{\widehat{\sigma}_2}{\sqrt{T_0}},\, \widehat{\zeta}_2 + 1.96\frac{\widehat{\sigma}_2}{\sqrt{T_0}}\right).
$$

Note that measures ζ_2 and ζ_{2k+1} do not depend on the univariate marginal distributions, but these measures do depend on the copula density c_0 and can detect changes in its shape. At the same time, c_0 does not need to be estimated to obtain estimates $\widehat{\zeta}_2$, $\widehat{\zeta}_{2k+1}$.

For the testing sample, we use the last K observations of the process $\mathbf{Z}$ to estimate ζ_k. For each $i = 1, \ldots, m$ and $t^* = T_0 + 1, \ldots, T_1$ and some $K > 0$ (moving window length), we define

$$
U_{i,t}^K = \frac{\mathtt{rank}(Z_{i,t}) - 0.5}{K}, \quad t = t^* - K + 1, \ldots, t^*,
$$

$$
\widehat{\zeta}_k^K(t^*) = \frac{1}{K}\sum_{t=t^*-K+1}^{t^*}\left(\frac{U_{1,t}^K + \ldots + U_{m,t}^K}{m} - \frac{1}{2}\right)^k. \tag{7}
$$

Let the estimate of ζ_k for the training set be ζ_k^0 (treated as a fixed value). Here we assume that, under the null hypothesis, the measure $\zeta_k^K(t^*)$ remains the same in the testing set, so that the difference $\widehat{\zeta}_k^K(t^*) - \zeta_k^0$ is close to zero and ζ_k^0 has very small variability (when the learning sample size, T_0, is large), so its estimation adds no uncertainty to (8). Then, the 95% confidence interval for $\widehat{\zeta}_k^K(t^*) - \zeta_k^0$ is:

$$
\left(-1.96\frac{\widehat{\sigma}_k}{\sqrt{K}},\, 1.96\frac{\widehat{\sigma}_k}{\sqrt{K}}\right). \tag{8}
$$

Using Monte Carlo simulations, we checked that some uncertainty that comes from estimating the asymptotic variance using the training set, $\widehat{\sigma}_k^2$, does not have a significant effect on the confidence interval even if the training sample size is not very large, and the false alarm rate is close to the nominal level. If $T_0, K \to \infty$, then $\widehat{\zeta}_k \to_P \zeta_k$, $\widehat{\sigma}_k \to_P \sigma_k$, and therefore

$$\Pr\left(\widehat{\zeta}_k^K(t^*) - 1.96\frac{\widehat{\sigma}_k}{\sqrt{K}} \leq \zeta_k^0 \leq \widehat{\zeta}_k^K(t^*) + 1.96\frac{\widehat{\sigma}_k}{\sqrt{K}}\right) \to 0.05.$$

Consequently, at time t^* an abnormality is detected if $|\widehat{\zeta}_k^K(t^*) - \zeta_k^0| > 1.96\widehat{\sigma}_k/\sqrt{K}$. With a large K, the confidence intervals will be narrower; however, the monitoring procedure based on $\zeta_k^K(t^*)$ will not be able to detect changes in density f_0 very quickly because it is constructed using many observations from the past. We found that a K between 80 and 100 can be a good choice in many scenarios; more details are given in the next section.

4. Simulation Study

In this section, we conduct simulation studies to show that the monitoring chart based on density levels is usually not very sensitive to changes in the shape of the density, unless there is a very large change in the mean of individual components of the observed process **Z**. The two testing charts based on the measures of skewness and overall dependence, ζ_{2k+1} and ζ_2, can be more sensitive to such changes, and they can also provide some information regarding the shape of the new density f_0^*.

We summarize the previous section to briefly describe how the control charts are constructed:

Step 1 Appropriate models are fitted to univariate marginals of the training data set of size T_0;

Step 2 Residuals from the fitted models are transformed to uniform ranks;

Step 3 Estimates of ζ_2 and ζ_{2k+1} are calculated using (6);

Step 4 Asymptotic variances σ_2^2 and σ_{2k+1}^2 are estimated by nonparametric methods, and the confidence intervals for the training data set are constructed using (8);

Step 5 For the testing data set, we calculate $\widehat{\zeta}_2^K(t^*)$ and $\zeta_{2k+1}^K(t^*)$ for $t^* = T_0 + 1, \ldots, T_1$ using (7). An anomaly is detected if these estimates fall outside the corresponding confidence intervals constructed in the previous step.

4.1. Change in overall dependence, no change in shape

We investigate the behavior of three monitoring charts: the first one constructed using negative log-density levels (density levels-based chart) and the other two based on dependence measures ζ_{2k+1} and ζ_2

(skewness-based chart and overall dependence-based chart, respectively). We assume that the new density f_0^* has the same shape, and that the copula c_0 is not changed, but that the strength of dependence changes. For example, for the multivariate normal distribution, testing data have stronger correlations. For the training sample, we generate data from 4 different copulas (normal, Frank, Gumbel, Clayton), and we assume that all marginal cdfs are standard normal. For the testing sample, we generate data from the same copulas but with different parameters. For each copula, we consider cases when the marginal distributions remain the same and when there is a shift in mean for all individual components (such that the new marginal distributions are all normal with some mean μ). For the monitoring chart based on the density level sets, we assume that the copula for the training data set is known, and it is used to construct the 95% confidence intervals. In practice, the density needs to be estimated from the data, and this can result in a worse performance of this monitoring chart.

