Novel Nonparametric Control Charts for Monitoring Dispersion of Count Data

Zhiqiong Wang[1], Qingxia Wu[1], Peihua Qiu[2*]

[1]*School of Management, Tianjin University of Technology, Tianjin 300384, China*

[2]*Department of Biostatistics, University of Florida, Gainesville, Florida 32610, United States*

Abstract

Count data are common in practice. Statistical process control for count data thus has attracted much attention in recent years. Most existing methods on this topic focus on the detection of mean shifts of count data based on parametric modeling. However, their assumed parametric models (e.g., the Poisson probability model) are often invalid in practice due mainly to the potential impact of some latent confounding risk factors, which would lead to unreliable performance of the related control charts. In addition, it is highly desirable and important to monitor the dispersion of count data when the Poisson probability model is invalid. To this end, new nonparametric cumulative sum control charts and their corresponding self-starting versions are suggested in this paper for monitoring the dispersion of count data based on data categorization and categorical data analysis. Numerical results show that the proposed method can provide more effective and robust monitoring of count data in comparison with some representative existing methods. A real-data example is used to demonstrate its implementation and application.

Keywords: Count data; Data categorization; Dispersion; Self-starting control chart; Statistical process control.

1 Introduction

Count data are common in practice (Wang & Qiu, 2018). For instance, in medical studies, we are often concerned about the number of registered patients in a unit time interval at a

*Corresponding author. Email: pqiu@ufl.edu

specific hospital. In road traffics, we usually need to monitor the number of vehicles passing through a high-speed toll station in a given time interval. In the manufacturing industry, the number of defective products found in a batch of sampled products is a key quality index to monitor. Therefore, proper monitoring of count data is an important research problem with broad applications, which is the focus of this paper.

To monitor a sequential process online, statistical process control (SPC) charts provide a major analytic tool (Qiu, 2014). In the SPC literature, many control charts for monitoring count data have been proposed, which include the classical c and u charts that are Shewhart type control charts based on the Poisson distributional assumption. Since Shewhart charts are good in detecting large shifts only and ineffective in detecting small shifts, some researchers proposed various cumulative sum (CUSUM) charts (Lucas, 1985; White & Keats, 1996) and exponentially weighted moving average (EWMA) charts (Gan, 1990; Borror et al., 1998). However, these control charts are all based on the Poisson distributional assumption, and are improper for monitoring underdispersed or overdispersed count data that are common in practice. To accommodate under- or over-dispersion of the count data, some control charts based on alternative parametric distributions have been proposed. For instance, Sellers (2012) developed a flexible Shewhart chart by assuming the observed count data to follow the Conway-Maxwell-Poisson (COM-Poisson) distribution (Shmueli et al., 2005), which is more flexible than the Poisson distribution. Saghir & Lin (2014a,b), among others, further extended that control chart to EWMA and CUSUM charts, respectively. Other existing control charts for monitoring under- or over-dispersed count data used the Katz distribution family (Fang, 2003), the generalized Poisson model (He et al., 2006), the BerG distribution (Bourguignon et al., 2021), the Touchard distribution (Ho et al., 2021), and one parameter Poisson mixture models (Jesus et al., 2022), to list a few. Readers are referred to Saghir & Lin (2015) for an overview of control charts for monitoring under- or over-dispersed count data.

Most control charts discussed above are designed for detecting mean shifts in the observed count data. Although some of them consider the data dispersion, they assume that the dispersion parameter does not change over time, which may not be true in practice. As a matter of fact, the dispersion of the observed count data changes over time in practice, reflecting the time-varying impact of certain latent confounding risk factors on the count data, and it is highly desirable to monitor the dispersion of the observed count data since the performance of the process under monitoring depends heavily on the dispersion level (Zaman,

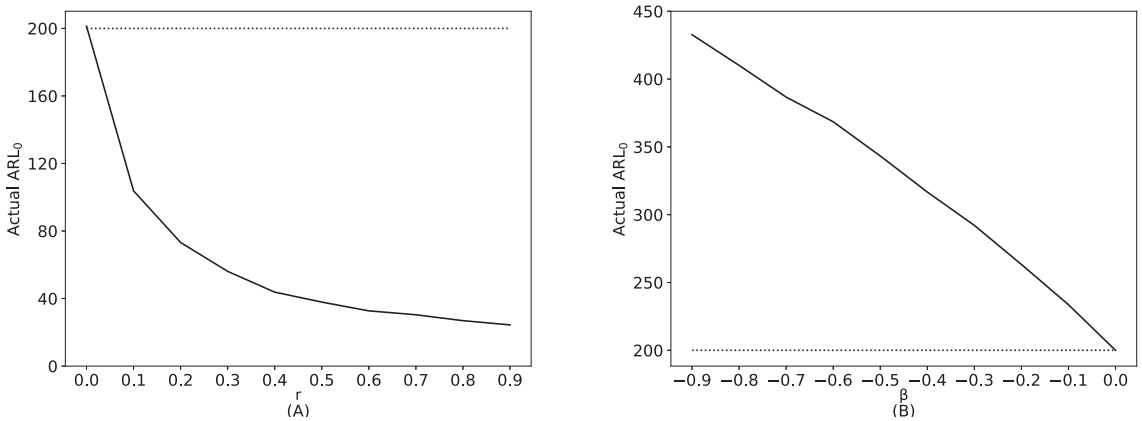

Figure 1: Actual ARL_0 values of the T-CUSUM chart in cases when the nominal ARL_0 value is 200, and the true IC distribution is a negative binomial distribution with the parameter r changing from 0 to 0.9 [plot (A)] or a generalized Poisson distribution with the parameter β changing from -0.9 to 0 [plot (B)]. See Section 3 for a detailed discussion of these two distributions.

2021). More specifically, an increase in dispersion often implies deterioration of the process, and a decrease indicates improvement in the process. In addition, it could be meaningless to detect a mean shift without ensuring that the dispersion is in-control (IC). It should be noted that several control charts mentioned above, such as those proposed by Lucas (1985), He et al. (2006), and Saghir & Lin (2014a), can address dispersion shifts. But, they rely on pre-specified parametric models, and are unreliable when their parametric models are invalid. As a demonstration, let us consider the traditional CUSUM chart, denoted by T-CUSUM, discussed in Lucas (1985), which is constructed and designed based on the assumption that the observed count data follow the regular Poisson distribution.

Figure 1 shows its actual IC average run length value, denoted as ARL_0, based on the assumption that the IC process distribution is Poisson(10), in cases when the true IC process distribution is a negative binomial distribution with the parameter r changing from 0 to 0.9 (i.e., a case of overdispersion) [plot (A)], or a generalized Poisson distribution with the parameter β changing from -0.9 to 0 (i.e., a case of underdispersion) [plot (B)]. The other parameter μ of the two true process distributions discussed is fixed at 10. The allowance constant of the T-CUSUM chart is chosen near the current IC mean level, and the nominal ARL_0 value is fixed at 200. From Figure 1, it can be seen that the actual ARL_0 values of the T-CUSUM chart are quite different from the nominal ARL_0 value in most cases considered. More specifically, the actual ARL_0 values are smaller than the nominal ARL_0 value when the true process distribution is the negative binomial, and larger than the nominal ARL_0

values when the true process distribution is the generalized Poisson distribution. In the former case, the related process would be stopped too often by the control chart when it is actually IC (Chatterjee & Qiu, 2009; Wang & Qiu, 2018). In the latter case, some true process distributional shifts would be detected much later than what we would expect, which is not good either because many defective products will be produced in such cases. This example shows that it is quite risky to develop a control chart for monitoring the dispersion of count data based on a parametric distribution assumption. In comparison, nonparametric or distribution-free control charts would be more desirable. See Qiu (2018) for an overview on nonparametric SPC.

Some nonparametric or distribution-free control charts have been proposed to monitor changes in data dispersion, including those based on the squared rank test (Das & Bhattacharya, 2008; Villanueva-Guerra et al., 2017), the sign statistic (Khilare & Shirke, 2012; Shirke & Barale, 2018; Godase et al., 2022), the Ansari-Bradley statistic (Zhou et al., 2016), and the statistic with a binomial distribution (Yang & Arnold, 2016; Haq, 2017). There are also nonparametric control charts that can jointly monitor process mean and dispersion, such as the ones based on the Lepage statistic (Ross et al., 2011; Mukherjee & Chakraborti, 2012; Chowdhury et al., 2015; Chong et al., 2018; Tercero-Gómez et al., 2020) and the Cucconi statistic (Chowdhury et al., 2014; Mukherjee & Marozzi, 2017; Liang et al., 2022), among others. A recent literature review on joint monitoring schemes can be found in Jalilibal et al. (2022).

Note that most nonparametric control charts mentioned above for monitoring data dispersion are based on the ordering or ranking information of process observations. To some extent, data ordering/ranking would result in a loss of information contained in the original observed data, and it is difficult to control the degree of information loss. In some applications, it is also difficult to use control charts constructed based on data ranking because of the discreteness of their charting statistics. As an alternative, Qiu & Li (2011) suggested a nonparametric control chart based on data categorization. Although data categorization would also result in information loss, the amount of lost information, however, can be controlled by the number of categories used. The larger the number of categories used, the less information would be lost. In addition, Li (2021) suggested a modification of the one by Qiu & Li (2011), in which an adaptive CUSUM chart was constructed by categorizing the data in a central-outward fashion for detecting arbitrary distributional changes, which was different from the approach in Qiu & Li (2011) to categorize the observed data from

the smallest to the largest values. Inspired by the data categorization approach in Qiu & Li (2011), Ye & Liu (2022) proposed a generic online nonparametric monitoring method for high-dimensional heterogeneous processes with partial observations, which only incorporated the small-to-large ordering information for detecting location shifts and thus impeded quick detection and limited broader applications for detecting dispersion and other shifts. To address this limitation, Ye et al. (2023) developed a method to use the center-outward ordering information (Li, 2021) for detecting dispersion shifts.

