A General Framework for Monitoring Mixed Data

Daniele Zago (daniele.zago.1@phd.unipd.it)
Department of Statistical Sciences, University of Padua, Padua, Italy
and
Zibo Tian (zibo.tian@ufl.edu)
Department of Biostatistics, University of Florida, Gainesville, USA
and
Giovanna Capizzi (capizzi@stat.unipd.it)
Department of Statistical Sciences, University of Padua, Padua, Italy
and
Peihua Qiu* (pqiu@ufl.edu)
Department of Biostatistics, University of Florida, Gainesville, USA

May 20, 2025

Abstract

Modern applications of statistical process monitoring involve checking the stability of multivariate processes with mixed data types, such as a combination of continuous, ordinal, and categorical quality variables. Appropriate statistical modeling for such data is often challenging, especially when the observed data are serially correlated, which explains why there is only a limited existing discussion on sequential monitoring of processes with mixed data. This paper introduces a general methodology to solve the problem. The main idea behind our approach is to sequentially transform the original observed data into continuous data through innovative data pre-processing, achieved by encoding the ordinal and categorical variables into continuous numerical variables using dummy and score variables and data transformation and decorrelation. Numerical studies show that the proposed method is effective in monitoring mixed data, in comparison with some state-of-the-art existing methods. The new method is illustrated in a case study involving online monitoring of hotel customers' behaviors. Computer codes in Julia for implementing the proposed methodology are provided in the supplemental material.

Keywords: Data decorrelation; Heterogeneous data; Mixed data; Self-starting charts; Statistical process monitoring; Transformation.

*corresponding author

1 Introduction

Statistical process control (SPC) charts provide a main analytic tool for online process monitoring that has broad applications in manufacturing, healthcare, environmental monitoring, and more (Montgomery, 2008; Qiu, 2013). Most existing SPC charts are designed for monitoring either continuous numerical quality variables or categorical quality variables, but not both. In practice, however, it is common to have both continuous (e.g., air temperature) and categorical (e.g., sunny, rainy, or cloudy weather) variables involved. This paper focuses on online monitoring of sequential processes with both continuous and categorical quality variables.

Traditional SPC charts assume that in-control (IC) process observations are independent and identically distributed (i.i.d.) and follow a parametric (e.g., normal) distribution (e.g., Crosier, 1988; Lowry et al., 1992). In the SPC literature, it has been well demonstrated that such charts are unreliable to use when their model assumptions are violated (Qiu and Hawkins, 2001; Apley and Lee, 2008; Capizzi and Masarotto, 2009). To handle cases when the normality assumption is violated, many distribution-free or nonparametric SPC charts have been developed based on ranks (Qiu and Hawkins, 2001; Zou and Tsung, 2010; Li et al., 2017; Chakraborti and Graham, 2019) or data categorization (Qiu, 2008; Li et al., 2012; Wang et al., 2017). While these charts can be used without strong parametric distributional assumptions, they only use partial (e.g., ranking) information in the observed data, reducing their effectiveness in detecting changes in the process distribution (Xie and Qiu, 2022). To monitor processes with serially correlated observations, many control charts have been developed using either parametric time series modeling (Apley and Lee, 2008; Capizzi and Masarotto, 2008; Lee and Apley, 2011; Prajapati and Singh, 2012) or nonparametric moment estimation and data decorrelation (Xue and Qiu, 2021; Qiu and Xie, 2022; Xie and Qiu, 2024).

All the SPC charts mentioned above are for monitoring processes with continuous quality variables only. In practice, however, there are many applications involving both con-

tinuous and categorical quality variables. Some examples are given below. In logistics, numerical data regarding the volume of freight managed by a contractor is often recorded alongside categorical quality variables like container type (Ning and Tsung, 2010, 2012). In manufacturing, ordinal categorical variables such as the qualitative leakage of material after welding are usually used in conjunction with some numerical product characteristics for improving the quality of manufactured products (Ding et al., 2016b). In meteorological applications, continuous variables like air temperature and ozone level are typically recorded along with ordinal (e.g. air quality indices) and categorical (e.g., presence or absence of extreme events) data. In healthcare, continuous, categorical, and ordinal data are usually recorded and monitored together for each admitted patient (Johnson, 2023). Additionally, there are many cases in practice when numerical quality variables are recorded as ordinal variables to reduce the cost of data collection (Tucker et al., 2002).

In the SPC literature, the problem to monitor processes with mixed data of both continuous and categorical quality variables has received limited attention due mainly to the challenges in modelling the mixed data properly. Existing methods to solve this problem typically require assumed latent structure on the observed ordinal categorical variables for monitoring processes with both continuous and ordinal categorical quality variables (Ning and Tsung, 2010; Ding et al., 2016a,b; Wang et al., 2017, 2018), or are constructed based on density estimation, multiple comparisons, and other statistical methods for monitoring processes with both continuous and categorical quality variables (Ning and Tsung, 2010, 2012; Huang et al., 2023). These methods either impose extra structure on the observed data or make memoryless decisions by Shewhart charts based on the i.i.d. and other model assumptions.

In this paper, we propose a general framework for the online monitoring of processes that involve mixed data, including continuous numerical, ordinal categorical, and nominal categorical quality variables. The core idea is briefly outlined as follows: In situations where all quality variables are numerical, flexible SPC charts have been developed for monitoring multivariate processes. These charts can handle cases where IC distributions do not con-

form to any parametric distribution families and observations might be serially correlated (cf., Wang et al., 2024; Xie and Qiu, 2022). Building on this, a natural approach is to transform all categorical quality variables into numerical variables without significant loss of information from the original data. For ordinal categorical variables, we recommend using rank transformation to convert them into numerical variables. For nominal categorical variables, we suggest replacing them with corresponding dummy variables. Once all categorical quality variables have been transformed into numerical variables, the transformed data should be pre-processed to remove serial correlation and further transformed to achieve a normal distribution, as suggested by Xie and Qiu (2022). Subsequently, a multivariate control chart can be applied to the transformed and pre-processed data for online process monitoring.

Our proposed method offers an effective way to monitor processes with mixed data without imposing additional structure on the categorical quality variables, as commonly done in previous literature (e.g., Wang et al., 2018). It also accommodates serial data correlation and complex data distributions. Numerical studies presented in this paper confirm that this method is effective in detecting mean shifts in processes with mixed data, outperforming some representative existing methods.

The remainder of the paper is organized as follows. In Section 2, our proposed methodology is described in detail. In Section 3, simulation studies are presented to evaluate its numerical performance, in comparison with some representative existing methods. The proposed method is illustrated in a case study about online monitoring of hotel customers' behaviors in Section 4. Some remarks conclude the paper in Section 5.

2 Methodology

At time $t \geq 1$, let $\boldsymbol{X}_t$ be the observation of p heterogeneous quality variables of the process under monitoring, and

$$\boldsymbol{X}_t = (\boldsymbol{Y}_t^\top, \boldsymbol{O}_t^\top, \boldsymbol{C}_t^\top)^\top, \tag{1}$$

where $\boldsymbol{Y}_t$ denotes the observation of a vector of d_Y continuous quality variables, $\boldsymbol{O}_t$ denotes the observation of a vector of d_O ordinal categorical quality variables with attribute levels being $\{h_{O,k}, k = 1, \ldots, d_O\}$, and $\boldsymbol{C}_t$ denotes the observation of a vector of d_C nominal categorical quality variables with the numbers of categories being $\{h_{C,l}, l = 1, \ldots, d_C\}$. The goal of our proposed method is to sequentially monitor the process $\{\boldsymbol{X}_t, t \geq 1\}$ and give a signal once a distributional shift is detected.

The core concept of our proposed method is to convert observations of ordinal categorical quality variables, $\boldsymbol{O}_t$, and nominal categorical quality variables, $\boldsymbol{C}_t$, into numerical data. To achieve this, we suggest replacing each element in $\boldsymbol{O}_t$ with its rank among all ordinal categories of the related variable. As a result, $\boldsymbol{O}_t$ can be relaced by $\boldsymbol{R}_t \in \{1, \ldots, h_{O,1}\} \times \cdots \times \{1, \ldots, h_{O,d_O}\}$.

For each nominal categorical variable in $\boldsymbol{C}_t$, we recommend substituting it with dummy variables. Specifically, consider the lth element of $\boldsymbol{C}_t$, which has $h_{C,l}$ categories. This element can be replaced by $h_{C,l} - 1$ dummy variables without losing any information. Let's assume the first category is chosen as the reference category. The $h_{C,l} - 1$ dummy variables can be defined as follows: the first dummy variable is set to 1 if the original categorical variable takes the value of the second category and 0 otherwise; the second dummy variable is 1 if the original variable takes the value of the third category and 0 otherwise, and so on. Then, $\boldsymbol{C}_t$ can be relaced by the vector of dummy variables, $\boldsymbol{I}_t \in \{0,1\}^{h_{C,1}-1} \times \cdots \times \{0,1\}^{h_{C,d_C}-1}$.

There are different methods to define dummy variables for a given nominal categorical variable. For instance, instead of choosing the first category as the reference, the last category could be selected as the reference. In this case, the first dummy variable would be 1 if the original variable takes the value of the first category and 0 otherwise, and so on. It is straightforward to verify that these two methods are equivalent in representing the observed data of the original categorical variable (Agresti, 2013). In most software packages (e.g., R), the default method for defining dummy variables is to use the first category as the reference.

As an example, consider an ordinal categorical variable with four possible values: "Low," "Medium," "Medium High," and "High," and a nominal categorical variable with three possible categories: "Black," "White," and "Red." At a given time point, let's assume the first variable has the observation "Medium," and the second variable takes the value "Red." In this example, the first variable is replaced by a numerical value of 2, representing its rank among the ordinal categories. The second variable is replaced by two dummy variables. When "Black" is chosen as the reference category, the dummy variables take the values (0, 1) at this time point, indicating the selection of "Red."

Our proposed online monitoring procedure is constructed in the transparent sequential learning framework (Qiu and Xie, 2022), and requires an initial IC dataset of $m > 0$ observations available in advance, which is denoted as $\mathcal{X}^{(0)} = \{\boldsymbol{X}_{-m+1}, \ldots, \boldsymbol{X}_0\}$. By the transformation described above, the tth observation in this dataset can be transformed into the numerical vector:

$$\boldsymbol{X}_t^* = (\boldsymbol{Y}_t^\top, \boldsymbol{R}_t^\top, \boldsymbol{I}_t^\top)^\top, \qquad \text{for } t = -m+1, -m+2, \ldots, 0,$$

where $\boldsymbol{X}_t^*$ is a d^*-dimensional vector with $d^* = d_Y + d_O + \sum_{l=1}^{d_C}(h_{C,l} - 1)$.