We assume that vector $\mathbf{Z}$ has three components so that the corresponding copula is trivariate (similar results are obtained for dimensions $m = 2, 3, \ldots, 10$). We use $T_0 = 200$ for the training data and $T_1 - T_0 = 200$ for the testing data. The copula parameters for the training/testing data sets are 0.25/0.5 for the normal copula, 1.5/3.3 for the Frank copula, 1.2/1.5 for the Gumbel copula, and 0.4/1.0 for the Clayton copula. This corresponds to a change in Spearman's rho from approximately 0.3 to 0.5 for all of these models.

In general, ζ_k depends on k-dimensional marginal distributions, and it is more sensitive to changes in the shape of the distribution when k is larger. However, the variability of this measure is also larger for larger k, and a large sample size is required to estimate it with a good accuracy. We use $k = 3$ for ζ_{2k+1} because Krupskii (2016) showed that the measure ζ_7 can be more sensitive to reflection asymmetry of copula density c_0 when $m \leq 10$, and statistical tests based on ζ_7 have good power for data sets with small to moderate sample size. For high-dimensional data with $m > 10$, $k = 1$ gives good results in simulations with $K \leq 200$. Larger k can be used if $K > 200$.

We calculate the number of detections in different scenarios for the three monitoring procedures, and we use $K = 80$ to obtain estimates of $\widehat{\zeta}_2$, $\widehat{\zeta}_7$, and the corresponding 95% confidence intervals. We use the following scenarios:

1. No change in the density function;
2. Change in the copula parameter, no shift in the mean;
3. Change in the copula parameter, shift in the mean (new mean $\mu = 1$);
4. Change in the copula parameter, shift in the mean (new mean $\mu = -1$).

Table 1: Proportion of detections for three testing procedures for different scenarios with copula parameters shown in parentheses. For each scenario, 1000 data sets were generated to calculate these values.

Training data set	Testing data set	Density levels-based chart	Skewness-based chart	Overall dependence-based chart
Normal (0.50) + $N(0,1)$ marginals	Normal (0.50) + $N(0,1)$ marginals	0.06	0.07	0.08
Normal (0.25) + $N(0,1)$ marginals	Normal (0.50) + $N(0,1)$ marginals	0.06	0.15	0.69
Frank (1.50) + $N(0,1)$ marginals	Frank (3.30) + $N(0,1)$ marginals	0.07	0.16	0.68
Gumbel (1.20) + $N(0,1)$ marginals	Gumbel (1.50) + $N(0,1)$ marginals	0.06	0.16	0.64
Clayton (0.40) + $N(0,1)$ marginals	Clayton (1.00) + $N(0,1)$ marginals	0.05	0.27	0.63
Normal (0.25) + $N(0,1)$ marginals	Normal (0.50) + $N(1,1)$ marginals	0.15	0.21	0.82
Frank (1.50) + $N(0,1)$ marginals	Frank (3.30) + $N(1,1)$ marginals	0.17	0.22	0.80
Gumbel (1.20) + $N(0,1)$ marginals	Gumbel (1.50) + $N(1,1)$ marginals	0.09	0.13	0.74
Clayton (0.40) + $N(0,1)$ marginals	Clayton (1.00) + $N(1,1)$ marginals	0.22	0.32	0.78
Normal (0.25) + $N(0,1)$ marginals	Normal (0.50) + $N(-1,1)$ marginals	0.15	0.22	0.82
Frank (1.50) + $N(0,1)$ marginals	Frank (3.30) + $N(-1,1)$ marginals	0.20	0.22	0.80
Gumbel (1.20) + $N(0,1)$ marginals	Gumbel (1.50) + $N(-1,1)$ marginals	0.19	0.18	0.77
Clayton (0.40) + $N(0,1)$ marginals	Clayton (1.00) + $N(-1,1)$ marginals	0.08	0.30	0.75

For each scenario, we generate 1000 data sets to estimate the proportion of detections for the testing data set. The results are presented in Table 1.

We see that the proportions of false alarms are close to nominal levels for all monitoring charts, as shown in the first line of the table. With a larger training sample, the accuracy of the estimated intervals can be increased. The density levels-based chart cannot detect changes in the dependence and, even when there is a shift in mean of individual components of the vector $\mathbf{Z}$, the proportions of detections are very small (8% to 22%). In contrast, the overall dependence-based chart has 63% to 69% of detections when there is no change in the mean. The detection rate is even higher when the mean is changed to $+1$ or -1 (74% to 82%). This is because the ζ_2 measure is sensitive to changes in overall dependence as measured by Spearman's rho, and therefore this measure performs quite well for all scenarios.

Two typical monitoring charts are shown in Figure 3. The black vertical lines correspond to the point at which the normal copula and the Gumbel copula parameters change from 0.5 to 0.25 and from 1.2 to 1.5, respectively. Two horizontal gray lines correspond to the 95% confidence intervals for each monitoring chart obtained from the training data set. For the testing set, the overall dependence-based monitoring chart goes below the lower bound for the normal copula (above the upper bound for the Gumbel copula) indicating weaker (stronger, respectively) overall dependence for the new distribution. There are also some detections in the T^2 density level charts, but the proportion of these detections is close to the nominal level of 5%.