It should be pointed out that the nonparametric control charts discussed above for monitoring data dispersion are all designed for cases when the quality variables are continuous numerical variables. So far, we could not find any existing charts that are designed for monitoring the dispersion of count data. Wang & Qiu (2018) modified the method originally proposed by Qiu & Li (2011) for monitoring the mean of count data by categorizing the observed count data in a small-to-large fashion. A similar chart was discussed in Tang & Li (2023). In nonparametric statistics, it is well demonstrated that rank-based tests constructed based on the small-to-large ranking information, e.g., the Mann-Whitney test (Mann & Whitney, 1947), would be powerful for detecting location differences, and those based on the center-outward ranking information, e.g., the Ansari-Bradley test (Ansari & Bradley, 1960), would be powerful for detecting dispersion differences. Similarly, the data categorizion scheme used in Wang & Qiu (2018) and Tang & Li (2023) would not be effective for detecting dispersion shifts in count data, which is confirmed by our simulation study in Section 3.3. To make the control chart more effective for detecting dispersion shifts, it should be helpful to consider the center-outward categorization scheme.

Inspired by the methods in Wang & Qiu (2018) and Li (2021), we propose a novel nonparametric CUSUM chart for detecting dispersion shifts in the observed count data. The proposed method first categorizes the observed data from the center outward and then integrates Pearson's Chi-squared test statistic or the likelihood ratio test statistic into a CUSUM charting scheme. Compared to the methods discussed in Wang & Qiu (2018) and Tang & Li (2023), we categorize the count data in a center-outward fashion in this paper, which should be more powerful for detecting dispersion shifts. In addition, because the new method is for monitoring count data and the methods in Tang & Li (2023) and Ye et al. (2023) are for monitoring continuous data, the categorization procedure used in this paper is also different from the ones used in the latter two papers (see Section 2.1 for more details). To implement the new method, a sufficiently large IC dataset is required to accurately estimate the

quantiles of the IC process distribution used to categorize the observed data, which may not be available in practice. To overcome this limitation, a self-starting version of the proposed CUSUM chart is also developed, in which quantile estimates are updated sequentially at each observation time after the process under monitoring is declared to be IC by the control chart.

The remainder of the paper is organized as follows. Our proposed nonparametric control chart is described in Section 2. Simulation studies to evaluate its numerical performance are presented in Section 3. A real-data application is discussed in Section 4 to demonstrate the proposed method in a real-world setting. Finally, some concluding remarks are given in Section 5.

2 Methodology

Our proposed method for detecting dispersion shifts of count data is described in Section 2.1. Its self-starting version is discussed in Section 2.2. Determination of the control limits of the two control charts is discussed in Section 2.3.

2.1 Proposed CUSUM charts for detecting dispersion shifts

Assume that there is an IC dataset available before online process monitoring, which consists of M independent and identically distributed (i.i.d.) IC process observations $X_{-M+1}, \ldots, X_0$ that take count values. The process observations under online monitoring are denoted as $\{X_n, n \geq 1\}$. To monitor the process online, we first partition the range of observation values, $[0, \infty)$, into the following $d (d > 1)$ intervals:

$$A_1 = (q_{d-1}, q_{d+1}],\ A_2 = (q_{d-2}, q_{d-1}] \cup (q_{d+1}, q_{d+2}],\ \ldots,\ A_d = [0, q_1] \cup (q_{2d-1}, \infty),$$

where $0 < q_1 < q_2 < \ldots < q_{2d-1} < \infty$ are $2d-1$ boundary points of the partitioning intervals. It is clear that $A_1, \ldots, A_d$ are ordered from the center outward. Define

$$Y_{n,j} = \mathrm{I}(X_n \in A_j), \quad \text{for } j = 1, 2, \ldots, d,$$

where $\mathrm{I}(a) = 1$ if a is "true" and 0 otherwise. We can see that $Y_{n,j}$ indicates whether X_n belongs to the jth interval A_j. So $\mathbf{Y}_n = (Y_{n,1}, Y_{n,2}, \ldots, Y_{n,j})'$ has one and only one component being 1, and the index of the component being 1 has a discrete distribution with probabilities $f_j = P(X_n \in A_j)$, for $j = 1, 2, \ldots, d$. For the convenience of presentation, the distribution $\mathbf{f} = (f_1, f_2, \ldots, f_d)'$ is called the distribution of $\mathbf{Y}_n$. Let $\mathbf{f}^{(0)} = (f_1^{(0)}, f_2^{(0)}, \ldots, f_d^{(0)})'$ be the

IC distribution of $\mathbf{Y}_n$ and $\mathbf{f}^{(1)} = (f_1^{(1)}, f_2^{(1)}, \ldots, f_d^{(1)})'$ be its out-of-control (OC) distribution. Under some regularity conditions, it can be checked that $\mathbf{f}^{(1)}$ is different from $\mathbf{f}^{(0)}$ if there is a dispersion shift in X_n. In Li (2021), it was suggested that the parameter q_j should be chosen to be the $(j/(2d))$-th quantile of the IC process distribution, for each j, and $\{q_1, q_2, \ldots, q_{2d-1}\}$ were chosen such that $\mathbf{f}^{(0)} \approx (1/d, 1/d, \ldots, 1/d)'$. For count data, this is often difficult to achieve because of the discreteness of the data. In such cases, we suggest choosing the parameters such that $\mathbf{f}^{(0)}$ is as close to a uniform distribution as possible.

The charting statistic of our proposed CUSUM chart is then defined as follow. Let $\mathbf{S}_0^{\text{obs}} = \mathbf{S}_0^{\text{exp}} = \mathbf{0}$ be two d-dimensional vectors, and

$$\begin{cases} \mathbf{S}_n^{\text{obs}} = \mathbf{0}, & \text{if } C_n \le k_{\text{P}}, \\ \mathbf{S}_n^{\text{exp}} = \mathbf{0}, & \text{if } C_n \le k_{\text{P}}, \\ \mathbf{S}_n^{\text{obs}} = (\mathbf{S}_{n-1}^{\text{obs}} + \mathbf{Y}_n)(C_n - k_{\text{P}})/C_n, & \text{if } C_n > k_{\text{P}}, \\ \mathbf{S}_n^{\text{exp}} = (\mathbf{S}_{n-1}^{\text{exp}} + \mathbf{f}^{(0)})(C_n - k_{\text{P}})/C_n, & \text{if } C_n > k_{\text{P}}, \end{cases}$$

where

$$C_n = \left[(\mathbf{S}_{n-1}^{\text{obs}} - \mathbf{S}_{n-1}^{\text{exp}}) + (\mathbf{Y}_n - \mathbf{f}^{(0)})\right]' \left[\text{diag}(\mathbf{S}_{n-1}^{\text{exp}} + \mathbf{f}^{(0)})\right]^{-1} \left[(\mathbf{S}_{n-1}^{\text{obs}} - \mathbf{S}_{n-1}^{\text{exp}}) + (\mathbf{Y}_n - \mathbf{f}^{(0)})\right],$$

$k_{\text{P}} \ge 0$ is the allowance constant, $\text{diag}(\mathbf{a})$ denotes a diagonal matrix with its diagonal elements being the corresponding elements of the vector $\mathbf{a}$, and the superscripts "obs" and "exp" denote observed and expected counts, respectively. Then, the chartng statistic is defined to be

$$u_{n,\text{P}} = (\mathbf{S}_n^{\text{obs}} - \mathbf{S}_n^{\text{exp}})' \left[\text{diag}(\mathbf{S}_n^{\text{exp}})\right]^{-1} (\mathbf{S}_n^{\text{obs}} - \mathbf{S}_n^{\text{exp}}),$$

and the chart gives a signal of dispersion shift if

$$u_{n,\text{P}} > h_{\text{P}}, \tag{1}$$

where $h_{\text{P}} > 0$ is a control limit. The proposed chart based on Equation (1) is denoted as P-CUSUM chart hereafter since it is constructed based on the Pearson's Chi-squared test. When $k_{\text{P}} = 0$, it is not difficult to check that $\mathbf{S}_n^{\text{obs}}$ is a frequency vector with its jth element being the cumulative observed count of observations in the jth interval A_j as of the time point n, for $j = 1, 2, \ldots, d$, and $\mathbf{S}_n^{\text{exp}}$ is the vector of the corresponding cumulative expected counts. Therefore, $u_{n,\text{P}}$ is the Pearson's Chi-squared test statistic in such cases to measure the difference between the cumulative observed and expected counts by time n.

Because $u_{n,\text{P}}$ can only take some discrete values on the positive number line, it could be hard to find a proper h_{P} value so that a pre-specified nominal ARL_0 value is reached within

a desired precision. To overcome this limitation, we can use the modification procedure as in Wang & Qiu (2018) by adding a small random number generated from $\mathrm{N}(0, s^2)$ to each component of $\mathbf{Y}_n$ to alleviate the discreteness of $u_{n,\mathrm{P}}$. As long as s is chosen small, the OC behavior of the chart would not change much and most nominal ARL_0 values can be reached within the desired precision after the modification. In all the simulation studies in this paper, we choose $s = 0.01$.