It is assumed in this paper that serial correlation in $\mathcal{X}^{(0)*} = \{\boldsymbol{X}_t^*, t = -m+1, -m+2, \ldots, 0\}$ is stationary. Namely, $\gamma(s) = \text{Cov}(\boldsymbol{X}_t^*, \boldsymbol{X}_{t+s}^*)$ depends on s only. In addition, the correlation structure is assumed to be short-ranged, namely, $\gamma(s) \approx 0$, for $s > b_{\max}$, where $b_{\max}$ denotes the range of autocorrelation. Under these assumptions, the IC mean $\boldsymbol{\mu}$ and the set of IC covariance matrices $\{\boldsymbol{\gamma}(s), 0 \le s \le b_{\max}\}$ can be initially estimated from the IC dataset by their moment estimates as follows:

$$\begin{aligned} \widehat{\boldsymbol{\mu}}^{(0)} &= \frac{1}{m}\sum_{i=-m+1}^{0} \boldsymbol{X}_i^*, \\ \widehat{\boldsymbol{\gamma}}^{(0)}(s) &= \frac{1}{m-s}\sum_{i=-m+1}^{-s} \left(\boldsymbol{X}_i^* - \widehat{\boldsymbol{\mu}}^{(0)}\right)\left(\boldsymbol{X}_{i+s}^* - \widehat{\boldsymbol{\mu}}^{(0)}\right)^\top, \quad \text{for } 0 \le s \le b_{\max}. \end{aligned} \tag{2}$$

It should be pointed out that the stationarity assumption would be reasonable in many

applications (e.g., manufacturing applications). In cases when this assumption is violated, the kernel estimation of the covariance matrices discussed in Xie and Qiu (2023) can be considered in place of the moment estimates defined in (2). In many applications, it is reasonable to assume that correlation between two process observations is weaker when the two observation times are farther away and thus the short-range autocorrelation assumption is also reasonable. See Qiu and You (2022) and Xie and Qiu (2023) for some real-data examples.

After calculating the IC parameter estimates, the initial IC data $\mathcal{X}^{(0)*}$ is then standardized and decorrelated using the algorithm similar to the one in Qiu and You (2022). More specifically, let $\boldsymbol{W}_i = ((\boldsymbol{X}^*_{i-b})^\top, (\boldsymbol{X}^*_{i-b+1})^\top, \ldots, (\boldsymbol{X}^*_i)^\top)^\top$ be the long vector of the observation $\boldsymbol{X}^*_i$ and all its previous observations that need to be decorrelated with $\boldsymbol{X}^*_i$, where $b = \min\{i+m-1, b_{\max}\}$ and $-m+1 \le i \le 0$. Then, the variance-covariance matrix $\mathrm{Cov}(\boldsymbol{W}_i, \boldsymbol{W}_i)$ can be written as

$$\Sigma_{i,i} = \begin{pmatrix} \widehat{\boldsymbol{\gamma}}^{(0)}(0) & \cdots & \widehat{\boldsymbol{\gamma}}^{(0)}(b) \\ \vdots & \ddots & \vdots \\ \widehat{\boldsymbol{\gamma}}^{(0)}(b)^\top & \cdots & \widehat{\boldsymbol{\gamma}}^{(0)}(0) \end{pmatrix} = \begin{pmatrix} \Sigma_{i-1,i-1} & \Sigma_{i-1,i} \\ \Sigma^\top_{i-1,i} & \widehat{\boldsymbol{\gamma}}^{(0)}(0) \end{pmatrix},$$

and the standardized and decorrelated observation at time i is defined to be

$$\boldsymbol{X}^{**}_i = \begin{cases} \widehat{\boldsymbol{\gamma}}^{(0)}(0)^{-1/2}\left(\boldsymbol{X}^*_i - \boldsymbol{\mu}^{(0)}\right), & \text{if } i = -m+1, \\ D_i^{-1/2}\left(\boldsymbol{X}^*_i - \boldsymbol{\mu}^{(0)} - \Sigma^\top_{i-1,i}\Sigma^{-1}_{i-1,i-1}\widehat{\boldsymbol{e}}_{i-1}\right), & \text{if } i > -m+1, \end{cases} \tag{3}$$

where $\widehat{\boldsymbol{e}}_{i-1} = \boldsymbol{W}_{i-1} - \boldsymbol{\mu}^{(0)}$ and $D_i = \widehat{\boldsymbol{\gamma}}^{(0)}(0) - \Sigma^\top_{i-1,i}\Sigma^{-1}_{i-1,i-1}\Sigma_{i-1,i}$. As pointed out by Xie and Qiu (2024), when the IC sample size m is small, the inverse matrices $\Sigma^{-1}_{i-1,i-1}$, $\widehat{\boldsymbol{\gamma}}^{(0)}(0)^{-1/2}$, and $D_i^{-1/2}$ may not exist. In such cases, a matrix modification is needed to make these matrices positive semidefinite. In this paper, we suggest using the matrix modification method discussed in Higham (1988) to modify the related matrices to positive semidefinite matrices, which can be implemented using the function **nearPD()** in the

R-package **Matrix**.

It can be checked that the standardized and decorrelated observation $\boldsymbol{X}_i^{**}$, for each i, would have the asymptotic mean $\mathbf{0}$ and asymptotic identity covariance matrix. In addition, the components of $\boldsymbol{X}_i^{**}$ would be asymptotically independent. Next, the following data transformation discussed originally in Xie and Qiu (2022) is considered. Let $F_j(\cdot)$ be the IC cumulative distribution function (cdf) of the jth decorrelated quality variable, for $j = 1, 2, \ldots, d^*$. Then, these cdf's can be estimated by the following empirical cdf's:

$$\widehat{F}_j^{(0)}(x) = \frac{1}{m} \sum_{i=-m+1}^{0} I(X_{ij}^{**} \leq x), \quad \text{for } j = 1, 2, \ldots, d^*,$$

where $I(A)$ is the indicator function of the event A, and X_{ij}^{**} denotes the jth element of $\boldsymbol{X}_i^{**}$. Then, the following Rosenblatt transformation (Rosenblatt, 1952; Nataf, 1962) is applied to the standardized and decorrelated data:

$$Z_{ij} = \Phi^{-1}\left[\widehat{F}_j^{(0)}(X_{ij}^{**})\right], \quad \text{for } i = -m+1, -m+2, \ldots, 0, \quad j = 1, 2, \ldots, d^*, \tag{4}$$

where Φ^{-1} denotes the inverse of the cdf of a standard normal distribution. Note that Equation (4) ensures that scale differences among different variables are accommodated well so that each of the d^* transformed variables has the asymptotic mean 0 and the asymptotic variance 1 under some regularity conditions.

As a side note, some researchers (e.g., Shen et al., 2016) pointed out that the distribution of the transformed quantity Z_{ij} by (4) may not be well approximated by a standard normal distribution when the original quantity X_{ij}^{**} is discrete. For the SPC problem focused in this paper, the main issue caused by this phenomenon is that the IC mean of the transformed data $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{id^*})^\top$ could be substantially different from zero. Consequently, control charts designed for monitoring the process mean could have the IC performance substantially different from what is expected. This issue, however, is not new in the SPC literature. In developing nonparametric control charts by using ranks or data categorization, the charting statistics are often discrete and the related charts can hardly

reach a pre-specified ARL_0 value. To overcome this difficulty, Qiu (2008) suggested adding small random numbers to some related discrete quantities used in computing the charting statistic value to reduce the discreteness of the charting statistic. This idea can also be used here. More specifically, small random numbers generated independently from the $N(0, \xi^2)$ distribution can be added to each element of $\boldsymbol{R}_t$ and $\boldsymbol{I}_t$ before data decorrelation and transformation, where $\xi > 0$ is a small number. In all numerical examples presented in Sections 3 and 4, ξ is chosen to be 0.01. After this small modification, all quantities in $\mathcal{X}^{(0)*}$ would have no ties, and the mean of the transformed data $\boldsymbol{Z}_t$ would be approximately zero.

Next, we discuss how to monitor the process $\{\boldsymbol{X}_t, t \geq 1\}$ (cf., (1)) by a self-starting chart. At the current observation time t, the observed mixed data $\boldsymbol{X}_t$ are first converted into numerical data $\boldsymbol{X}_t^*$. Then, $\boldsymbol{X}_t^*$ needs to be standardized and decorrelated with all previous data, as discussed above for the initial IC data. The standardized and decorrelated observation $\boldsymbol{X}_t^{**}$ is then transformed by the Rosenblatt transformation as follows:

$$\boldsymbol{Z}_t = \left(\Phi^{-1}[\widehat{F}_1^{(t-1)}(X_{t1}^{**})], \Phi^{-1}[\widehat{F}_2^{(t-1)}(X_{t2}^{**})], \ldots, \Phi^{-1}[\widehat{F}_{d^*}^{(t-1)}(X_{td^*}^{**})]\right), \tag{5}$$

where $\widehat{F}_j^{(t-1)}$ is the empirical cdf of the j-th standardized and decorrelated variable computed from the IC data at time $t-1$ (see Equation (7) below).

After the data transformation by (5), the random vector $\boldsymbol{Z}_t$ should have the asymptotic joint distribution $N(\boldsymbol{0}, I_{d^* \times d^*})$. Then, a conventional multivariate chart can be applied to $\boldsymbol{Z}_t$ for online process monitoring. For this purpose, many existing multivariate SPC charts can be considered (e.g., Huang and Yeh, 2024; Qiu, 2013, Chapter 7). In this paper, we use the MCUSUM chart suggested by Crosier (1988) whose charting statistic is defined as follows. Let

$$\boldsymbol{S}_t = \begin{cases} \boldsymbol{0}, & \text{if } C_t \leq k, \\ (1 - k/C_t)(\boldsymbol{S}_{t-1} + \boldsymbol{Z}_t), & \text{if } C_t > k, \end{cases}$$

where $k > 0$ is a pre-specified allowance, and

$$C_t = \sqrt{(\boldsymbol{S}_{t-1} + \boldsymbol{Z}_t)^\top (\boldsymbol{S}_{t-1} + \boldsymbol{Z}_t)}.$$

Then, the chart gives a signal when

$$Y_t = \sqrt{\boldsymbol{S}_t^\top \boldsymbol{S}_t} > \rho, \tag{6}$$

where $\rho > 0$ is a control limit chosen to achieve a pre-specified ARL_0 value. The chart (6) is called R-MCUSUM chart hereafter, where the first letter "R" indicates the use of the Rosenblatt transformation in Equation (5).