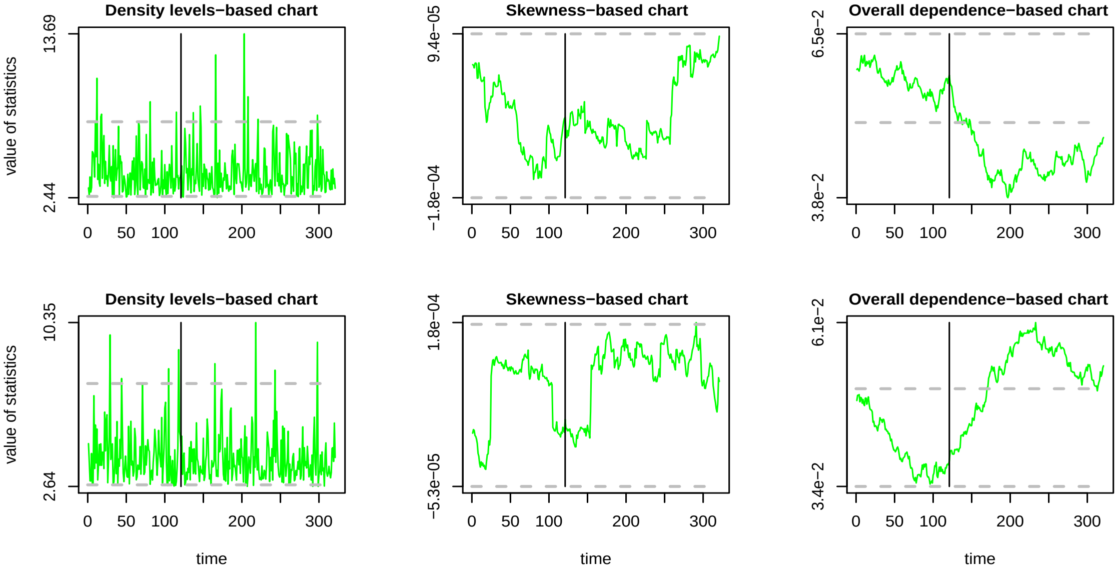

Figure 3: Monitoring results for density levels-based chart (left), skewness-based chart (middle) and overall dependence-based chart (right). The 95% confidence intervals (gray) are constructed from the training set: normal copula with parameter 0.5 (top) and Gumbel copula with parameter 1.2 (bottom). The copula parameter changes to 0.25 for the normal copula and to 1.5 for the Gumbel copula for the testing set (to the right of the black vertical line). Univariate marginals are $N(0,1)$ for both the training set and for the testing set.

4.2. Change in shape, no change in overall dependence

We investigate the performance of the three monitoring charts from the previous section when there is no change in the overall dependence of the testing data set, that is, when the Spearman's rho for all pairs of the observed variables in $\mathbf{Z}$ does not change. We again assume that vector $\mathbf{Z}$ has three components so that the corresponding copula is trivariate (similar results are obtained for dimensions $m = 2, 3, \ldots, 10$), and we use $T_0 = 200$ for the training data and $T_1 - T_0 = 200$ for the testing data.

We select copula parameters such that the Spearman's correlations for each pair of variables equal 0.5, both for the training set and for the testing set. We consider three cases: when the normal copula changes to the Gumbel copula, when the normal copula changes to the Clayton copula, and when the Gumbel copula changes to the Clayton copula. As before, we consider cases when the marginal distributions remain the same and when there is a shift in mean for all individual components. We calculate the number of detections in different scenarios for the three monitoring procedures, and we use $K = 80$ to obtain estimates of ζ_2, ζ_7, and the corresponding 95% confidence intervals. We use the three scenarios:

1. Change in the copula density, no shift in the mean;
2. Change in the copula density, shift in the mean (new mean $\mu = 1$);

Table 2: Average proportion of detections for three testing procedures for different scenarios with copula parameters shown in parentheses. For each scenario, 1000 data sets were generated to calculate these values.

Training data set	Testing data set	Density levels-based chart	Skewness-based chart	Overall dependence-based chart
Normal (0.50) + $N(0,1)$ marginals	Gumbel (1.50) + $N(0,1)$ marginals	0.07	0.37	0.09
Normal (0.50) + $N(0,1)$ marginals	Clayton (1.00) + $N(0,1)$ marginals	0.07	0.51	0.10
Gumbel (1.50) + $N(0,1)$ marginals	Clayton (1.00) + $N(0,1)$ marginals	0.08	0.80	0.08
Normal (0.50) + $N(0,1)$ marginals	Gumbel (1.50) + $N(1,1)$ marginals	0.11	0.36	0.12
Normal (0.50) + $N(0,1)$ marginals	Clayton (1.00) + $N(1,1)$ marginals	0.14	0.52	0.17
Gumbel (1.50) + $N(0,1)$ marginals	Clayton (1.00) + $N(1,1)$ marginals	0.19	0.80	0.18
Normal (0.50) + $N(0,1)$ marginals	Gumbel (1.50) + $N(-1,1)$ marginals	0.14	0.39	0.16
Normal (0.50) + $N(0,1)$ marginals	Clayton (1.00) + $N(-1,1)$ marginals	0.12	0.49	0.14
Gumbel (1.50) + $N(0,1)$ marginals	Clayton (1.00) + $N(-1,1)$ marginals	0.14	0.81	0.10

3. Change in the copula density, shift in the mean (new mean $\mu = -1$).

For each scenario, we generate 1000 data sets to estimate the proportion of detections for the testing data set. The results are presented in Table 2.