Besides the CUSUM chart (1), we can also construct a CUSUM chart based on the likelihood ratio test. More specifically, let $\tilde{\mathbf{S}}_n^{\mathrm{obs}}$ and $\tilde{\mathbf{S}}_n^{\mathrm{exp}}$ be quantities defined in the same way as $\mathbf{S}_n^{\mathrm{obs}}$ and $\mathbf{S}_n^{\mathrm{exp}}$ used in chart (1), except that k_{P} is replaced by another constant k_{L} and C_n is replaced by

$$\widetilde{C}_n = 2(\widetilde{\mathbf{S}}_{n-1}^{\mathrm{obs}} + \mathbf{Y}_n)' \log\left(\frac{\widetilde{\mathbf{S}}_{n-1}^{\mathrm{obs}} + \mathbf{Y}_n}{\widetilde{\mathbf{S}}_{n-1}^{\mathrm{exp}} + \mathbf{f}^{(0)}}\right),$$

where $\boldsymbol{a}/\boldsymbol{b}$ denotes a vector obtained by component-wise division of the vector $\boldsymbol{a}$ by the vector $\boldsymbol{b}$ and $\log(\boldsymbol{a}/\boldsymbol{b})$ denotes a component-wise operation as well. Then, the new charting statistic is defined as

$$u_{n,\mathrm{L}} = 2(\widetilde{\mathbf{S}}_n^{\mathrm{obs}})' \log\left(\frac{\widetilde{\mathbf{S}}_n^{\mathrm{obs}}}{\widetilde{\mathbf{S}}_n^{\mathrm{exp}}}\right).$$

It gives a signal when

$$u_{n,\mathrm{L}} > h_{\mathrm{L}}, \tag{2}$$

where $h_{\mathrm{L}} > 0$ is the control limit. The chart based on Equation (2) is denoted as L-CUSUM.

2.2 Self-starting version of the proposed charts

To implement the proposed control charts, we need to know the quantiles $\{q_1, q_2, \ldots, q_{2d-1}\}$ for categorizing the original data. In subsection 2.1, we assume that the size of the IC dataset, M, is large enough such that the quantiles could be accurately estimated during online process monitoring. When such IC dataset is unavailable, we suggest the self-starting version of the proposed charts, in which the estimates of the quantiles are constantly updated. More specifically, at time n, we have the IC observations $X_{-M+1}, \ldots, X_0, X_1, \ldots, X_{n-1}$ collected in the past. Let $X_{n,(1)} < X_{n,(2)} < \ldots X_{n,(M+n-1)}$ denote their order statistics. For a given j, find the integer $l_{n,j}$ such that

$$\left| \frac{l_{n,j}}{M+n} - \frac{j}{2d} \right| = \min_{1 \le l \le M+n-1} \left| \frac{l}{M+n} - \frac{j}{2d} \right|.$$

Then, the updated estimate of q_j at time n is defined to be

$$\widehat{q}_{n,j} = X_{n,(l_{n,j})}.$$

Based on these estimates, we can partition $[0,\infty)$ into the following $d(d>1)$ center-outward intervals:

$$\widehat{A}_{n,1} = (\widehat{q}_{n,d-1}, \widehat{q}_{n,d+1}],\ \widehat{A}_{n,2} = (\widehat{q}_{n,d-2}, \widehat{q}_{n,d-1}]\cup(\widehat{q}_{n,d+1}, \widehat{q}_{n,d+2}],\ \ldots,\ \widehat{A}_{n,d} = [0, \widehat{q}_{n,1}]\cup(\widehat{q}_{n,2d-1}, \infty).$$

Define $\widehat{Y}_{n,j} = \mathrm{I}(X_n \in \widehat{A}_{n,j})$ and $\widehat{\mathbf{Y}}_n = (\widehat{Y}_{n,1}, \widehat{Y}_{n,2}, \ldots, \widehat{Y}_{n,j})'$, for $j = 1, 2, \ldots, d$. In the self-starting control chart, we replace $\mathbf{Y}_n$ in our proposed CUSUM statistic $u_{n,\mathrm{P}}$ described in Section 2.1 by $\widehat{\mathbf{Y}}_n$. Let $\widehat{\mathbf{S}}_n^{\mathrm{obs}}$ and $\widehat{\mathbf{S}}_n^{\mathrm{exp}}$ be quantities defined in the same way as $\mathbf{S}_n^{\mathrm{obs}}$ and $\mathbf{S}_n^{\mathrm{exp}}$ used in chart (1), except that k_{P} is replaced by another constant k_{PS} and C_n is replaced by

$$\widehat{C}_n = \left[(\widehat{\mathbf{S}}_{n-1}^{\mathrm{obs}} - \widehat{\mathbf{S}}_n^{\mathrm{exp}}) + (\widehat{\mathbf{Y}}_n - \mathbf{f}^{(0)})\right]' \left[\mathrm{diag}(\widehat{\mathbf{S}}_n^{\mathrm{exp}} + \mathbf{f}^{(0)})\right]^{-1} \left[(\widehat{\mathbf{S}}_{n-1}^{\mathrm{obs}} - \widehat{\mathbf{S}}_{n-1}^{\mathrm{exp}}) + (\widehat{\mathbf{Y}}_n - \mathbf{f}^{(0)})\right].$$

The charting statistic of the self-starting CUSUM chart is then defined to be

$$u_{n,\mathrm{PS}} = (\widehat{\mathbf{S}}_n^{\mathrm{obs}} - \widehat{\mathbf{S}}_n^{\mathrm{exp}})' \left[\mathrm{diag}(\widehat{\mathbf{S}}_n^{\mathrm{exp}})\right]^{-1} (\widehat{\mathbf{S}}_n^{\mathrm{obs}} - \widehat{\mathbf{S}}_n^{\mathrm{exp}}).$$

It gives a signal when

$$u_{n,\mathrm{PS}} > h_{\mathrm{PS}}, \tag{3}$$

where $h_{\mathrm{PS}} > 0$ is the control limit. The chart based on Equation (3) is denoted by PS-CUSUM. The self-starting version of the L-CUSUM chart can be constructed similarly, which is denoted as LS-CUSUM.

2.3 Determination of the control limit

In the above CUSUM chart, the allowance constant is usually specified beforehand. Then, the control limit is chosen such that a pre-specified ARL_0 value is reached. Taking the P-CUSUM chart as an example, to compute h_{P} from the IC dataset, the following iterative bisection searching algorithm based on the bootstrap resampling idea can be used.The control limit h_{L} or h_{PS} can be chosen similarly to h_{P}.

Step 1: In the ith iteration, h_{P} is searched in the interval $[L^{(i)}, U^{(i)}]$. When $i = 1$, $L^{(1)}$ and $U^{(1)}$ are the pre-specified lower and upper bounds, respectively.

Step 2: A sequence of observations is selected randomly with replacement from the IC dataset and this sequence is regarded as process observations for online monitoring. Then, the P-CUSUM chart with $h_{\mathrm{P}} = h^{(i)} = (L^{(i)} + U^{(i)})/2$ is applied to this sequence and the run length (RL), denoted as $\mathrm{RL}^{(i)}$, is recorded, where RL is defined to be the number of observation times from the beginning of process monitoring to the first signal time.

Step 3: Step 2 is repeated for N times, and the actual ARL_0 value is approximated by the sample mean ${\mathrm{ARL}_0}^{(i)}$ of the N values of $\mathrm{RL}^{(i)}$.

Step 4: If $|{\mathrm{ARL}_0}^{(i)} - \mathrm{ARL}_0| < \varepsilon$, where ε is a small number and denotes the required searching accuracy, then the whole algorithm stops and the searched value of h_{P} is $h^{(i)}$. Otherwise, set $L^{(i+1)} = h^{(i)}$, $U^{(i+1)} = U^{(i)}$, and $h^{(i+1)} = (L^{(i+1)} + U^{(i+1)})/2$ if ${\mathrm{ARL}_0}^{(i)} < \mathrm{ARL}_0$, and set $L^{(i+1)} = L^{(i)}$, $U^{(i+1)} = h^{(i)}$, and $h^{(i+1)} = (L^{(i+1)} + U^{(i+1)})/2$ if ${\mathrm{ARL}_0}^{(i)} > \mathrm{ARL}_0$. In this latter case, the algorithm executes the next iteration until the maximum number Q of iterations is reached.

The above searching algorithm usually converges quickly. Although it is rare, if it does not stop before the Qth iteration, then define $h_{\mathrm{P}} = h^{(Q)}$.

3 Simulation studies

In this section, we investigate the OC performance of the proposed control charts and compare them with some existing parametric and nonparametric control charts under different discrete distributions. For measuring the OC performance of a chart, the ARL can be computed under either the zero-state or the steady-state setting. In this paper, we choose to use the steady-state OC ARL, denoted as ARL_1, as the metric to compare the OC performance of different control charts. For describing the distribution of count data, the regular Poisson distribution Poisson(λ) is routinely used, where $\lambda > 0$ is a parameter. This distribution has an important property that its mean and variance are both equal to λ. In the simulation study, the following two discrete distributions are considered, which are both different from the regular Poisson distribution.

I. Negative binomial distribution: cases with overdispersion

The negative binomial distribution has two parameters μ and r and is denoted as $\mathrm{NB}(\mu, r)$. After an uncommon but useful parameterization (Bliss & Owen, 1958; Collings & Margolin,

1985), the random variable $X \sim \text{NB}(\mu, r)$ has mean μ and the probability mass function (PMF) being

$$\Pr\{X = x\} = \frac{\Gamma(x + r^{-1})}{x!\Gamma(r^{-1})}(\frac{r\mu}{1 + r\mu})^x(\frac{1}{1 + r\mu})^{r^{-1}},$$

where $x = 0, 1, 2, \ldots$, $0<\mu<\infty$. The variance of X is $\mu(1 + r\mu)$. So, the index of dispersion, defined as the ratio of variance to mean, is $(1+r\mu)$, which implies that the negative binomial distribution is overdispersed.

II. Generalized Poisson distribution: cases with mixed dispersion

The generalized Poisson distribution (Consul & Jain, 1973) has two parameters μ and β and is denoted as $\text{GP}(\mu, \beta)$. Its PMF is

$$\Pr\{X = x\} = \begin{cases} \mu(1-\beta)\left[\mu(1-\beta) + \beta x\right]^{x-1} \frac{exp[-\mu(1-\beta)-\beta x]}{x!} & \text{, for } x = 0, 1, 2, \ldots, \\ 0 & \text{, for } x \geq m, \text{if } \mu(1-\beta) + m\beta \leq 0. \end{cases}$$

where $\mu>0$, $-1 \leq \beta \leq 1$. The mean and the variance of this distribution are μ and $\frac{\mu}{(1-\beta)^2}$, respectively. Thus, the index of dispersion is $1/(1-\beta)^2$. Compared with the Poisson distribution, the additional parameter β controls the degree of dispersion. With $\beta = 0$, $\text{GP}(\mu, \beta)$ is just the Poisson distribution Poisson(μ). When $0<\beta<1$, the index of dispersion is larger than 1, implying that the generalized Poisson distribution is overdispersed. When $-1<\beta<0$, the index of dispersion is smaller than 1, implying that the generalized Poisson distribution is underdispersed.