In cases when the R-MCUSUM chart does not give a signal at the current time t, the process under monitoring is declared to be IC and the current observation $\boldsymbol{X}_t$ is combined with the IC dataset. The IC data at time t is denoted as $\mathcal{X}^{(t)} = \{\boldsymbol{X}_{-m+1}, \ldots, \boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t\}$, and the numerical version is denoted as $\mathcal{X}^{(t)*} = \{\boldsymbol{X}^*_{-m+1}, \ldots, \boldsymbol{X}^*_0, \boldsymbol{X}^*_1, \ldots, \boldsymbol{X}^*_t\}$. Then, estimates of the IC quantities can be updated recursively as follows: for $j = 1, 2, \ldots, d^*$ and $s = 0, 1, \ldots, b_{\max}$,

$$\begin{aligned}
\widehat{F}_j^{(t)}(x) &= \frac{m+t-1}{m+t}\widehat{F}_j^{(t-1)}(x) + \frac{1}{m+t} I(X^{**}_{tj} \le x), \\
\widehat{\boldsymbol{\mu}}^{(t)} &= \frac{m+t-1}{m+t}\widehat{\boldsymbol{\mu}}^{(t-1)} + \frac{1}{m+t}\boldsymbol{X}^*_t, \\
\widehat{\boldsymbol{\gamma}}^{(t)}(s) &= \frac{m+t-s-1}{m+t-s}\widehat{\boldsymbol{\gamma}}^{(t-1)}(s) + \frac{1}{m+t-s}\left(\boldsymbol{X}^*_t - \widehat{\boldsymbol{\mu}}^{(n)}\right)\left(\boldsymbol{X}^*_{t-s} - \widehat{\boldsymbol{\mu}}^{(t)}\right)^\top,
\end{aligned} \tag{7}$$

where X^{**}_{tj} is the jth element of $\boldsymbol{X}^{**}_t$, and $\boldsymbol{X}^{**}_t$ is the standardized and decorrelated observation at time t obtained from $\mathcal{X}^{(t)*}$ in a similar way to that described in (3).

To use the R-MCUSUM chart (6), its control limit ρ should be chosen properly. To this end, numerical methods based on Monte Carlo simulations, such as the ones based on bisection search (see, for example, Qiu, 2013) and stochastic approximations (Capizzi and Masarotto, 2016), can be considered. However, because the standardized and decorrelated process observations $\{\boldsymbol{X}^{**}_t, t \ge 1\}$ would still contain some residual serial correlation and

the data distribution would not be exactly normally distributed, in the current research problem we suggest using the following circular block bootstrap procedure (Politis and Romano, 1992) with the block length $l > b_{\max}$ to determine the value of ρ. Interested readers can see Bühlmann (2002) and Lahiri (2003) for overviews on bootstrap methods for analyzing autocorrelated data. To use the circular block bootstrap procedure, the initial IC data are first wrapped around a circle. Namely, we first define $\boldsymbol{Q}_i = \boldsymbol{Z}_{(i+m) \bmod m}$, for $-m+1 \le i \le 0$, where "$a \bmod b$" indicates "a modulo b", and $\{\boldsymbol{Z}_i, -m+1 \le i \le 0\}$ are defined in (4). Then, a sequence of integers $i_0, i_1, \ldots$ is drawn with replacement from the uniform distribution on the set $\{-m+1, -m+2, \ldots, 0\}$. The process observations generated by the circular block bootstrap procedure for online process monitoring are then defined to be

$$\boldsymbol{Z}^*_{h \cdot l + j} = \boldsymbol{Q}_{i_h + j - 1}, \quad \text{for } j = 1, 2, \ldots, l, \ h = 0, 1, 2, \ldots.$$

Then, the R-MCUSUM chart with a given value of ρ can be applied to each sequence of process observations generated by the circular block bootstrap procedure, and the run length (RL) value can be recorded. This process is then repeated for B times, and the average of the B RL values is used for estimating the ARL_0 value. The ρ value is then searched by the bisection searching algorithm or other alternative algorithms so that a pre-specified ARL_0 value is reached.

It should be pointed out that the computational burden is quite heavy to determine the control limit value ρ by the above procedure using the bisection searching algorithm, because B sequences of process observations need to be generated for each given ρ value and many such ρ values need to be considered in the searching process. To reduce the computational burden, we suggest using a modified bisection searching algorithm to determine ρ, which is described in Appendix A.

In the proposed R-MCUSUM chart, there are a few parameters to choose in advance. The decorrelation window size $b_{\max}$ can usually be determined by the autocorrelation function (ACF) of the initial IC dataset in practice. In the next section, we also report some numerical results about its impact on the R-MCUSUM chart. Once $b_{\max}$ is determined, it

is found that the block size used in the circular block bootstrap procedure can be chosen to be two times $b_{\max}$ to have reasonably good results.

3 Simulation Studies

In this section, we evaluate the numerical performance of the R-MCUSUM chart for monitoring processes with mixed data, compared to three representative existing methods based on data categorization described below. To make the comparison fair, all four charts are applied to the standardized and decorrelated process observations $\{\boldsymbol{X}_t^{**}, t \geq 1\}$.

- The LLCUSUM chart discussed in Xue and Qiu (2021). This chart is constructed by categorizing each quality variable into a binary variable and then describing the relationship among different binary variables by a log-linear model.

- The LLD chart suggested by Li et al. (2012). This chart generalizes the LLCUSUM chart in several aspects. First, it categorizes each quality variable into a categorical variable with multiple categories. Then, an ordinal log-linear model with second-order interactions is considered for describing the joint distribution of all the categorical variables. In this paper, each continuous or ordinal categorical quality variable is categorized into 3 categories to use this chart.

- The MOC chart proposed by Wang et al. (2017) for monitoring processes with ordinal categorical quality variables. This chart employs the same data categorization procedure as the one in the LLD chart, followed by constructing an EWMA chart based on an approximation of the GLR test statistic derived from an estimated log-linear model with the second-order interactions. To use this method, each continuous or ordinal categorical quality variable is categorized into 3 categories in this paper.

In all examples considered in this section, we consider the problem of monitoring $p = 3$ heterogeneous quality variables. Their observations are generated in the following way. First, observations of three continuous variables, denoted as $\{\boldsymbol{S}_t = (S_{t1}, S_{t2}, S_{t3}) \in \mathbb{R}^3, t \geq$

1} are generated. Then, observations of the three heterogeneous quality variables, denoted as $\{\boldsymbol{X}_t = (X_{t1}, X_{t2}, X_{t3}), t \geq 1\}$, are defined to be

$$\boldsymbol{X}_t = \left(S_{t1}, \min\left\{[S_{t2}], 2\right\}, I(S_{t3} > 0)\right), \qquad \text{for } t \geq 1,$$

where $[a]$ denotes the integer part of a. Namely, observations of the first quality variable are just $\{S_{t1}, t \geq 1\}$ which are continuous numerical, observations of the second quality variable are $\{\min\{[S_{t2}], 2\}, t \geq 1\}$ which are ordinal categorical, and observations of the third quality variable are $\{I(S_{t3} > 0), t \geq 1\}$ which are binary. The main consideration to generate observations of the three heterogeneous quality variables in this way is that serial correlation in $\{\boldsymbol{S}_t, t \geq 1\}$ is easier to control, compared to the serial correlation in $\{\boldsymbol{X}_t, t \geq 1\}$.

Then, the following four scenarios with different correlation structures and/or data distributions in $\{\boldsymbol{S}_t, t \geq 1\}$ are considered:

- Case I: S_{t1}, S_{t2} and S_{t3} are independent of each other, $\{S_{t1}, t \geq 1\}$ are generated independently from $N(0, 1)$, and $\{S_{t2}, t \geq 1\}$ and $\{S_{t3}, t \geq 1\}$ are generated independently from the standardized version of the χ^2_2 distribution.

- Case II: Same as Case I, except that $\{S_{t1}, t \geq 1\}$ are generated independently from the standardized version of the t_3 distribution.

- Case III: The vector $\boldsymbol{S}_t$ is defined to be $\boldsymbol{S}_t = C^{1/2}\boldsymbol{\varepsilon}_t$, where each component of $\boldsymbol{\varepsilon}_t$ is generated independently from the standardized version of the χ^2_2 distribution, and

$$C = \begin{pmatrix} 1 & 0.5 & 0.5^2 \\ 0.5 & 1 & 0.5 \\ 0.5^2 & 0.5 & 1 \end{pmatrix}.$$

- Case IV: The vector $\boldsymbol{S}_t$ is defined to be $\boldsymbol{S}_t = A\boldsymbol{S}_{t-1} + C^{1/2}\boldsymbol{\varepsilon}_t$, where each element of $\boldsymbol{\varepsilon}_t$ is generated independently from the standardized version of the χ^2_2 distribution,

C is the same as the one in Case III, and $A = \text{diag}(0.3, 0.2, 0.1)$.

Obviously, in Case I, there is no serial correlation in $\{\boldsymbol{S}_t, t \geq 1\}$, the three continuous variables are independent of each other, the first variable has a standard normal distribution, and the second and third variables have skewed distributions. Case II is the same as Case I, except that the first continuous variable has a heavy-tail distribution. In Case III, there is still no serial correlation in the data, but the three continuous variables are correlated with each other and each of them has a skewed distribution. In Case IV, there is vector AR(1) serial correlation in the data, the three continuous variables are also correlated with each other, and each of them has a skewed distribution. Since the process observations $\{\boldsymbol{X}_t, t \geq 1\}$ are generated from $\{\boldsymbol{S}_t, t \geq 1\}$, the former should have similar correlation structures to those in the latter.