We can see that density levels-based and overall dependence-based charts cannot detect changes in the shape of the density f_0; but the skewness-based measure is more sensitive to these changes, with a detection rate ranging from 36% to 81%, depending on how strong the change of the underlying copula density c_0 is. Two typical monitoring charts are shown in Figure 4. The black vertical lines correspond to the points at which the normal copula with parameter 0.5 changes to the Gumbel copula with parameter 1.5 (top) and to the Clayton copula with parameter 1 (bottom). Two horizontal gray lines correspond to the 95% confidence intervals for each monitoring procedure obtained from the training data set. For the testing set, the skewness-based monitoring chart goes above the upper bound (below the lower bound) indicating that the new distribution is skewed to the right (left, respectively).

From the plot in the top middle of Figure 4, it can be seen that the first signal is given by this chart at around sample 240 and goes back below the upper bound again at around sample 270. The missed detection regions of the former chart might be due to the large variability when the moving window is not large. Indeed, with a larger moving window, variability is less, but it can take longer for the chart to detect any changes in the distribution.

4.3. Real data application: Monitoring of a decentralized wastewater treatment plant at Golden, USA

Monitoring of treatment processes is needed not only for evaluating the process operating conditions, but also for inspecting product quality. The developed monitoring charts are illustrated here with data collected from the Mines Park Water Reclamation Test Site, a decentralized wastewater treatment plant in Golden,

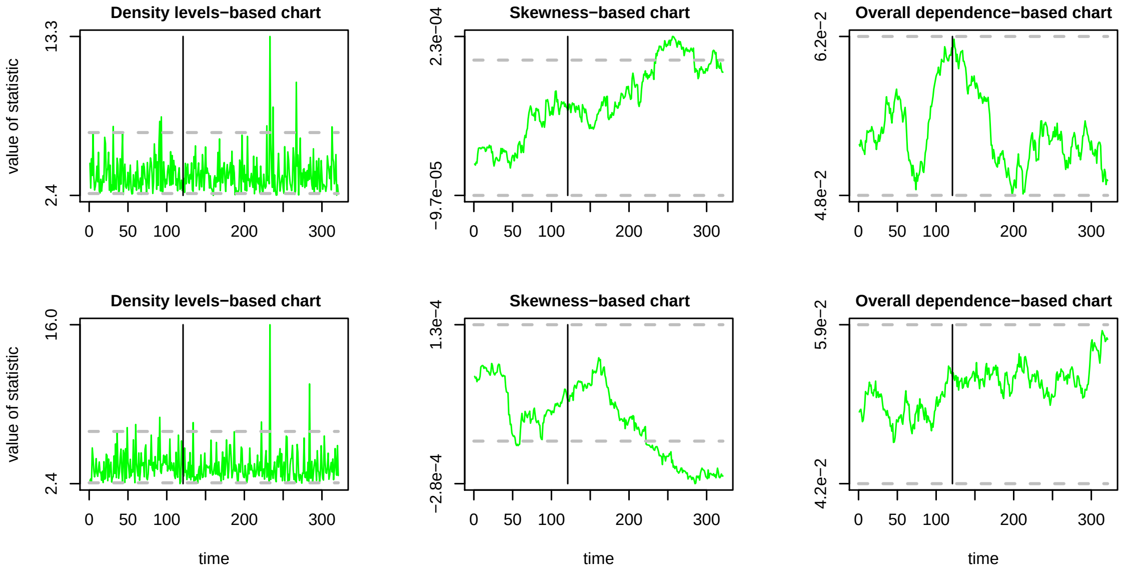

Figure 4: Monitoring results for density levels-based chart (left), skewness-based chart (middle), and overall dependence-based chart (right). The 95% confidence intervals (dashed gray lines) are constructed from the training set (Normal copula with parameter 0.5), and the copula changes to the Gumbel copula with parameter 1.5 (top) and to the Clayton copula with parameter 1 (bottom) for the testing set to the right of the black vertical line. Univariate marginals are $N(0, 1)$ both for the training set and for the testing set.

CO. Wastewater treatment processes are extremely important for community health, removing pollutants from wastewater so that it can be reused or safely discharged. Decentralized facilities such as this one will become more common in the future as communities begin to reuse their wastewater locally, and with greater influent variability, quality, and quantity, decentralized systems require tighter control and faster response to changes or malfunctioning.

This dataset includes 28 variables for which ten minute averages were recorded from April 10, 2010, to May 10, 2010, resulting in a total of 4,464 observations. During this 31-day period, a fault is known to have occurred that affected both pH and salinity, and the microoganisms in the system took over two months to recover. Here, we select some key variables involved in the process to monitor, namely a subset of the following six variables: membrane bioreactor (MBR) permeate pressure, permeate turbidity, return activated sludge (RAS) dissolved oxygen content, RAS pH, RAS total suspended solids, and permeate tank conductivity. There are two MBRs in this particular system, and since they operate similarly, we only monitor the second one here. The RAS contains active biological organisms that do much of the work of eliminating waste from the water. The design of the system is described in detail in Vuono et al. (2013).