Next, we first evaluate the performance of the proposed P-CUSUM and L-CUSUM charts in comparison with their self-starting versions for monitoring count data dispersion in Subsection 3.1 when a sufficiently large IC dataset is unavailable. The impact of the size of the IC dataset, the number of categories used, and the allowance constant on the performance of the P-CUSUM chart is studied in Subsection 3.2. In Subsection 3.3, we evaluate the performance of the proposed charts P-CUSUM and L-CUSUM in comparison with the corresponding methods discussed in Wang & Qiu (2018) where the observed data are categorized from the smallest to the largest. Finally, in Subsection 3.4, the proposed control charts are compared with the traditional parametric control chart proposed by (Lucas, 1985) and the nonparametric control charts proposed by (Shirke & Barale, 2018).

3.1 Comparison with self-starting version

The performance of the proposed P-CUSUM and L-CUSUM charts depends on the IC data size M. When M is small, their self-starting version should be used, as discussed in Subsection

2.2. In this part, we briefly compare the P-CUSUM and L-CUSUM charts with their self-starting version PS-CUSUM and LS-CUSUM in cases when $M = 50$. In the four charts, their allowance constants are all chosen to be 0.01, their assumed ARL_0 values are set to be 200, and the number of categories d is fixed at 5 in all charts. Their control limits are searched by the bootstrap procedure with $Q = 100$ maximum iterations. In each iteration, the ARL_0 value is computed from 10,000 replicated simulation runs. The IC distribution is specified to be one of the following four distributions, each of which has a mean of 10: Poisson(10), NB(10, 0.4), GP(10, 0.4), and GP(10, -0.4). These four cases represent different scenarios when the true IC distribution is equidispersed, overdispersed, or underdispersed. For the negative binomial distribution and the generalized Poisson distributions, the parameters r and β are used to simulate the dispersion changes of the count data. The calculated ARL_1 values and their standard errors of the related charts are shown in Table 1. From the table, it can be seen that the PS-CUSUM (LS-CUSUM) chart is outperformed by the P-CUSUM (L-CUSUM) chart for detecting small shifts when the true IC distribution is the Poisson distribution, but the PS-CUSUM chart becomes better when the shift gets larger. When the true IC distribution is the negative binomial distribution or the generalized Poisson distribution, the PS-CUSUM (LS-CUSUM) chart performs much better than the P-CUSUM (L-CUSUM) chart for detecting positive dispersion shifts (dispersion increase) and has worse performance than the P-CUSUM (L-CUSUM) chart for detecting negative dispersion shifts (dispersion decrease). This result can be explained by the fact that the self-starting charts could miss some dispersion shifts, especially when the shifts are negative. In practice, we are usually more concerned about dispersion increase since it often indicates process deterioration. The proposed self-starting control charts are more effective and robust to use in such cases.

3.2 Impact of parameter settings on the P-CUSUM chart

In this subsection, we further study the impact of parameter settings in the P-CUSUM chart on its numerical performance. First, we study the impact of the IC dataset size M on the performance of the P-CUSUM chart, by changing M among 200, 500, 1,000, and 2,000, and keeping the other setups of the chart to be the same as those in the example of Table 1. Its calculated ARL_1 values are displayed in Figure 2. From the figure, it can be seen that the ARL_1 values tend to be smaller when M increases, do not change much when $M \geq 500$, and are almost identical when $M \geq 1,000$ in most cases considered. Based on this example, we

Table 1: ARL_1 values and their standard errors (in parentheses) of the P-CUSUM, L-CUSUM, PS-CUSUM, and LS-CUSUM charts when their nominal ARL_0 values are all fixed at 200, $M = 50$, and the true IC distribution is Poisson(10), NB(10, 0.4), GP(10, 0.4), or GP (10, -0.4).

Underlying distribution	Dispersion shifts	P-CUSUM	PS-CUSUM	L-CUSUM	LS-CUSUM
Poisson (10)	-8	13.4(0.10)	4.1(0.04)	23.5(0.11)	19.8(0.16)
	-6	15.8(0.13)	5.6(0.07)	27.3(0.14)	36.0(0.51)
	-4	28.0(0.26)	34.1(1.71)	47.2(0.29)	141.4(1.24)
	-2	91.8(1.06)	189.6(5.22)	127.2(0.88)	198.7(0.75)
	0	208.6(2.42)	207.9(5.58)	203.4(1.30)	201.4(0.67)
	2	94.1(1.05)	159.9(4.85)	126.9(0.86)	188.6(0.83)
	4	37.6(0.36)	54.6(2.75)	61.2(0.39)	145.6(1.13)
	6	22.0(0.20)	9.9(0.49)	37.7(0.23)	76.9(0.99)
	8	17.1(0.14)	5.8(0.07)	29.5(0.16)	36.9(0.49)
NB (10, 0.4)	-0.39	22.3(0.28)	16.4(0.26)	43.8(0.29)	73.2(0.43)
	-0.3	35.2(0.51)	41.2(1.49)	69.4(0.49)	120.0(0.70)
	-0.2	60.7(1.02)	115.3(3.70)	108.0(0.81)	172.1(0.76)
	-0.1	120.3(2.55)	194.1(5.22)	161.2(1.19)	198.4(0.69)
	0	193.0(4.31)	197.7(5.27)	201.3(1.35)	199.7(0.71)
	0.1	178.4(3.89)	162.4(4.85)	203.9(1.32)	185.6(0.79)
	0.2	123.0(2.30)	112.3(3.80)	182.4(1.19)	166.0(0.84)
	0.3	87.0(1.46)	73.9(2.66)	154.4(1.04)	146.1(0.83)
	0.4	68.8(1.07)	49.4(1.80)	132.1(0.91)	130.2(0.83)
GP (10, 0.4)	-0.4	32.0(0.44)	34.5(1.19)	64.3(0.44)	109.8(0.66)
	-0.3	39.3(0.57)	56.5(2.08)	77.4(0.55)	134.7(0.75)
	-0.2	52.9(0.83)	112.5(3.75)	99.7(0.73)	165.4(0.79)
	-0.1	85.9(1.59)	186.8(5.07)	140.1(1.03)	194.2(0.72)
	0	200.8(4.21)	201.9(5.34)	198.0(1.34)	200.0(0.70)
	0.1	176.2(3.24)	112.5(3.73)	189.5(1.17)	165.7(0.84)
	0.2	68.0(0.93)	32.6(1.10)	115.7(0.75)	110.2(0.75)
	0.3	36.0(0.41)	13.5(0.21)	68.0(0.43)	64.4(0.48)
	0.4	22.5(0.23)	8.3(0.10)	43.3(0.25)	39.0(0.27)
GP (10, -0.4)	-0.4	123.0(1.42)	115.2(3.74)	153.8(0.97)	171.5(0.77)
	-0.3	142.7(1.72)	148.1(4.39)	168.6(1.08)	185.5(0.74)
	-0.2	165.8(2.11)	185.6(5.13)	182.8(1.17)	196.0(0.72)
	-0.1	186.4(2.45)	200.2(5.35)	194.7(1.24)	201.6(0.68)
	0	200.3(2.81)	198.6(5.29)	201.2(1.28)	200.0(0.70)
	0.1	204.8(2.90)	169.2(4.83)	199.2(1.24)	190.1(0.75)
	0.2	178.5(2.51)	117.5(3.81)	185.1(1.14)	172.3(0.82)
	0.3	138.5(1.76)	69.5(2.48)	162.6(1.00)	146.1(0.83)
	0.4	102.3(1.44)	52.0(1.97)	149.6(0.92)	132.4(0.81)

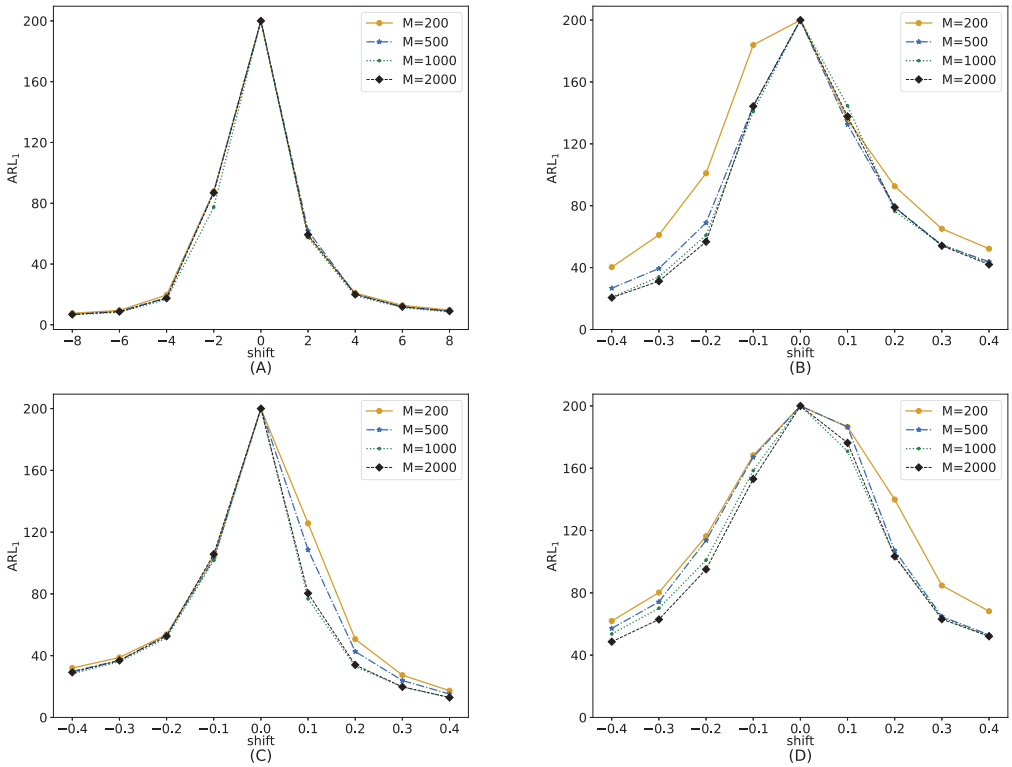

Figure 2: ARL_1 values of the P-CUSUM chart when M=200, 500, 1,000, or 2,000, $k_{\text{P}} = 0.01$, $d = 5$, and the true IC process distribution is (A) Poisson(10), (B) NB(10, 0.4), (C) GP(10, 0.4), or (D) GP (10, -0.4).

suggest that an IC dataset with size $M \geq 500$ is needed if the P-CUSUM chart is used.