3.1 Evaluation of the IC performance

We first investigate the IC performance of the four control charts R-MCUSUM, LLCUSUM, LLD, and MOC in cases when the initial IC sample size m changes its value among $\{250, 500, 750, 1000, 2000\}$. In the four charts, the smoothing parameter of the R-MCUSUM chart is set to be 0.25, the allowance constant of the LLCUSUM chart is chosen to be 0.01, and the smoothing parameters of the MOC and LLD charts are set to be 0.05, as suggested in the related papers. For all charts, $b_{\max}$ is chosen to be 5 for data decorrelation.

The actual ARL_0 value of each chart is calculated in the following way in a given case. First, an initial IC dataset of size m is generated from the IC process distribution to estimate the related IC process parameters. Second, the circular block bootstrap procedure with block size of $l = 10$ and $B = 1,000$ bootstrap samples, together with the modified bisection searching algorithm discussed in Appendix A, is used for determining the control limit such that the nominal $\text{ARL}_0 = 200$ is reached. Third, the control chart with the searched control limit is used to monitor 100 sequences of process observations generated from the true IC process distribution directly, and the average of the 100 resulting run length values is used to estimate the actual ARL_0 value of the chart conditional on the

initial IC data. Finally, the entire simulation mentioned above, from generation of the initial IC data, determination of the control limit, to computation of the conditional ARL_0 value, is repeated for 100 times, and the actual ARL_0 value of the chart is estimated by the average of the 100 conditional ARL_0 values. In the above simulation, the length of individual sequence is truncated at 2,000. Namely, a total of 2,000 process observations are used in each simulation run of online process monitoring.

Figure 1 presents the actual ARL_0 values of the four control charts in various cases considered. From the figure, it can be seen that the R-MCUSUM chart has a reasonably good IC performance in all cases considered, since its actual ARL_0 values are all within 15% of the nominal ARL_0 value of 200 for all considered values of m. As a comparison, the three competing methods could have unreliable IC performance in some cases when m is relatively small. For instance, the charts LLCUSUM and MOC do not perform well in all cases when $m = 250$. The figure also shows that the IC performance of all charts become more reliable when m gets larger, which is intuitively reasonable.

3.2 Evaluation of the OC performance

In this part, we evaluate the OC performance of the four charts R-MCUSUM, LLCUSUM, LLD, and MOC using various simulation examples. To make the comparison among different control charts fair, their control limits have been adjusted properly so that their actual ARL_0 values are all the same to be 200. For detecting a given shift, their optimal ARL_1 values are considered, which are obtained by changing their parameter values such that their ARL_1 values are minimized. In the first example, the initial IC data size m is fixed at 500, and the three continuous variables in $\boldsymbol{S}_t$ have the same shift size δ that can change among $\{\pm 0.25, \pm 0.5, \pm 0.75, \pm 1.0\}$. The results are presented in Figure 2. From the figure, it can be seen that the R-MCUSUM chart is the best or close to the best in all cases considered. The other three charts do not perform well in some cases, especially in cases when the shift is negative, which is mainly due to the information loss during data categorization in these methods.

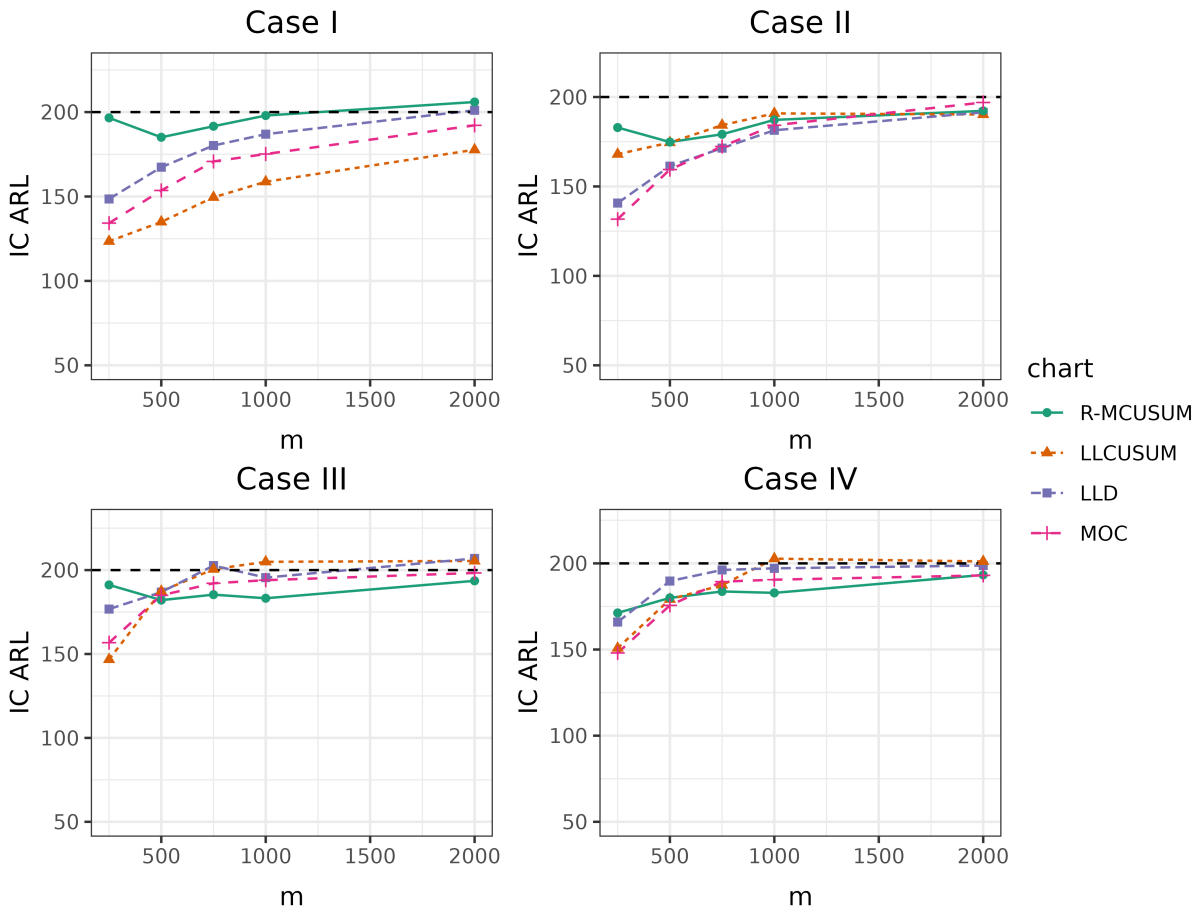

Figure 1: Estimated actual ARL_0 values of the four control charts when their nominal ARL_0 values are all fixed at 200 and the initial IC sample size m changes in $\{250, 500, 750, 1000, 2000\}$.

For the proposed R-MCUSUM chart, we also study its OC performance when its the allowance constant k is pre-specified to be one of $\{0.05, 0.1, 0.25, 0.5\}$ and all other setups are kept to be the same as those in Figure 2. The ARL_1 values are shown in Figure 3. From the figure, it can be seen that the IC performance of the proposed chart is quite robust with respect to the choice of the allowance constant, although k should not be chosen too large for detecting relatively small shifts. Based on the results in this example, it appears that a value of $k \in [0.1, 0.25]$ can provide a satisfactory OC performance for detecting a wide range of shifts.

In the previous two examples, the decorrelation window size $b_{\max}$ is fixed at 5. To study the impact of $b_{\max}$ on the OC performance of the proposed R-MCUSUM chart, we next consider an example when $b_{\max}$ can change among $\{5, 10, 20\}$ and all other setups are the

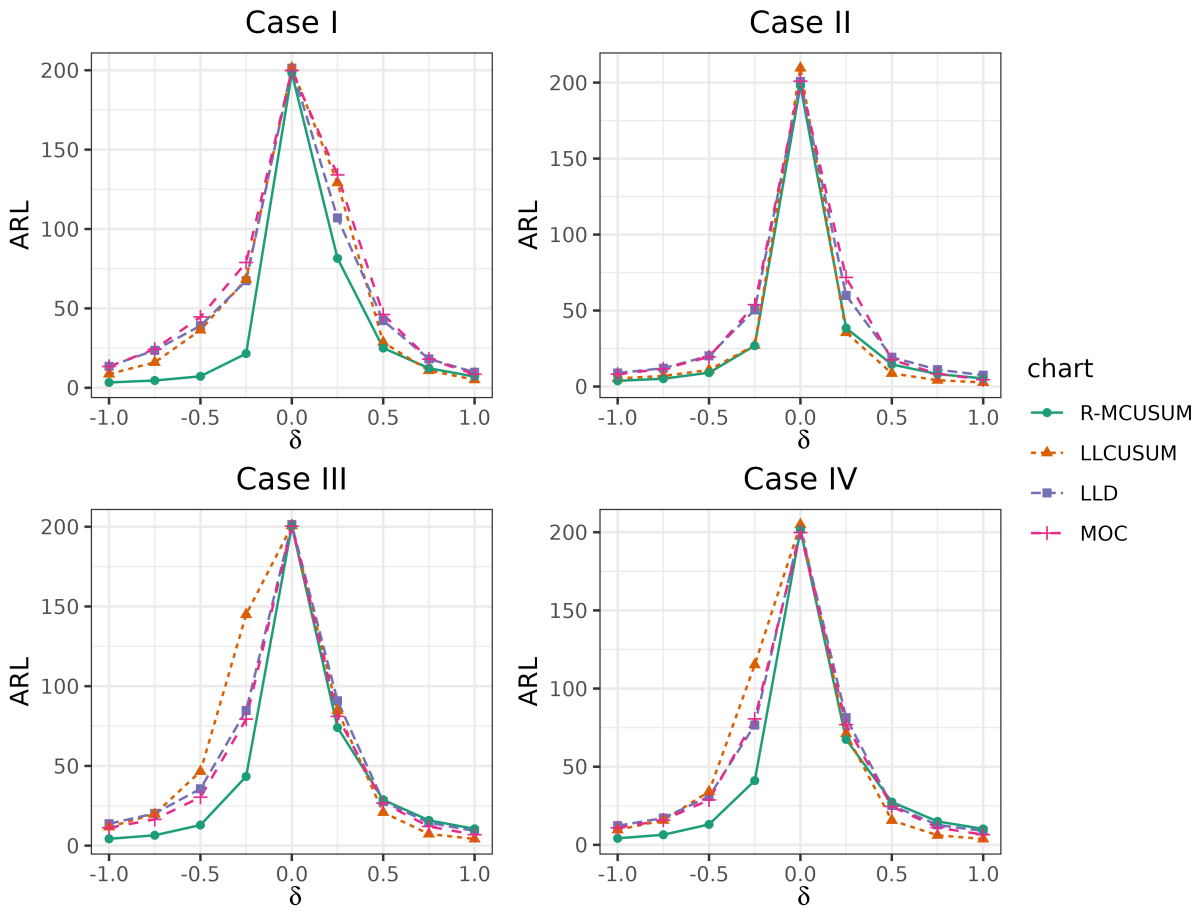

Figure 2: Optimal ARL_1 values of the four control charts in cases when their actual ARL_0 values are fixed at 200, the initial IC sample size m is fixed at 500, and all continuous variables in $\boldsymbol{S}_t$ are shifted by $\delta \in \{\pm 0.25, \pm 0.5, \pm 0.75, \pm 1.0\}$.