Note that the proposed monitoring charts assume that the data are uncorrelated. However, the data collected from the inspected MBR system show serial dependency. Time-series models are commonly utilized for modelling autocorrelated processes. If the model is adequate, the residuals are approximately uncorrelated and can be monitored using the developed charts. Therefore, different non-seasonal autoregressive (AR) models have been fit to each variable. We use the first 500 fault-free observations shown in Figure 5 as training data to build AR models. An AR(1), AR(3), AR(13), AR(12), AR(12), and an AR(2) adequately capture the autocorrelative structure in a time series of the MBR 2 permeate pressure, permeate turbidity, RAS dissolved oxygen content, RAS pH, RAS total suspended solids, and permeate tank conductivity, respectively. To find the best AR model to fit each variable, Box and Jenkins's methodology (Box and Jenkins, 1976) was used to analyze the simple autocorrelation and partial autocorrelation functions of each time series. The maximum likelihood estimates of nonzero AR coefficients with their sampling standard deviations are shown in Table 3.

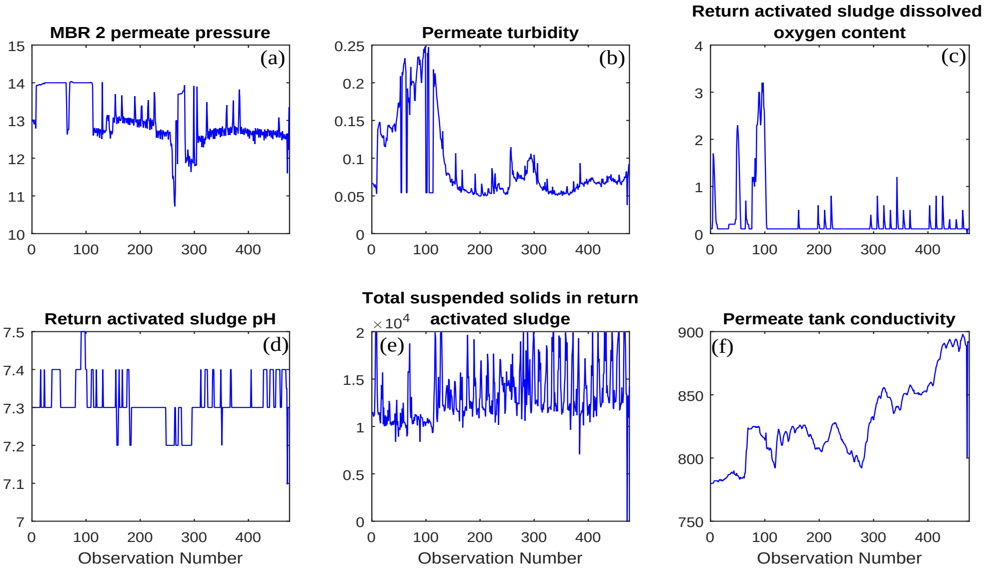

Figure 5: Training data time series of: (a) MBR permeate pressure, (b) permeate turbidity, (c) RAS dissolved oxygen content, (d) RAS pH, (e) RAS total suspended solids, and (f) permeate tank conductivity.

The sample autocorrelation function (ACF) and the sample partial autocorrelation function (PACF) are used to check the presence of serial correlation in the AR-based residuals. The ACF and PACF of the residuals for the selected AR models are presented, respectively, in Figures 6 and 7. Figure 6 indicates that the residuals are not significantly different from a white noise series. It can be seen from Figure 7 that the residuals are approximately uncorrelated.

After reference AR models are identified based on fault-free data, the residuals are used to monitor the abnormal events (faults) in the MBR system that may lead the system to depart from its normal state. To do so, we use three monitoring procedures:

Table 3: The details of the parameter estimations and their associated standard errors for the six time series under study.

MBR process variables	Time series model	Estimated parameters	Coefficient	Standard error
MBR permeate pressure	AR(1)	Constant Lag 1	2.00631 0.845	0.31680 0.02434
Permeate turbidity	AR(3)	Constant Lag 1 Lag 3	0.008433 0.810329 0.098051	0.002292 0.034843 0.034852
RAS dissolved oxygen content	AR(13)	Constant Lag 1 Lag 12 Lag 13	0.023708 0.93640 0.356009 -0.380977	0.009263 0.016134 0.042993 0.042546
RAS pH	AR(12)	Constant Lag 1 Lag 2 Lag 12	0.74819 0.70432 0.07127 0.12219	0.23714 0.04544 0.04512 0.02987
RAS total suspended solids	AR(12)	Constant Lag 1 Lag 12	2602 0.4315 0.3768	519.8 0.0377 0.0379
Permeate tank conductivity	AR(2)	Constant Lag 1 Lag 2	29041 0.58608 40436	6.69599 0.04206 0.04210