Next, we study the impact of the number of categories d used in categorizing the observed count data (cf., Subsection 2.1) on the performance of the P-CUSUM chart. As discussed in Section 1, the information loss due to data categorization is controlled by d. The information loss would be smaller when d is larger. At the same time, the computation would be more time-consuming and there would be more procedure parameters (e.g., the quantiles $\{q_1, q_2, \ldots, q_{2d-1}\}$) to estimate in that case. To study the selection of d, we compute the ARL_1 values of the P-CUSUM chart in cases when d is chosen to be 5, 10, 20, or 30. The IC data size M is fixed at 500 for all charts based on the results in Figure 2, and all other setups are the same as those in the example of Table 1. The results are shown in Figure 3. From the figure, it can be seen that the performance of the P-CUSUM chart is indeed affected by d. In general, the P-CUSUM chart with a larger d would have a better ARL_1 than those with smaller d for detecting dispersion increases, and vice versa for detecting dispersion decreases. In practice, if dispersion increase is our major concern, then we suggest choosing $d \geq 20$. Otherwise, d can be chosen as small as 5.

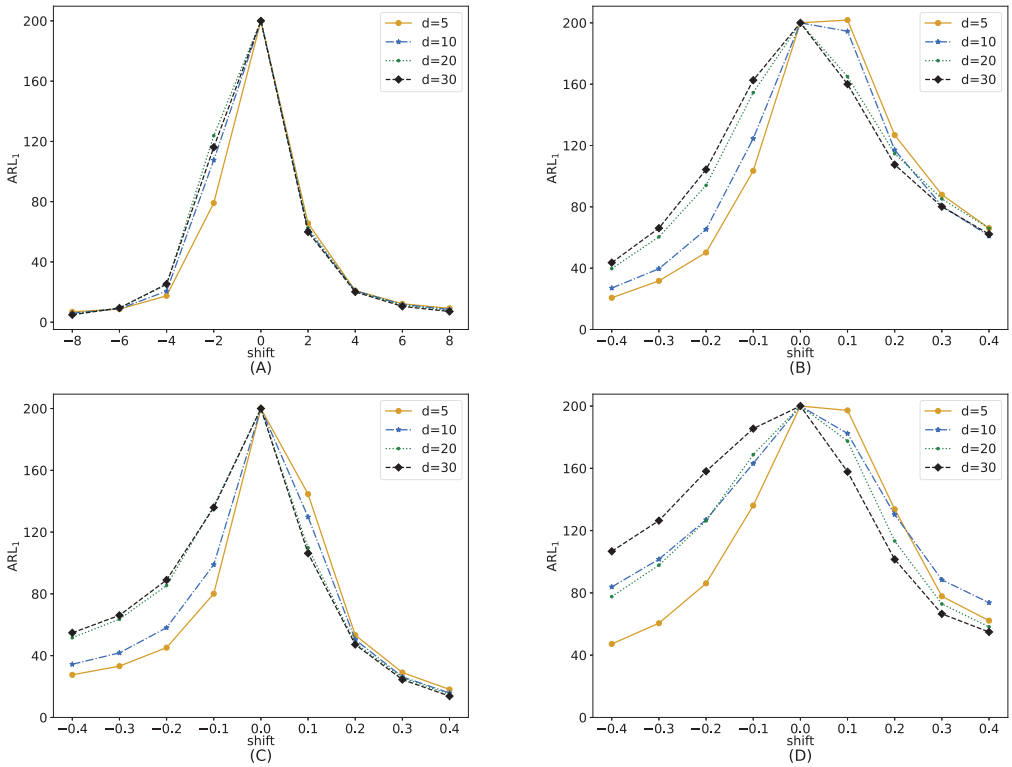

Figure 3: ARL_1 values of the P-CUSUM chart when $d = 5$, 10, 20, or 30, $k_{\text{P}} = 0.01$, $M = 500$, and the true IC process distribution is (A) Poisson(10), (B) NB(10, 0.4), (C) GP(10, 0.4), or (D) GP(10, -0.4).

Finally, we study the impact of the allowance constant k_{P} on the OC performance of the P-CUSUM chart, by changing the value of k_{P} among 0.01, 0.05, 0.1, and 0.2. The IC data size M is fixed at 500, the number of categories d is chosen to be 5, and all other setups are the same as those in the example of Table 1. The calculated ARL_1 values of the P-CUSUM chart are shown in Figure 4. From the figure, it can be seen that the ARL_1 values are uniformly larger when $k_{\text{P}} = 0.2$, suggesting that a smaller k_{P} value should be chosen, which is consistent with the conclusion in Qiu & Li (2011). Based on this example, we suggest choosing $k_{\text{P}} \leq 0.1$.

3.3 Comparison with the methods in Wang & Qiu (2018)

Recall that the observed count data can be categorized either in a small-to-large fashion or in a center-outward fashion. The proposed charts P-CUSUM and L-CUSUM in this paper use the center-outward data categorization, and the two corresponding charts in Wang & Qiu (2018), denoted as P0-CUSUM and L0-CUSUM, use the small-to-large data categorization. In this subsection, we compare their OC performance using the same setting as those in Subsection

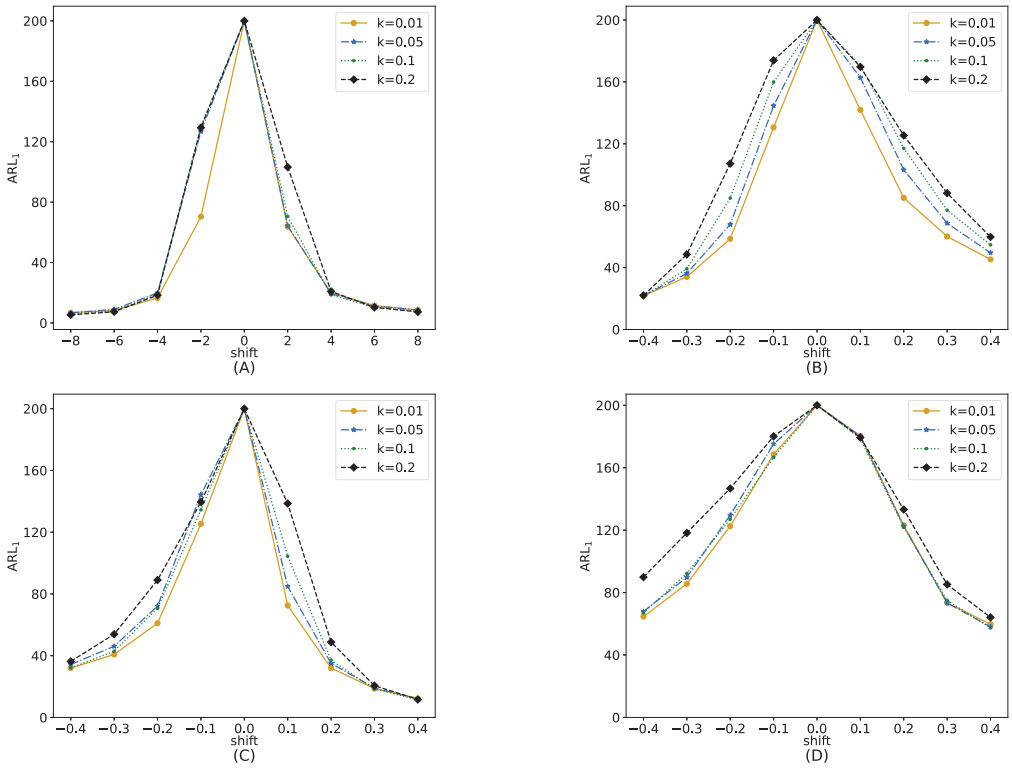

Figure 4: ARL_1 values of the P-CUSUM chart when k_P=0.01, 0.05, 0.1, or 0.2, $d = 5$, $M = 500$, and the true IC process distribution is (A) Poisson(10), (B) NB(10, 0.4), (C) GP(10, 0.4), or (D) GP(10, -0.4).

3.2. In the four charts, their allowance constants are all chosen to be 0.01, their assumed ARL_0 values are set to be 200, and the number of categories d is fixed at 5 in all charts. Their control limits are also searched by the bootstrap procedure with $Q = 100$ maximum iterations. In each iteration, the ARL_0 value is computed from 10,000 replicated simulation runs, in each of which the bootstrap resampling procedure is applied to an IC dataset of size $M = 500$. The IC distribution is still the four cases mentioned above. The ARL_1 values and their standard errors of the four charts in various cases considered are shown in Table 2. From 2, it can be seen that i) all control charts can be used to monitor data dispersion as expected, ii) the P-CUSUM (P0-CUSUM) chart almost always outperforms the L-CUSUM (L0-CUSUM) chart, which is consistent with the conclusion in Wang & Qiu (2018), iii) the P-CUSUM (L-CUSUM) chart performs better than the P0-CUSUM (L0-CUSUM) chart in most cases when the IC process distribution is overdispersed or underdispersed, indicating that the control charts based on center-outward data categorization are generally more powerful for detecting dispersion shifts of count data than the ones based on small-to-large data categorization in such cases, and iv) when the IC process distribution is equidispersed, the

P0-CUSUM (L0-CUSUM) chart could perform better than the P-CUSUM (L-CUSUM) chart.