same as those in the example of Figure 2. As pointed out by Apley and Tsung (2002) and You and Qiu (2019), decorrelating the process observations can potentially mask the process shift, leading to decreased effectiveness of the related control chart in detecting the shift. The optimal ARL_1 values of the proposed R-MCUSUM chart in the cases considered are shown in Figure 4. From the results, it can be seen that a larger decorrelation window size appears to adversely impact the detection power of the control chart in this example, which is consistent with the results in Apley and Tsung (2002) and You and Qiu (2019). In addition, the impact of the decorrelation window size on the optimal ARL_1 values seems to be more obvious for smaller shifts, and this impact might also be aggravated by the well-known fact that small shifts are more difficult to detect by self-starting control charts (Wardell et al., 1994; Tsung and Apley, 2002; Zantek, 2005; Capizzi and Masarotto, 2010).

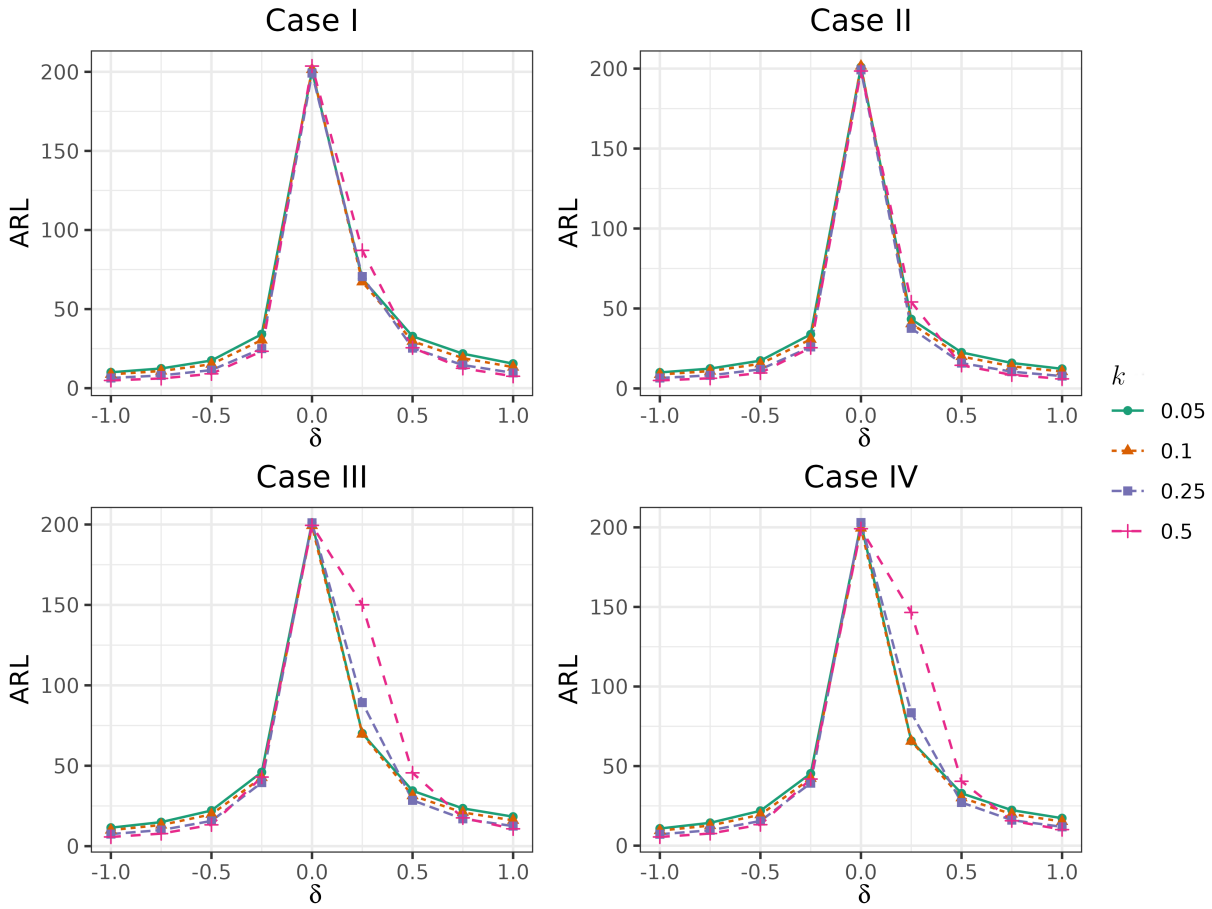

Figure 3: ARL_1 values of the proposed R-MCUSUM chart in cases when its allowance constant k changes among $\{0.05, 0.1, 0.25, 0.5\}$, its actual ARL_0 value is fixed at 200, the initial IC sample size m is set to to 500, and all continuous variables in $\boldsymbol{S}_t$ are shifted by $\delta \in \{\pm 0.25, \pm 0.5, \pm 0.75, \pm 1.0\}$.

All numerical examples discussed above focus on cases when there are only 3 heterogeneous quality variables, although the proposed R-MCUSUM chart can be applied to a process of arbitrary dimensions. It is expected that when the dimensionality of the process increases, the initial IC sample size should also increase to have a satisfactory performance. To study these issues, we consider an example in which the dimensionality of the process could be 3, 6, or 9. In the case where the dimension is 6, the vector $\boldsymbol{X}_t$ is generated by sampling a vector $\boldsymbol{S}_t$ of six latent variables. The observations $\boldsymbol{X}_t$ are then defined by selecting the first two latent variables, transforming the second two latent variables into ordinal categorical, and discretizing the last two latent variables into binary variables. Similarly, in the case where the dimension is 9, the same procedure is followed, but with nine latent

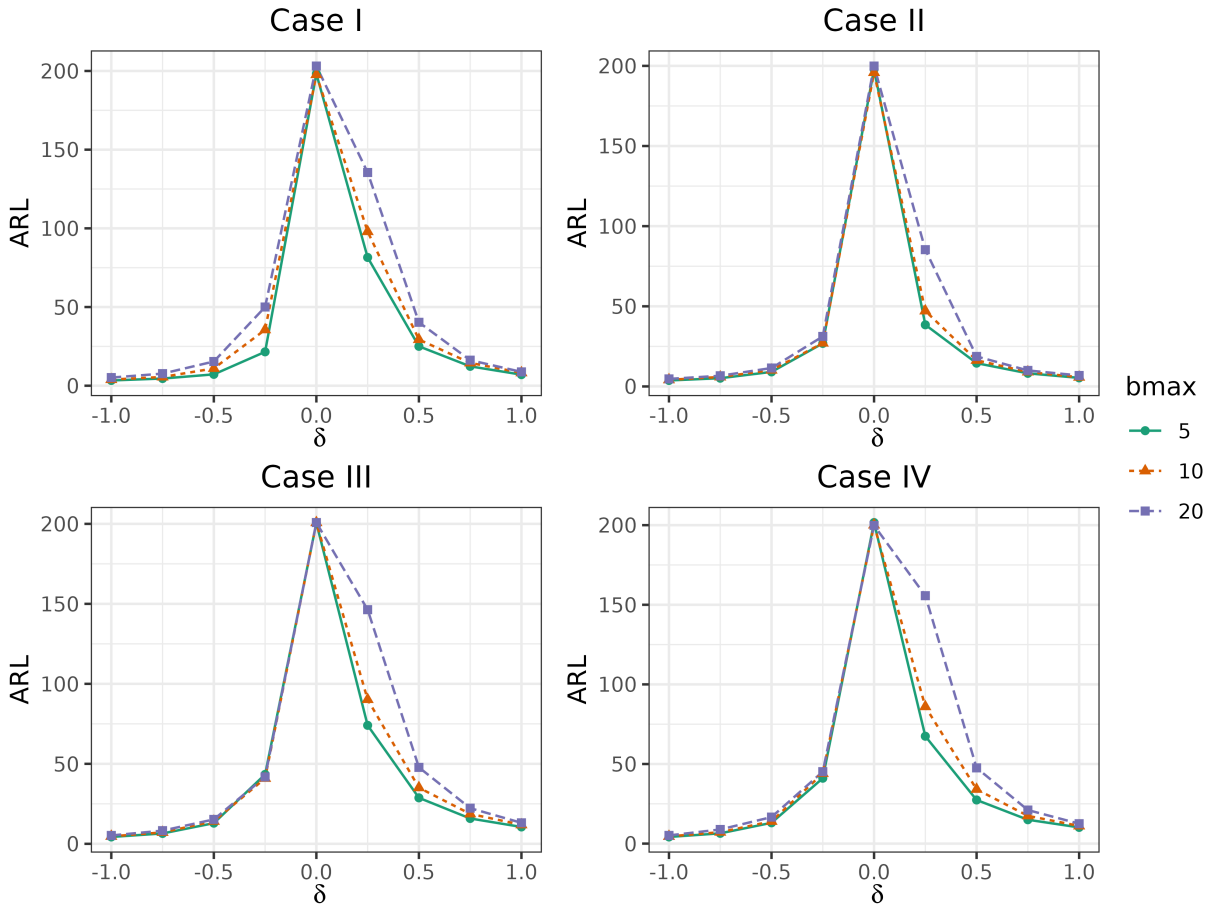

Figure 4: Optimal ARL_1 values of the proposed R-MCUSUM chart in cases when $b_{\max} \in \{5, 10, 20\}$, its actual ARL_0 value is fixed at 200, the initial IC sample size m is set to to 500, and all continuous variables in $\boldsymbol{S}_t$ are shifted by $\delta \in \{\pm 0.25, \pm 0.5, \pm 0.75, \pm 1.0\}$.

variables being sampled and then processed three at a time.