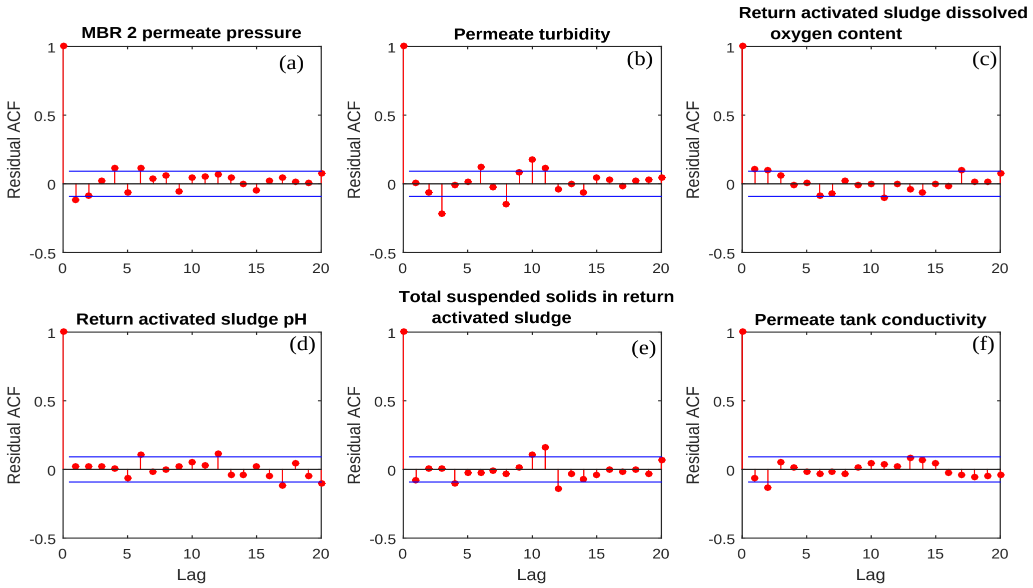

Figure 6: ACF of residual errors: (a) MBR permeate pressure, (b) permeate turbidity, (c) RAS dissolved oxygen content, (d) RAS pH, (e) RAS total suspended solids, and (f) permeate tank conductivity.

1. Negative log-density levels assuming multivariate normality of data (Hotelling T^2 rule);

2. Procedure based on the measure of skewness ζ_7;

3. Procedure based on the measure of overall dependence ζ_2.

For the three monitoring charts, we use the first 500 residuals from the fitted autoregressive models in

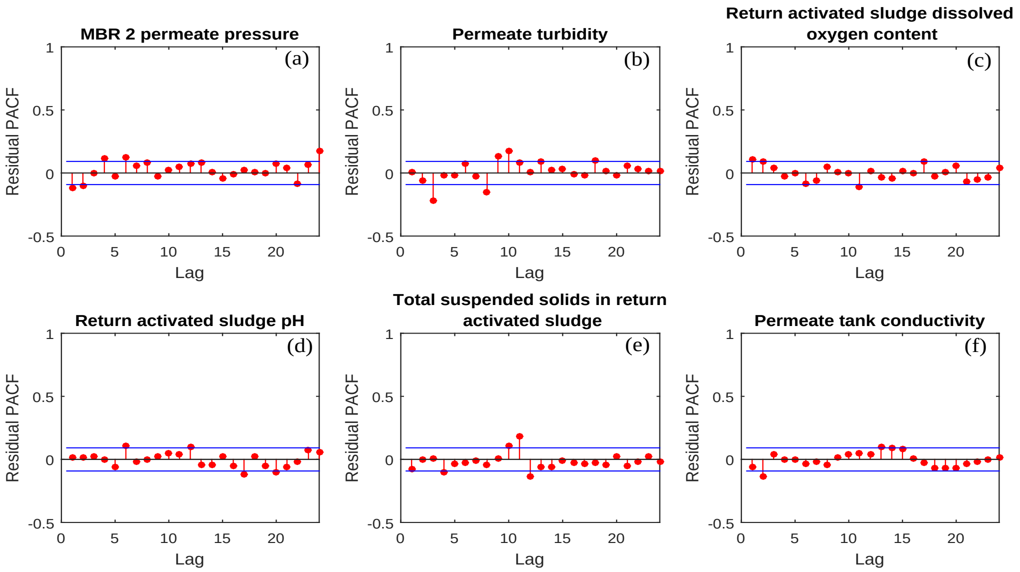

Figure 7: PACF of residual errors: (a) MBR permeate pressure, (b) permeate turbidity, (c) RAS dissolved oxygen content, (d) RAS pH, (e) RAS total suspended solids, and (f) permeate tank conductivity.

Table 3 as training data to construct confidence intervals as explained in Section 3.4. We use the remaining residuals as testing data. For the skewness-based chart and the overall dependence-based chart, we use the moving window length $K = 500$ because there are many observations; with a larger K, the variability of the measures ζ_2 and ζ_7 is smaller, so that these charts are more sensitive to anomalies. The top row of Figure 8 shows monitoring results for the three charts for the last 3,464 observations.

We can see that the Hotelling T^2 and skewness-based charts do not detect any significant anomalies. The proportion of detections when using the Hotelling T^2 is quite close to the nominal level. This suggests that there is no big change in the shape of the joint distribution of the data. However, the overall dependence-based monitoring chart indicates that the overall dependence among the six variables weakens. This confirms findings from Kazor et al. (2016) where the authors monitored the process using all 28 variables. They applied linear and nonlinear dimension reduction methods to the variables first before constructing T^2 and monitoring it based on nonparametric thresholds. In nonparametric T^2, the density of this statistic under normal operating conditions (fault-free) is estimated via kernel density estimation. After obtaining the estimated density, the decision threshold is then set to the corresponding $(1 - \alpha)$-th quantile Kazor et al. (2016). With their method, an alarm would have been issued at 1:40 p.m. on April 21st and then again on April 22nd. Here, our method begins to flag the fault even earlier, on April 18th. The very large deviation below the lower bound in Figure 8 coincides with the initial discovery of the problem by operators on April 24th.