In summary, although control charts based on categorizing the data from the smallest to the largest can be applied to monitor process changes, they are less powerful for detecting dispersion differences. Therefore, to develop a CUSUM statistic that is more powerful for dispersion changes, we consider making use of the center-outward ordering of the data.

3.4 Comparison with some other alternative methods

In this part, we compare our proposed charts P-CUSUM and L-CUSUM with some representative charts designed for monitoring process dispersion. The first existing method considered is the T-CUSUM chart suggested by Lucas (1985), which is constructed and designed based on the assumption that the process under monitoring has a regular Poisson distribution. We also consider the nonparametric CUSUM chart suggested by Shirke & Barale (2018) for monitoring process dispersion based on a sign test, which is denoted as S-CUSUM. The control limit of T-CUSUM is determined based on the Poisson distribution assumption. For the P-CUSUM, L-CUSUM, and S-CUSUM charts, their control limits are obtained by the bootstrap procedure discussed in Subsection 2.3 with $Q = 100$. All other settings are kept to be the same as those in the example of Table 2. Based on 10,000 replicated simulations, the ARL_1 values of the four charts are presented in Figure 5. From the figure, it can be seen that i) the proposed P-CUSUM chart outperforms the S-CUSUM and L-CUSUM charts in all cases, ii) the S-CUSUM chart has a better performance than the L-CUSUM chart in most cases considered, and iii) the parametric T-CUSUM chart performs the best among the four charts in plot (A) when its Poisson distribution assumption is valid, and performs poorly in all three other plots when its distributional assumption is violated.

This example shows that nonparametric charts can generally provide more robust and effective monitoring of count data dispersion when the pre-specified parametric distribution assumption is violated. Among the three nonparametric charts considered here, the proposed P-CUSUM chart performs the best in most cases considered.

4 An application

In this section, we apply the proposed control charts discussed in the previous sections to a real data example about the daily confirmed cases of Coronavirus Disease 2019 (COVID-19) in China from October 31, 2020 to January 31, 2021. This dataset can be found on the

Table 2: The ARL_1 values and their standard errors (in parentheses) of the P-CUSUM, L-CUSUM, P0-CUSUM, and L0-CUSUM charts when their nominal ARL_0 values are all fixed at 200 and the true IC distributions is Poisson (10), NB (10, 0.4), GP (10, 0.4), or GP (10, -0.4).

Underlying distribution	Dispersion shifts	P-CUSUM	P0-CUSUM	L-CUSUM	L0-CUSUM
Poisson (10)	-8	7.1(0.08)	9.5(0.09)	17.5(0.09)	20.4(0.09)
	-6	9.4(0.12)	10.8(0.10)	23.0(0.14)	23.3(0.11)
	-4	19.6(0.31)	17.5(0.18)	48.5(0.35)	36.7(0.20)
	-2	112.8(3.16)	46.4(0.61)	165.7(1.05)	92.5(0.56)
	0	204.2(6.07)	196.8(4.88)	200.2(0.99)	198.4(1.27)
	2	56.2(1.26)	22.2(0.35)	119.5(0.89)	53.6(0.41)
	4	19.2(0.30)	11.8(0.17)	48.3(0.36)	29.3(0.20)
	6	11.7(0.16)	9.0(0.11)	29.0(0.20)	22.3(0.14)
	8	9.1(0.12)	8.0(0.10)	22.2(0.14)	19.4(0.12)
NB (10, 0.4)	-0.39	21.7(0.32)	20.1(0.29)	51.5(0.32)	46.2(0.31)
	-0.3	34.0(0.55)	32.6(0.55)	77.8(0.55)	74.2(0.55)
	-0.2	58.6(1.21)	60.0(1.29)	119.1(0.86)	119.1(0.87)
	-0.1	130.6(3.53)	133.9(3.54)	175.2(1.08)	171.8(1.08)
	0	194.5(5.46)	192.2(5.64)	200.8(1.04)	200.2(1.04)
	0.1	141.9(3.98)	164.5(4.60)	180.1(1.05)	187.4(1.06)
	0.2	85.1(2.09)	109.7(2.86)	150.1(1.02)	164.4(1.02)
	0.3	60.1(1.29)	73.4(1.66)	124.0(0.93)	139.4(0.98)
	0.4	45.3(0.91)	57.5(1.14)	103.8(0.77)	119.7(0.87)
GP (10, 0.4)	-0.4	32.0(0.51)	34.7(0.58)	73.6(0.49)	79.3(0.55)
	-0.3	40.8(0.69)	46.7(0.84)	90.8(0.61)	100.1(0.70)
	-0.2	61.0(1.17)	71.2(1.50)	120.6(0.81)	134.4(0.90)
	-0.1	125.4(3.28)	147.6(3.93)	169.4(1.02)	182.7(1.02)
	0	191.5(5.74)	207.9(5.99)	198.3(0.95)	200.8(0.99)
	0.1	72.5(1.75)	85.6(2.20)	136.2(1.01)	147.2(1.04)
	0.2	32.0(0.61)	39.1(0.74)	77.8(0.64)	89.3(0.72)
	0.3	18.7(0.31)	23.5(0.38)	47.1(0.37)	56.5(0.45)
	0.4	12.3(0.18)	15.0(0.23)	31.2(0.23)	37.2(0.27)
GP (10,-0.4)	-0.4	57.2(1.14)	63.6(1.36)	115.6(0.80)	128.7(0.93)
	-0.3	74.2(1.65)	81.4(1.91)	136.3(0.93)	148.2(1.03)
	-0.2	113.7(2.89)	114.6(3.07)	161.3(1.03)	169.9(1.11)
	-0.1	176.1(4.84)	164.3(4.65)	187.0(1.04)	190.5(1.16)
	0	209.4(6.02)	193.8(5.45)	201.7(1.00)	202.0(1.14)
	0.1	186.6(5.28)	186.1(5.22)	193.1(1.02)	199.3(1.14)
	0.2	107.1(2.67)	138.1(3.62)	163.4(1.02)	182.7(1.11)
	0.3	64.8(1.36)	90.8(2.04)	129.2(0.91)	155.6(1.05)
	0.4	53.0(1.05)	75.2(1.61)	114.1(0.83)	142.4(0.97)

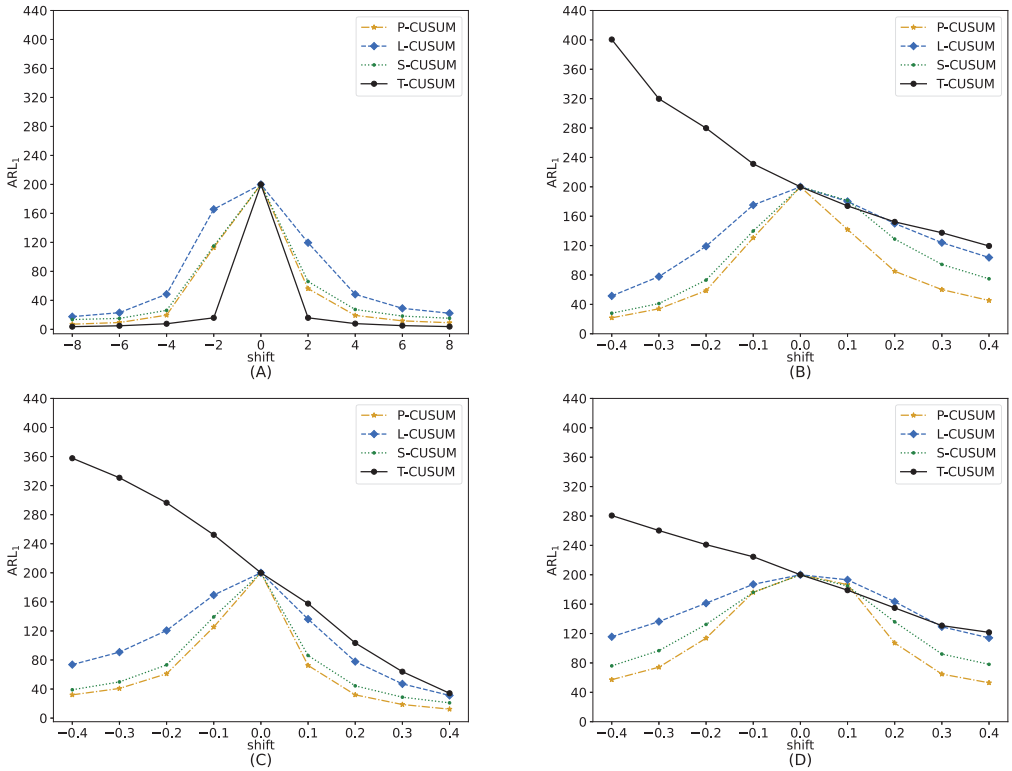

Figure 5: ARL_1 values of the four control charts when $ARL_0 = 200$, $M = 500$, $d = 5$, and the true IC process distribution is (A) Poisson(10), (B) NB(10, 0.4), (C) GP(10, 0.4), or (D) GP(10, -0.4).

web page http://www.nhc.gov.cn/, and is shown in Figure 6. It can be seen from the figure that the fluctuation in the observed counts of daily confirmed COVID-19 cases from October 31, 2020 to January 5, 2021 is much smaller than the fluctuation in the observed counts after January 5, 2021. It seems that the process mean and dispersion changed after January 5, 2021. In fact, the data variance of the daily confirmed cases from October 31, 2020 to January 5, 2021 is 67.53, and it increases to 1,108.40 in the time period afterward since there was a big disease outbreak in a region of Shijiazhuang, Hebei Province of China on January 6, 2021. Thus, the data collected between October 31, 2020 and January 5, 2021 are used as the IC data in this example, and the data collected afterward are used for online process monitoring. The vertical dashed line in Figure 6 separates the IC dataset from the data for process monitoring.