The correlation matrix C of the latent variables is defined, for dimensions $p = 6$ and 9, as

$$C = (c_{ij})_{i,j=1,\ldots,d^*} = \left(0.5^{|i-j|}\right)_{i,j=1,\ldots,d^*}.$$

When $p = 6$, the matrix A is given by $A = \text{diag}\,(0.3, 0.3, 0.2, 0.2, 0.1, 0.1)$, and when $p = 9$, $A = \text{diag}\,(0.3, 0.3, 0.3, 0.2, 0.2, 0.2, 0.1, 0.1, 0.1)$. All other settings are the same as those in the example of Figure 4, except that only Case IV is considered here for simplicity. All latent variables are therefore sampled from the standardized version of the χ^2_2 distribution.

The optimal ARL_1 values of the proposed R-MCUSUM chart are shown in Figure 5. From the plot, it can be seen that i) the OC performance of the proposed chart is quite ro-

bust to the dimensionality when detecting large shifts, and ii) its OC performance becomes worse when the dimensionality increases when detecting positive small shifts. As seen in Figure 2, negative shifts appear to be easier to detect than positive shifts of the same magnitudes, because of the fact that the distribution of each quality variable is skewed to the right in Case IV. Consequently, the dimensionality appears to have a smaller impact on the OC performance of R-MCUSUM when detecting negative shifts in this example.

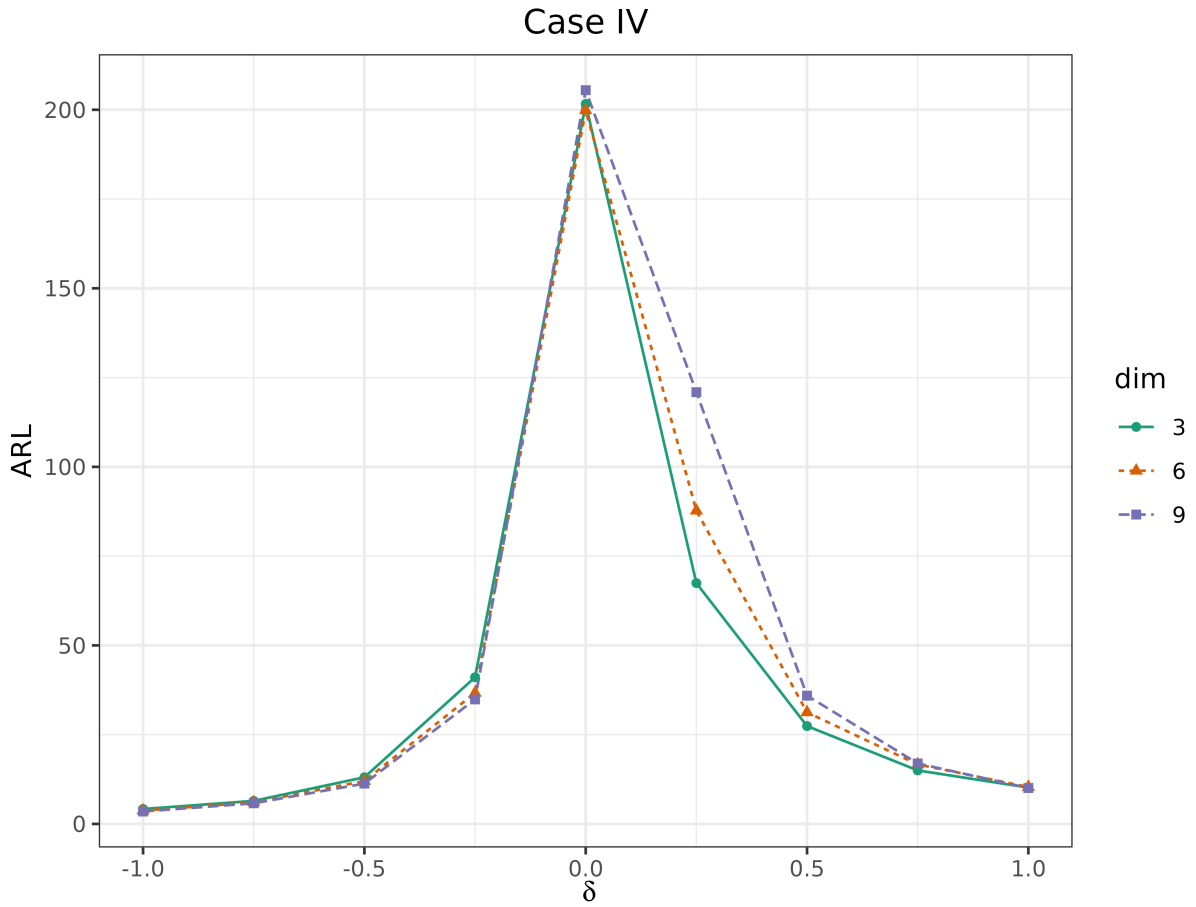

Figure 5: Optimal ARL_1 values of the proposed R-MCUSUM chart in cases when process observations are generated in Case IV when the dimension of the process varies in $\{3, 6, 9\}$, its actual ARL_0 value is fixed at 200, the initial IC sample size m is set to to 500, and all continuous variables in $\boldsymbol{S}_t$ are shifted by $\delta \in \{\pm 0.25, \pm 0.5, \pm 0.75, \pm 1.0\}$.

4 An Application

In this section, we demonstrate our proposed R-MCUSUM chart using an example of monitoring the behaviors of hotel customers. The dataset is publicly available as part of the `R` package `modeldata` (Kuhn, 2023) that is described in Antonio et al. (2019). It contains 28 variables for each customer, including the date of arrival, family composition, and information related to the hotel booking. For the purpose of demonstrating our proposed method, the following three quality variables are used, which are relevant for identifying patterns in the customer base:

1. Meal: Indicates the type of meal package requested alongside the room. It is an ordinal categorical variable with the following possible values: "none" (1), "breakfast" (2), "breakfast and one meal" (3), "breakfast, lunch and dinner" (4).

2. Special requests: A numerical variable indicating the number of special requests made by the customer.

3. Repeated guest: A binary variable indicating whether the customer of a booking is a repeated guest (1) or not (0).

To facilitate the analysis, we first identify a stable subset of the data as the initial IC data, and then use the remaining observations for online process monitoring. The original data are shown in Figure 6. From the figure, it is evident that the variable "Special requests" takes values from the finite set $\{0, 1, 2, 3, 4, 5\}$. Therefore, it can be viewed as either an ordinal categorical variable or a continuous variable. In fact, these two different treatments would result in the same outcomes, as explained below. If "Special requests" is treated as an ordinal categorical variable, it would be transformed according to the suggested method discussed in the second paragraph of Section 2. This transformation results in a numerical variable taking values from $\{1, 2, 3, 4, 5, 6\}$, which is equivalent to adding 1 to the "Special requests" values treated as a continuous variable. As a result, the standardized and decorrelated data of this variable remain identical in both setups. Consequently, all

process monitoring results would remain unchanged.

In this example, the first 1,900 observations of the three variables are used as the initial IC data, while the remaining observations are used for process monitoring. These two segments are separated by a vertical dashed line in each plot of Figure 6. For the initial IC data, the Box-Pierce tests show significant autocorrelation in the observed data of all three variables, with p-values of $\leq 2.2 \times 10^{-16}$, 2.9×10^{-11}, and $\leq 2.2 \times 10^{-16}$, respectively. The Augmented Dickey-Fuller Test (ADF) indicates that the serial correlation in each sequence is stationary, with p-values smaller than 0.01 (Note: stationarity is the alternative hypothesis of this test and thus confirmed when the test is significant). The Shapiro-Wilk normality tests show that none of the variables are normally distributed with all three p-values smaller than 2.2×10^{-16}.

After conducting data standardization, decorrelation, and the Rosenblatt transformation, as discussed in Section 2, the observed data are shown in Figure 7. For the initial IC data, the Box-Pierce tests confirm that the autocorrelation in the observed data has been mostly removed by the data transformations, with p-values for the three variables being 0.9941, 0.9947, and 0.9776, respectively. The normality of the data has improved, as indicated by the substantially increased test statistic values of the Shapiro-Wilk normality tests. However, the tests cannot confirm the normality of the transformed data, which is expected, as discussed in the second-to-last paragraph of Section 2.

To assess the performance of our proposed method, we apply the four control charts R-MCUSUM, LLCUSUM, LLD, and MOC to this dataset. In this example, the nominal ARL_0 values of all charts are fixed at 500, the data decorrelation window is set to be $b_{\max} = 20$, and the block size of the circular block bootstrap procedure is set to be 40. All other setups are the same as those in the simulation studies. The four control charts are shown in Figure 8. From the figure, it can be seen that the LLCUSUM control chart gives its first signal at the 8th observation. Based on our numerical experience, the run length distribution of the LLCUSUM chart tends to have a large variance, leading to a large probability to have small run length values. By checking the transformed data shown