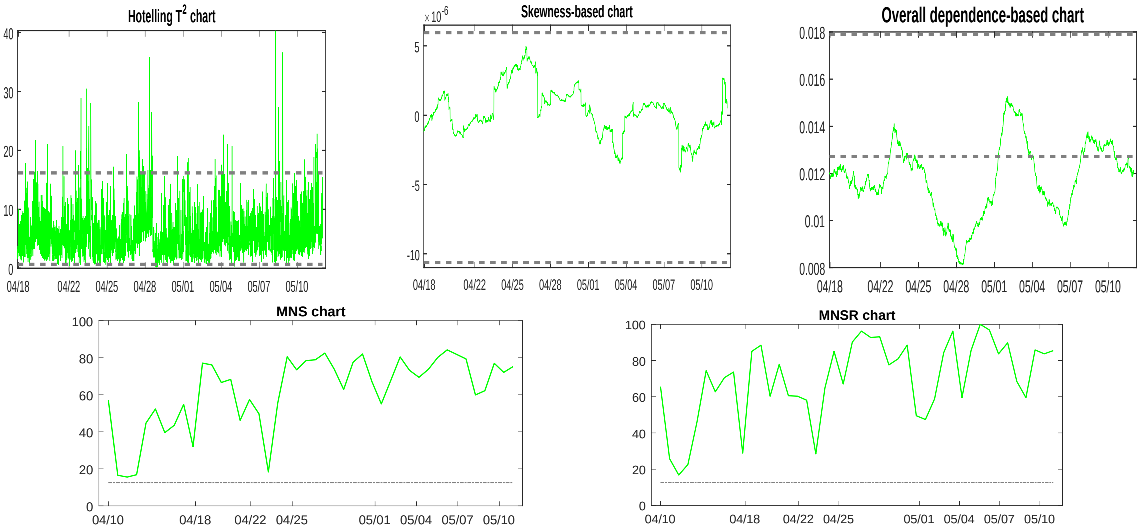

Figure 8: Monitoring results of Hotelling T^2 chart (top left), skewness-based chart (top middle), overall dependence-based chart (top right), the multivariate nonparametric sign (MNS) chart (bottom left), and the multivariate nonparametric signed-rank (MNSR) chart (bottom right). The 95% confidence intervals (dashed gray lines) are constructed from the training set (first 500 observations).

In the bottom row of Figure 8, we also compare the detection efficiency of the proposed copula-based charts to that of two nonparametric multivariate control charts based on the ranking or ordering information of the observed data. Specifically, the first chart is the multivariate nonparametric sign (MNS) chart (Boone and Chakraborti, 2012), which is based on the multivariate form of the sign test (Hettmansperger, 2006), and it uses an asymptotic chi-squared distribution to determine the value of the decision threshold. The second chart is called the multivariate nonparametric signed-rank (MNSR) chart, and it is based on the componentwise Wilcoxon signed-rank statistics (Boone and Chakraborti, 2012). In the MNSR chart, Wilcoxon sign-rank statistics for each component of the multivariate data are considered to design Hotelling's T^2-type chart. We used $n = 100$ consecutive observations to calculate each value of the two test statistics, resulting in 44 test statistics for the two charts. Figure 8 indicates that these two charts can detect the presence of faults but with several false alarms when using the training dataset (first four values). In addition, these charts – like many others – do not provide any information about possible changes in the shape of the underlying distribution.

5. Conclusion

In this paper, we showed that the multivariate control chart based on density level sets cannot usually detect changes in the shape of the distribution for a multivariate process. Moreover, this chart is not sensitive to changes in the mean of individual components if the shape of the distribution also changes. Therefore, we constructed two monitoring procedures for a multivariate process that are sensitive to changes in the shape of the underlying density function. No copula density estimation is needed for these procedures, and therefore, the corresponding control charts are easy to construct. If there is a change in the shape or overall dependence in the joint distribution, these new procedures can provide this information, thus helping to locate the detected abnormality. We applied the proposed monitoring procedures to a real data set. One of the procedures was able to detect anomalies in the data that were consistent with results obtained earlier for the same data set.

The proposed charts are constructed using data from the past and therefore might not detect changes in the distribution very quickly. If an anomaly occurs during a short period of time, the proposed methods might be not very sensitive. Thus, some adjustment is needed to detect these changes more quickly; for example, using exponential downweighting. This is a topic for future research.

Acknowledgment

The research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No: OSR-2015-CRG4-2582. Professor Tzahi Cath of Colorado School of Mines provided the decentralized wastewater treatment data.

References

Basseville, M., Nikiforov, I., 1993. Detection of Abrupt Change: Theory and Application. Prentice Hall, Information and System Sciences Series.

Bissell, D., 1994. Statistical Methods for SPC and TQM. Vol. 26. CRC Press.

Boone, J., Chakraborti, S., 2012. Two simple Shewhart-type multivariate nonparametric control charts. Applied Stochastic Models in Business and Industry 28 (2), 130–140.

Box, G., Jenkins, G., 1976. Time series analysis: Forecasting and control. Holden-Day.