Figure 7 shows the density histogram of the daily confirmed cases of COVID-19 in China from October 31, 2020 to January 5, 2021 along with the estimated density curve (solid) and the density curve of a Poisson distribution (dashed) with the same mean as that of the IC data. From the figure, it can be seen that the distribution of the observed IC data is quite

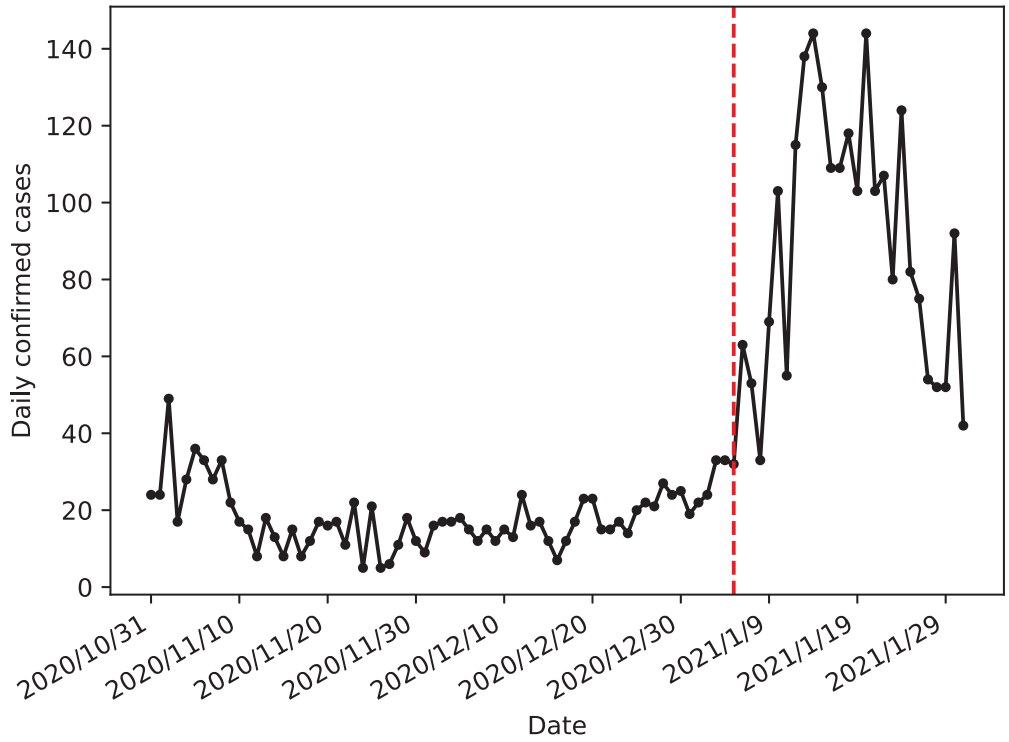

Figure 6: Daily confirmed cases of COVID-19 in China from October 31, 2020 to January 31, 2021.

different from the Poisson distribution. The Poisson goodness-of-fit test gives a P value of 0.000, implying that the distribution of the observed data is indeed significantly different from a Poisson distribution. Therefore, the three nonparametric charts P-CUSUM, L-CUSUM, and S-CUSUM are considered here.

When implementing the related control charts, we fix $\mathrm{ARL}_0 = 200$ for all charts, and choose $d = 5$ for the P-CUSUM and L-CUSUM charts, the subgroup size of 3 for the SCUSUM chart, and the allowance constants to be 0.01 for all three nonparametric charts. The three control charts are shown in Figure 8, where the red horizontal dashed lines denote the control limits of the corresponding control charts. From the plots, it can be seen that the P-CUSUM, L-CUSUM, and S-CUSUM charts give their first signals at the 9th, 26th, and 15th observation times, respectively, during online process monitoring. So the P-CUSUM chart is the most effective one for detecting process dispersion shift in this example.

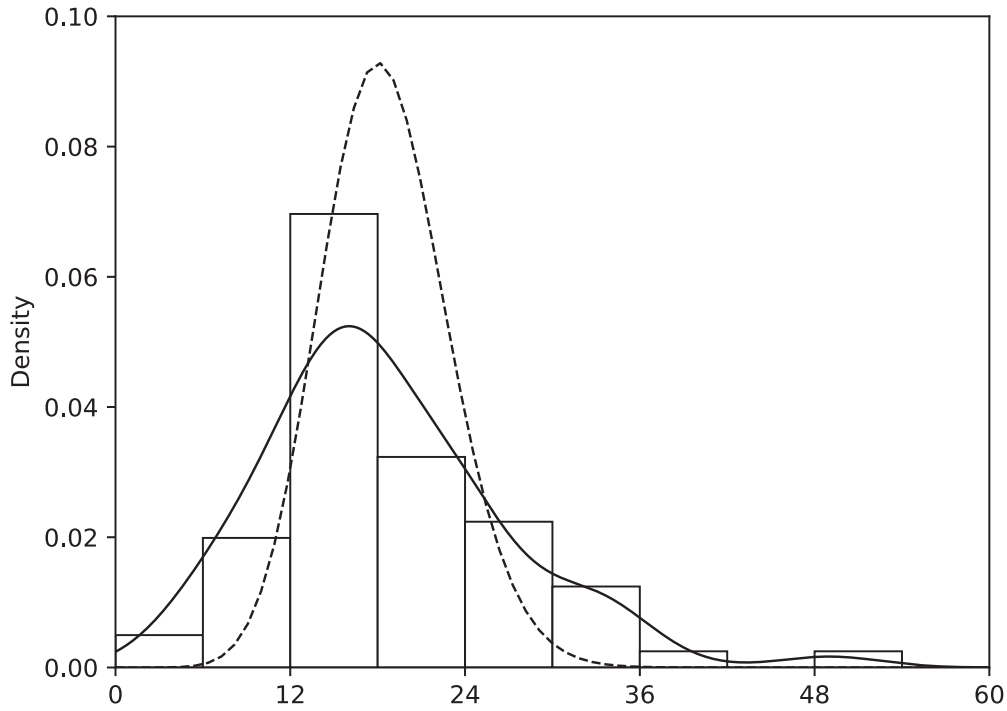

Figure 7: Density histogram with the estimated density curve (solid) of the daily confirmed cases of COVID-19 in China from October 31, 2020 to January 5, 2021, together with the density curve of a Poisson distribution (dashed) with the same mean as that of the IC data.

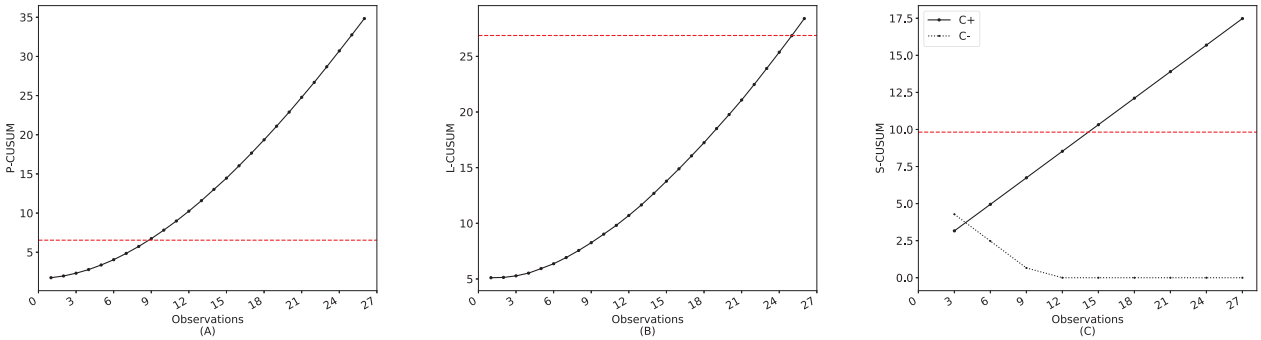

Figure 8: The P-CUSUM, L-CUSUM, and S-CUSUM charts for monitoring the observed counts of daily confirmed cases of COVID-19 in China are shown in Figure 6. The horizontal dashed line in each plot denotes the control limit of the related chart.

5 Concluding remarks

Statistical process control for count data is important because count data are a basic data format in practice. In the literature, most existing control charts for monitoring count data are based on assumed parametric probability models and focus more on detecting location shifts. In various applications, however, the assumed parametric models are rarely valid, due to the complicated impact of various confounding factors (e.g., the environment, weather, and

so forth) on the count data. We have shown in this paper that parametric control charts are unreliable to use in such cases, as their actual ARL_0 values could be substantially different from a pre-specified nominal ARL_0 level. In this paper, we suggest using nonparametric charts instead. Moreover, it is also highly desirable and important to monitor the dispersion of count data. To this end, new nonparametric control charts and their self-starting versions are suggested in this paper for monitoring the dispersion of count data. Numerical results show that the proposed P-CUSUM control chart is more effective and robust than the competing charts L-CUSUM, S-CUSUM, and T-CUSUM in various cases considered.

Several issues have not been discussed thoroughly in the current paper about the proposed method, and are worth further study in future research. First, we focus solely on online monitoring of the process dispersion in this paper. In some applications, simultaneous online monitoring of the process mean and dispersion would be our interest. Second, our proposed control charts do not require the specification of the IC distribution. However, it is assumed that an IC dataset is available for estimating certain IC parameters, which requires a proper Phase I analysis that has not been discussed in this paper. Third, the proposed control charts involve some procedure parameters such as the allowance constant. Proper selection of these parameters is important, and adaptive CUSUM charts to accommodate their data-driven estimates might be of interest for future research.

Acknowledgements

The authors thank the editors and one referee for many helpful comments and suggestions, which improved the quality of the paper greatly. This work was supported in part by the National Natural Science Foundation of China under grant numbers 71902138, 72002066, 72231005, and 72032005; the Humanity and Social Science Youth Foundation of the Ministry of Education of China under grant number 19YJC630181; and the State Administration of Science, Technology and Industry for National Defense under grant numbers JSZL2021204B001 and JSZL2022204B005.