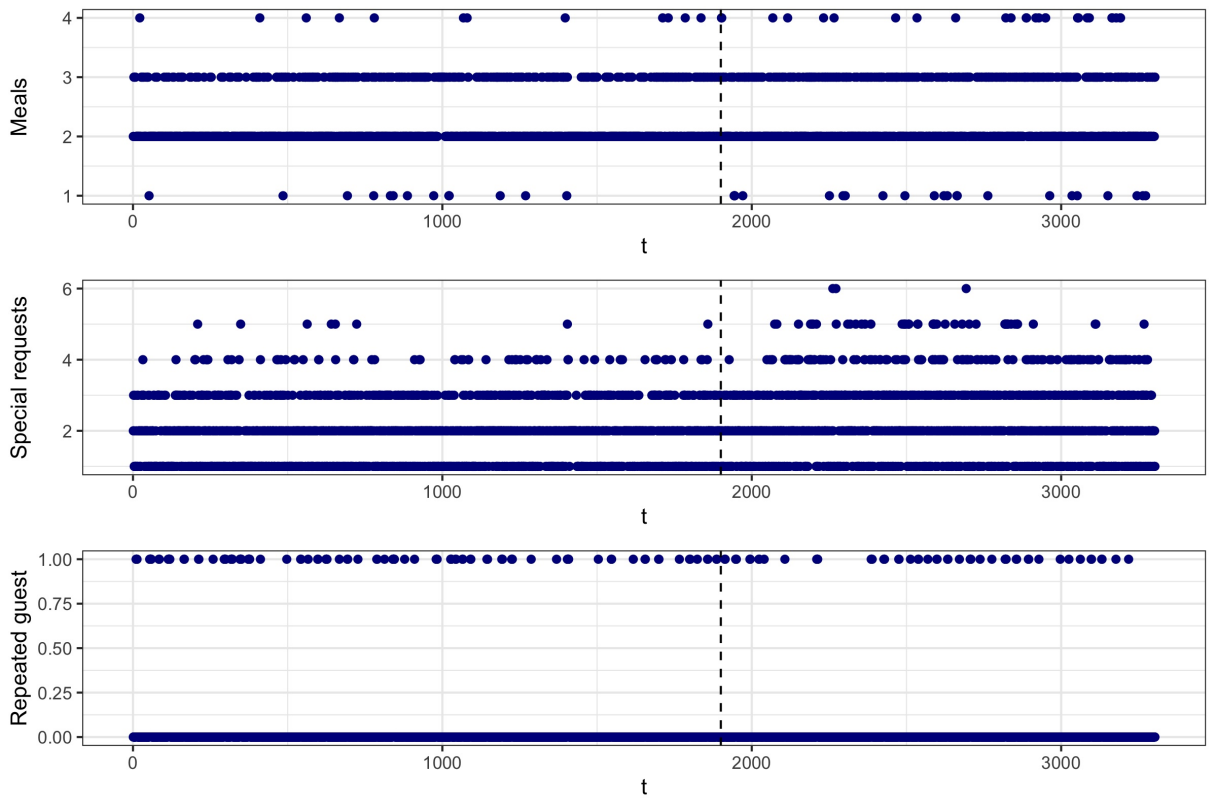

Figure 6: Original observations of the three quality variables in the hotel customer example. The dashed vertical line in each plot separates the initial IC data from the data for online process monitoring.

in Figure 7, it seems that there is no obvious shifts at or before the 8th observation time after the online process monitoring starts. Therefore, this signal by the LLCUSUM chart could be a spurious signal.

Among the remaining three charts, the proposed R-MCUSUM chart gives the earliest signal at the 302nd observation time during process monitoring, while the charts LLD and MOC give their signals at the 304th and 324th observation times, respectively. From Figure 7, it can be seen that there is a positive shift in the average number of special requests. This shift is successfully detected by the charts R-MCUSUM, LLD and MOC, and R-MCUSUM gives the earliest signal among them.

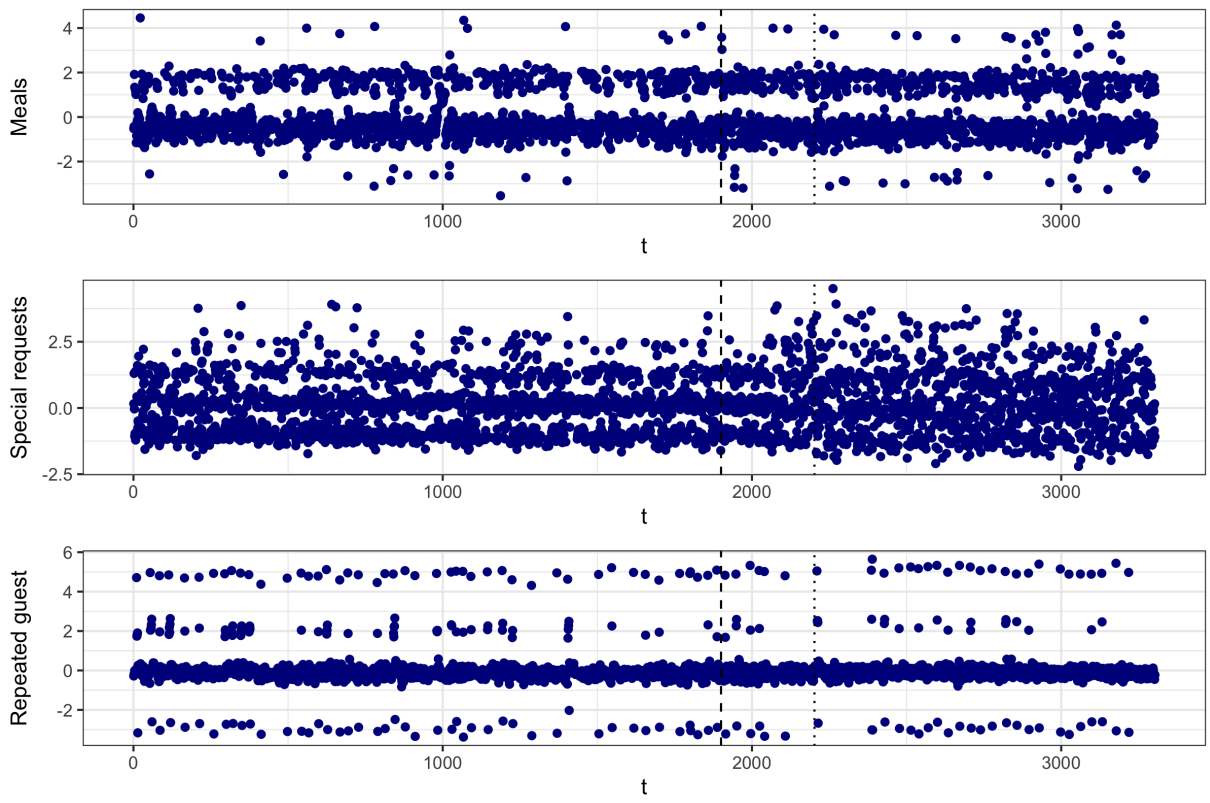

Figure 7: Transformed observations of the three quality variables in the hotel customer example. The dashed vertical line in each plot separates the initial IC data from the data for online process monitoring, and the dotted vertical line indicates the first signal time of the proposed R-MCUSUM chart.

5 Conclusions

Monitoring processes with mixed data of heterogeneous quality variables can be challenging due to the difficulty in properly modelling the observed data. In this paper, a general framework has been developed for this purpose. Instead of relying on data categorization that would lead to information loss, our method uses various data transformations to simplify the related problem into a process monitoring problem for processes with multivariate numerical quality variables. Numerical studies presented in Section 3 and Section 4 have shown that it is effective for monitoring mixed data in various cases considered. In practice, there could be many quality variables involved. Although our proposed method can handle such cases by theory, it may need a quite large initial IC dataset in order to have a reliable performance. In such cases, it may be beneficial to incorporate a dimensionality reduc-

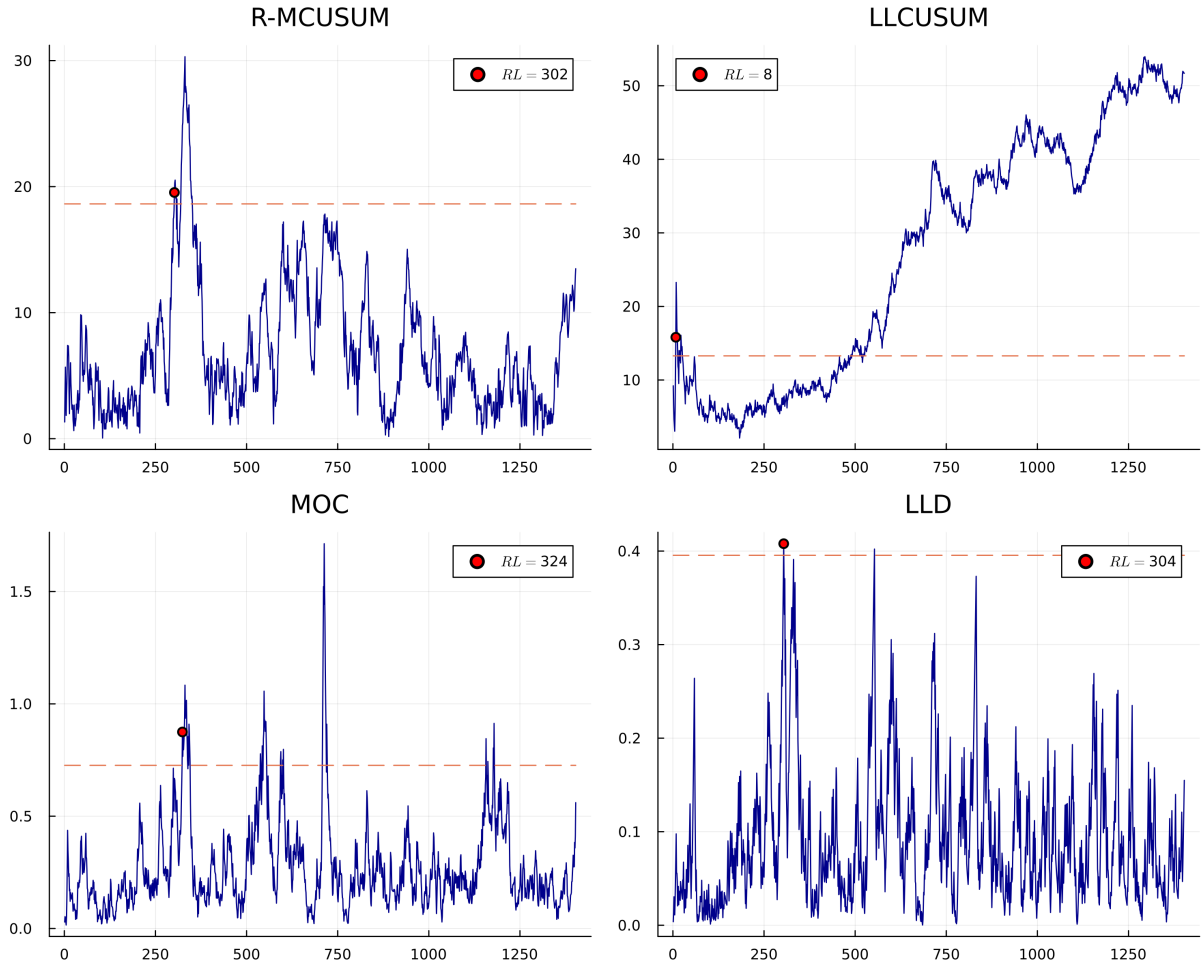

Figure 8: Control charts R-MCUSUM, LLCUSUM, MOC, and LLD when they are applied to the properly transformed hotel customer data. The horizontal dashed line in each plot indicates the control limit of the related chart when $\text{ARL}_0 = 500$. The red point in each plot indicates the first signal time by the related control chart.

tion technique, such as the variable selection (Hastie et al., 2013) and principal component analysis (Xie and Qiu, 2023) procedures. This requires much future research.