Box, G., Jenkins, G., Reinsel, G., 1994. Time Series Analysis: Forecasting and Control. Prentice-Hall.

Clayton, D. G., 1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65 (1), 141–151.

Crosier, B., 1988. Multivariate generalizations of cumulative sum quality-control schemes. Technometrics 30 (3), 291–303.

Deng, H., Runger, G., Tuv, E., 2012. System monitoring with real-time contrasts. Journal of Quality Technology 44 (1), 9–27.

Fatahi, A., Dokouhaki, P., Moghaddam, B., 2011. A bivariate control chart based on copula function. In: Quality and Reliability (ICQR), 2011 IEEE International Conference on. IEEE, pp. 292–296.

Frank, M. J., 1979. On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. Aequationes Mathematicae 19, 194–226.

Genest, C., Rémillard, B., Beaudoin, D., 2009. Goodness-of-fit tests for copulas: A review and a power study. Insurance: Mathematics and Economics 44 (2), 199–213.

Gertler, J., 1998. Fault Detection and Diagnosis in Engineering Systems. CRC press.

Gumbel, E. J., 1960. Distributions des valeurs extrêmes en plusieurs dimensions. Publ. Inst. Statist. Univ. Paris 9, 171–173.

Han, F., Liu, H., 2013. Principal component analysis on non-Gaussian dependent data. In: Proceedings of the 30th International Conference on Machine Learning. pp. 240–248.

Harrou, F., Fillatre, L., Nikiforov, I., 2014. Anomaly detection/detectability for a linear model with a bounded nuisance parameter. Annual Reviews in Control 38 (1), 32–44.

He, Q., Wang, J., 2007. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes. IEEE Transactions on Semiconductor Manufacturing 20 (4), 345–354.

Hettmansperger, T., 2006. Multivariate location tests. Encyclopedia of Statistical Sciences.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24, 417–441.

Hotelling, H., 1947. Multivariate quality control illustrated by the air testing of sample bomb sights, Techniques of Statistical Analysis, Ch. II.

Isermann, R., 2006. Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance. Springer Science & Business Media.

Joe, H., 2014. Dependence Modeling with Copulas. Chapman & Hall/CRC, Boca Raton, FL.

Kazor, K., Holloway, R. W., Cath, T. Y., Hering, A. S., 2016. Comparison of linear and nonlinear dimension reduction techniques for automated process monitoring of a decentralized wastewater treatment facility. Stochastic Environmental Research and Risk Assessment 30 (5), 1527–1544.

Kojadinovic, I., Yan, J., 2011. A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems. Statistics and Computing 21, 17–30.

Krupskii, P., 2016. Copula-based measures of reflection and permutation asymmetry and statistical tests. Statistical Papers, 1–23.

Kuvattana, S., Sukparungsee, S., Busababodhin, P., Areepong, Y., 2015. Efficiency of bivariate copula on the CUSUM chart. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. Vol. 2.

Liu, R., Singh, K., Teng, J., 2004. Ddma-charts: Nonparametric multivariate moving average control charts based on data depth. Allgemeines Statistisches Archiv 88 (2), 235–258.

Lowry, C. A., Woodall, W. H., Champ, C. W., Rigdon, S. E., 1992. A multivariate exponentially weighted moving average control chart. Technometrics 34 (1), 46–53.

Lucas, J., Saccucci, M., 1990. Exponentially weighted moving average control schemes: Properties and enhancements. Technometrics 32 (1), 1–12.

Montgomery, D. C., 2005. Introduction to Statistical Quality Control. John Wiley & Sons, New York.

Nelsen, R. B., 2006. An Introduction to Copulas, 2nd Edition. Springer, New York.

Qiu, P., 2013. Introduction to Statistical Process Control. CRC Press.

Rabhu, S., Runger, G., 1997. Designing a multivariate EWMA control chart. Journal of Quality Technology 29 (1), 8–15.

Russell, E., Chiang, L., Braatz, R., 2012. Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes. Springer Science & Business Media.

Shao, J., Wu, C. F. J., 1989. A general theory for jackknife variance estimation. Annals of Statistics 17(3), 1176–1197.

Shewhart, W., 1930. Economic quality control of manufactured product. Bell System Technical Journal 2, 364–389.

Sklar, A., 1959. Fonctions de répartition à n dimensions et leurs marges. Institute of Statistics of the University of Paris 8, 229–231.

Sukchotrat, T., Kim, S., Tsung, F., 2009. One-class classification-based control charts for multivariate process monitoring. IIE Transactions 42 (2), 107–120.

Verdier, G., 2013. Application of copulas to multivariate control charts. Journal of Statistical Planning and Inference 143 (1), 2151–2159.

Vuono, D., Henkel, J., Benecke, J., Cath, T. Y., Reid, T., Johnson, L., Drewes, J. E., 2013. Flexible hybrid membrane treatment systems for tailored nutrient management: A new paradigm in urban wastewater treatment. Journal of Membrane Science 446, 34–41.

Yin, S., Ding, S. X., Xie, X., Luo, H., 2014. A review on basic data-driven approaches for industrial process monitoring. Industrial Electronics, IEEE Transactions on 61 (11), 6418–6428.

Zhang, C., Chen, N., Zou, C., 2016. Robust multivariate control chart based on goodness-of-fit test. Journal of Quality Technology 48 (2), 139–161.