References

Ansari, A. R. & Bradley, R. A. (1960). Rank-sum tests for dispersions. *The Annals of Mathematical Statistics*, 31(4), 1174–1189.

Bliss, C. I. & Owen, A. R. (1958). Negative binomial distributions with a common k. *Biometrika*, 45(1/2), 37–58.

Borror, C. M., Champ, C. W., & Rigdon, S. E. (1998). Poisson EWMA control charts. *Journal of Quality Technology*, 30(4), 352–361.

Bourguignon, M., Medeiros, R. M., Fernandes, F. H., & Lee Ho, L. (2021). Simple and useful statistical control charts for monitoring count data. *Quality and Reliability Engineering International*, 37(2), 541–566.

Chatterjee, S. & Qiu, P. (2009). Distribution-free cumulative sum control charts using bootstrap-based control limits. *The Annals of Applied Statistics*, 3(1), 349–369.

Chong, Z. L., Mukherjee, A., & Khoo, M. B. (2018). Some distribution-free Lepage-type schemes for simultaneous monitoring of one-sided shifts in location and scale. *Computers & Industrial Engineering*, 115, 653–669.

Chowdhury, S., Mukherjee, A., & Chakraborti, S. (2014). A new distribution-free control chart for joint monitoring of unknown location and scale parameters of continuous distributions. *Quality and Reliability Engineering International*, 30(2), 191–204.

Chowdhury, S., Mukherjee, A., & Chakraborti, S. (2015). Distribution-free Phase II CUSUM control chart for joint monitoring of location and scale. *Quality and Reliability Engineering International*, 31(1), 135–151.

Collings, B. J. & Margolin, B. H. (1985). Testing goodness of fit for the Poisson assumption when observations are not identically distributed. *Journal of the American Statistical Association*, 80(390), 411–418.

Consul, P. C. & Jain, G. C. (1973). A generalization of the Poisson distribution. *Technometrics*, 15(4), 791–799.

Das, N. & Bhattacharya, A. (2008). A new non-parametric control chart for controlling variability. *Quality Technology & Quantitative Management*, 5(4), 351–361.

Fang, Y. (2003). C-charts, X-charts, and the Katz family of distributions. *Journal of Quality Technology*, 35(1), 104–114.

Gan, F. F. (1990). Monitoring Poisson observations using modified exponentially weighted moving average control charts. *Communication in Statistics- Simulation and Computation*, 19(1), 103–124.

Godase, D. G., Rakitzis, A. C., Mahadik, S. B., & Khoo, M. B. (2022). Deciles-based EWMA-type sign charts for process dispersion. *Quality and Reliability Engineering International*, 38(7), 3726–3740.

Haq, A. (2017). A new nonparametric EWMA control chart for monitoring process variability. *Quality and Reliability Engineering International*, 33(7), 1499–1512.

He, B., Xie, M., Goh, T., & Tsui, K. (2006). On control charts based on the generalized Poisson model. *Quality Technology & Quantitative Management*, 3(4), 383–400.

Ho, L. L., Andrade, B., Bourguignon, M., & Fernandes, F. H. (2021). Monitoring count data with Shewhart control charts based on the Touchard model. *Quality and Reliability Engineering International*, 37(5), 1875–1893.

Jalilibal, Z., Amiri, A., & Khoo, M. B. (2022). A literature review on joint control schemes in statistical process monitoring. *Quality and Reliability Engineering International*, 38(6), 3270–3289.

Jesus, B. D., Ferreira, P. H., Boaventura, L. L., Fiaccone, R. L., Bertoli, W., Ramos, P. L., & Louzada, F. (2022). Statistical process control of overdispersed count data based on one-parameter Poisson mixture models. *Quality and Reliability Engineering International*, 38(5), 2324–2344.

Khilare, S. & Shirke, D. (2012). Nonparametric synthetic control charts for process variation. *Quality and Reliability Engineering International*, 28(2), 193–202.

Li, J. (2021). Nonparametric adaptive CUSUM chart for detecting arbitrary distributional changes. *Journal of Quality Technology*, 53(2), 154–172.

Liang, W., Mukherjee, A., Xiang, D., & Xu, Z. (2022). A new nonparametric adaptive EWMA procedures for monitoring location and scale shifts via weighted Cucconi statistic. *Computers & Industrial Engineering*, 170, 108321.

Lucas, J. M. (1985). Counted data CUSUM's. *Technometrics*, 27(2), 129–144.

Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1), 50–60.

Mukherjee, A. & Chakraborti, S. (2012). A distribution-free control chart for the joint monitoring of location and scale. *Quality and Reliability Engineering International*, 28(3), 335–352.

Mukherjee, A. & Marozzi, M. (2017). A distribution-free Phase-II CUSUM procedure for monitoring service quality. *Total Quality Management & Business Excellence*, 28(11-12), 1227–1263.

Qiu, P. (2014). *Introduction to statistical process control*. Chapman & Hall/CRC: Boca Raton, FL.

Qiu, P. (2018). Some perspectives on nonparametric statistical process control. *Journal of Quality Technology*, 50(1), 49–65.

Qiu, P. & Li, Z. (2011). On nonparametric statistical process control of univariate processes. *Technometrics*, 53(4), 390–405.

Ross, G. J., Tasoulis, D. K., & Adams, N. M. (2011). Nonparametric monitoring of data streams for changes in location and scale. *Technometrics*, 53(4), 379–389.

Saghir, A. & Lin, Z. (2014a). Cumulative sum charts for monitoring the COM-Poisson processes. *Computers & Industrial Engineering*, 68, 65–77.

Saghir, A. & Lin, Z. (2014b). A flexible and generalized exponentially weighted moving average control chart for count data. *Quality and Reliability Engineering International*, 30(8), 1427–1443.

Saghir, A. & Lin, Z. (2015). Control charts for dispersed count data: an overview. *Quality and Reliability Engineering International*, 31(5), 725–739.

Sellers, K. F. (2012). A generalized statistical control chart for over-or under-dispersed data. *Quality and Reliability Engineering International*, 28(1), 59–65.

Shirke, D. T. & Barale, M. S. (2018). A nonparametric CUSUM chart for process dispersion. *Quality and Reliability Engineering International*, 34(5), 858–866.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1), 127–142.

Tang, L. & Li, J. (2023). A nonparametric control chart for monitoring count data mean. *Quality and Reliability Engineering International*, 39(7), 1–15.

Tercero-Gómez, V., Aguilar-Lleyda, V., Cordero-Franco, A., & Conover, W. (2020). A distribution-free CUSUM chart for joint monitoring of location and scale based on the combination of Wilcoxon and Mood statistics. *Quality and Reliability Engineering International*, 36(4), 1422–1453.

Villanueva-Guerra, E. C., Tercero-Gómez, V. G., Cordero-Franco, A. E., & Conover, W. J. (2017). A control chart for variance based on squared ranks. *Journal of Statistical Computation and Simulation*, 87(18), 3537–3562.

Wang, Z. & Qiu, P. (2018). Count data monitoring: Parametric or nonparametric? *Quality and Reliability Engineering International*, 34(8), 1763–1774.

White, C. H. & Keats, J. B. (1996). ARLs and higher-order run-length moments for the Poisson CUSUM. *Journal of Quality Technology*, 28(3), 363–369.

Yang, S.-F. & Arnold, B. C. (2016). A new approach for monitoring process variance. *Journal of Statistical Computation and Simulation*, 86(14), 2749–2765.

Ye, H. & Liu, K. (2022). A generic online nonparametric monitoring and sampling strategy for high-dimensional heterogeneous processes. *IEEE Transactions on Automation Science and Engineering*, 19(3), 1503–1516.

Ye, H., Zheng, Z., Cheng, J.-R. C., Hable, B., & Liu, K. (2023). Online monitoring of high-dimensional asynchronous and heterogeneous data streams for shifts in location and scale. *International Journal of Production Research*, (pp. 1–17).

Zaman, B. (2021). Efficient adaptive CUSUM control charts based on generalized likelihood ratio test to monitor process dispersion shift. *Quality and Reliability Engineering International*, 37(8), 3192–3220.

Zhou, M., Zhou, Q., & Geng, W. (2016). A new nonparametric control chart for monitoring variability. *Quality and Reliability Engineering International*, 32(7), 2471–2479.

Author biographies

Zhiqiong Wang is currently an associate professor of the School of Management at the Tianjin University of Technology, China. He received his Ph.D. degree in management science and engineering from Tianjin University, China, in 2018. His major research interests include quality control and management, statistical process control, and change-point detection.

Qingxia Wu is a Master Degree Candidate of the School of Management at the Tianjin University of Technology, China. Her major research interests are quality management and statistical process control.

Peihua Qiu received his PhD in statistics from the Department of Statistics at the University of Wisconsin at Madison in 1996. He worked as a senior research consulting statistician of the Biostatistics Center at Ohio State University during 1996-1998. Then, he worked as an assistant professor (1998-2002), an associate professor (2002-2007), and a full professor (2007-2013) of the School of Statistics at the University of Minnesota. He has been a professor and the founding chair of the Department of Biostatistics at the University of Florida since 2013. Qiu is an elected fellow of the American Association for the Advancement of Science (AAAS), an elected fellow of the American Society for Quality (ASQ), an elected fellow of the American Statistical Association (ASA), an elected fellow of the Institute of Mathematical Statistics (IMS), and an elected member of the International Statistical Institute (ISI). He served as an associate editor for a number of top journals in statistics, including Journal of the American Statistical Association, Biometrics, and Technometrics. He was the editor of Technometrics during 2014-2016. Qiu has made substantial contributions in the areas of jump regression analysis, image processing, statistical process control, survival analysis, disease screening, and disease surveillance. So far, he has published two research monographs and over 160 research papers in refereed journals.