Acknowledgments

The authors thank the editor and two referees for many constructive comments and suggestions that improved the quality of the paper greatly.

Data availability statement

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

Competing interests statement

The authors have no relevant financial or non-financial interests to disclose

Funding

No funding was received for this research.

References

Agresti, A. (2013), *Categorical Data Analysis*, Hoboken, NJ: John Wiley & Sons Inc, 3 edition.

Antonio, N., de Almeida, A., and Nunes, L. (2019), "Hotel Booking Demand Datasets," *Data in Brief*, 22, 41–49.

Apley, D. W. and Lee, H. C. (2008), "Robustness Comparison of Exponentially Weighted Moving-Average Charts on Autocorrelated Data and on Residuals," *Journal of Quality Technology*, 40, 428–447.

Apley, D. W. and Tsung, F. (2002), "The Autoregressive T2 Chart for Monitoring Univariate Autocorrelated Processes," *Journal of Quality Technology*, 34, 80–96.

Bühlmann, P. (2002), "Bootstraps for Time Series," *Statistical Science*, 17, 52–72.

Capizzi, G. and Masarotto, G. (2008), "Practical Design of Generalized Likelihood Ratio Control Charts for Autocorrelated Data," *Technometrics*, 50, 357–370.

Capizzi, G. and Masarotto, G. (2009), "Bootstrap-Based Design of Residual Control Charts," *IIE Transactions*, 41, 275–286.

Capizzi, G. and Masarotto, G. (2010), "Self-Starting CUSCORE Control Charts for Individual Multivariate Observations," *Journal of Quality Technology*, 42, 136–151.

Capizzi, G. and Masarotto, G. (2016), "Efficient Control Chart Calibration by Simulated Stochastic Approximation," *IIE Transactions*, 48, 57–65.

Chakraborti, S. and Graham, M. (2019), "Nonparametric (Distribution-Free) Control Charts: An Updated Overview and Some Results," *Quality Engineering*, 31, 523–544.

Crosier, R. B. (1988), "Multivariate Generalizations of Cumulative Sum Quality-Control Schemes," *Technometrics*, 30, 291–303.

Ding, D., Tsung, F., and Li, J. (2016a), "Directional Control Schemes for Processes with Mixed-Type Data," *International Journal of Production Research*, 54, 1594–1609.

Ding, D., Tsung, F., and Li, J. (2016b), "Rank-Based Process Control for Mixed-Type Data," *IIE Transactions*, 48, 673–683.

Hastie, T., Tibshirani, R., and Friedman, J. (2013), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer Nature.

Higham, N. (1988), "Computing a Nearest Symmetric Positive Semidefinite Matrix," *Linear Algebra and its Applications*, 103, 103–118.

Huang, W. and Yeh, A. (2024), "Multiple Exponentially Weighted Moving Average Control Schemes for Monitoring and Diagnostics of Correlated Quality Variables of Different Types with Individual Observations," *Computers & Industrial Engineering*, 194, 110344.

Huang, W.-H., Sun, J., and Yeh, A. B. (2023), "Monitoring and Diagnostics of Correlated Quality Variables of Different Types," *Journal of Quality Technology*, 55, 1–33.

Johnson, S. G. (2023), *The NLopt Nonlinear-Optimization Package*, https://github.com/stevengj/nlopt.

Kuhn, M. (2023), *Modeldata: Data Sets Useful for Modeling Examples*, https://github.com/tidymodels/modeldata.

Lahiri, S. N. (2003), *Resampling Methods for Dependent Data*, New York: Springer.

Lee, H. C. and Apley, D. W. (2011), "Improved Design of Robust Exponentially Weighted Moving Average Control Charts for Autocorrelated Processes," *Quality and Reliability Engineering International*, 27, 337–352.

Li, J., Tsung, F., and Zou, C. (2012), "Directional Control Schemes for Multivariate Categorical Processes," *Journal of Quality Technology*, 44, 136–154.

Li, W., Pu, X., Tsung, F., and Xiang, D. (2017), "A Robust Self-Starting Spatial Rank Multivariate EWMA Chart Based on Forward Variable Selection," *Computers and Industrial Engineering*, 103, 116–130.

Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E. (1992), "A Multivariate Exponentially Weighted Moving Average Control Chart," *Technometrics*, 34, 46–53.

Montgomery, D. C. (2008), *Introduction to Statistical Quality Control*, Hoboken, N.J: Wiley, 6 edition.

Nataf, A. (1962), "Determination Des Distributions Dont Les Marges Sont Donnees," *Comptes rendus de l'Académie des Sciences*, 42–43.

Ning, X. and Tsung, F. (2010), "Monitoring a Process with Mixed-Type and High-Dimensional Data," in *2010 IEEE International Conference on Industrial Engineering and Engineering Management*.

Ning, X. and Tsung, F. (2012), "A Density-Based Statistical Process Control Scheme for High-Dimensional and Mixed-Type Observations," *IIE Transactions*, 44, 301–311.

Politis, D. N. and Romano, J. P. (1992), "A Circular Block-resampling Procedure for Stationary Data," in LePage, R. and Billard, L. (editors), *Exploring the Limits of Bootstrap*, New York: Wiley-Interscience, 263–270.

Prajapati, D. and Singh, S. (2012), "Control Charts for Monitoring the Autocorrelated Process Parameters: A Literature Review," *International Journal of Productivity and Quality Management*, 10, 207.

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

Qiu, P. (2013), *Introduction to Statistical Process Control*, Boca Raton, FL: CRC Press.

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

Qiu, P. and Xie, X. (2022), "Transparent Sequential Learning for Statistical Process Control of Serially Correlated Data," *Technometrics*, 64, 487–501.

Qiu, P. and You, L. (2022), "Dynamic Disease Screening by Joint Modelling of Survival and Longitudinal Data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71, 1158–1180.

Rosenblatt, M. (1952), "Remarks on a Multivariate Transformation," *The Annals of Mathematical Statistics*, 23, 470–472.

Shen, X., Tsui, K.-L., Zou, C., and Woodall, W. H. (2016), "Self-Starting Monitoring Scheme for Poisson Count Data With Varying Population Sizes," *Technometrics*, 58, 460–471.

Tsung, F. and Apley, D. W. (2002), "The Dynamic T2 Chart for Monitoring Feedback-Controlled Processes," *IIE Transactions*, 34, 1043–1053.

Tucker, G. R., Woodall, W. H., and Tsui, K.-L. (2002), "A Control Chart Method for Ordinal Data," *American Journal of Mathematical and Management Sciences*, 22, 31–48.

Wang, J., Li, J., and Su, Q. (2017), "Multivariate Ordinal Categorical Process Control Based on Log-Linear Modeling," *Journal of Quality Technology*, 49, 108–122.

Wang, J., Su, Q., Fang, Y., and Zhang, P. (2018), "A Multivariate Sign Chart for Monitoring Dependence Among Mixed-Type Data," *Computers & Industrial Engineering*, 126, 625–636.

Wang, Z., Li, X., Ma, Y., and Xue, L. (2024), "Monitoring of high-dimensional and high-frequency data streams: A nonparametric approach," *Quality Technology & Quantitative Management*, 22, 506–525.

Wardell, D. G., Moskowitz, H., and Plante, R. D. (1994), "Run-Length Distributions of Special-Cause Control Charts for Correlated Processes," *Technometrics*, 36, 3–17.

Xie, X. and Qiu, P. (2022), "Robust Monitoring of Multivariate Processes With Short-Ranged Serial Data Correlation," *Quality and Reliability Engineering International*, 38, 4196–4209.

Xie, X. and Qiu, P. (2023), "Control Charts for Dynamic Process Monitoring with an Application to Air Pollution Surveillance," *The Annals of Applied Statistics*, 17, 47–66.

Xie, X. and Qiu, P. (2024), "A General Framework for Robust Monitoring of Multivariate Correlated Processes," *Technometrics*, 66, 40–54.

Xue, L. and Qiu, P. (2021), "A Nonparametric CUSUM Chart for Monitoring Multivariate Serially Correlated Processes," *Journal of Quality Technology*, 53, 396–409.

You, L. and Qiu, P. (2019), "Fast Computing for Dynamic Screening Systems When Analyzing Correlated Data," *Journal of Statistical Computation and Simulation*, 89, 379–394.

Zantek, P. F. (2005), "Run-Length Distributions of Q-chart Schemes," *IIE Transactions*, 37, 1037–1045.

Zou, C. and Tsung, F. (2010), "Likelihood Ratio-Based Distribution-Free EWMA Control Charts," *Journal of Quality Technology*, 42, 174–196.

A Modified Bisection Search Algorithm

Calculating the control limit ρ to obtain an average run length $E_0[\text{RL}] = \text{ARL}_0$ can be time-consuming due to the heavy computational requirements of the decorrelation procedure (3). To address this concern, a workaround is implemented by simulating and storing a large number ($B = 1,000$) of Phase II decorrelated datasets of length $10 \cdot \text{ARL}_0$ from the IC process. This initial simulation employs the block bootstrap procedure discussed in Section 2. Next, a bisection search algorithm (see Qiu, 2013, for details) is employed. For each control chart, the Phase II samples are resampled using a bootstrap approach to calculate the ARL_0 value for the current estimate of the control limit. This iterative process continues until the bisection algorithm terminates. Furthermore, the initial interval required by the bisection search can be set to

$$[h_L, h_U] = \left[\min_{C_{i,b} \in \mathcal{C}} C_{i,t}, \max_{C_{i,b} \in \mathcal{C}} C_{i,t} \right], \tag{A.1}$$

where $\mathcal{C} = \{C_{i,n} : i = 1, \ldots, 10 \cdot \text{ARL}_0, b = 1, \ldots, B\}$ is the set of all values that the monitoring statistic takes in the B simulated Phase II datasets. Note that, by using the initial interval defined in (A.1), the property $1 = \text{ARL}_0(h_L) \leq \text{ARL}_0 \leq \text{ARL}_0(h_U) = 10 \cdot \text{ARL}_0$

required by the bisection algorithm is trivially verified. By storing the decorrelated data in advance and using this choice of initial interval in the bisection search, the computational cost of finding the control limit is greatly reduced.