Weighted Likelihood Ratio Charts for Statistical Monitoring of Queueing Systems

Dequan Qi\(^1\), Zhonghua Li\(^1\), Xuemin Zi\(^2\), and Zhaojun Wang\(^1\)*

\(^1\)LPMC and Institute of Statistics, Nankai University, China
\(^2\)School of Science, Tianjin University of Technology and Education, China

Abstract

In recent years, statistical monitoring for effective detection of queueing system has increasingly attracted attention of researchers. However, most existing research did not consider the data autocorrelation, nor evaluate rigorously the performance either. In this paper, considering the data autocorrelation and motivated by giving higher weights to recent data, a new control chart based on the weighted likelihood ratio test (WLRT) is proposed to efficiently monitor the utilization of M/M/1 queueing system. Our approach can be naturally extended to other general queueing systems if we can obtain the likelihood function based on information available according the queueing theory. Simulation results show, as expected, that the performance of the proposed WLRT-chart is satisfactory by comparing with several alternative methods.

**Keywords**: Statistical Process Control; Average Run Length; Cumulative Sum Charts; Markov Chain; Queue Length.

1 Introduction

Statistical monitoring for effective detection of the parameters of queueing system, such as the utilization, the service rate and the arrival rate, has increasingly attracted attention of researchers. Out of question, such detection can be further used to assist root-cause identification and decision making for service-operation improvement. However, there are several challenges to apply the statistical process control (SPC) in the queueing systems. Firstly, the distribution of the observations from queueing systems is typically unknown and is often highly skewed (Shore (2006)). Secondly, the observations collected from queueing

*Corresponding author. Email: zjwang@nankai.edu.cn
systems are often auto-correlated (Chen et al. (2011)). Regrettably, ignoring the date autocorrelation may influence the monitoring performance effectively (Tsung et al. (2006)). At last but not least, the observations from queueing systems are only partially available in many situations (Chen and Zhou (2013)). Because of these challenges, only limited literatures focus on monitoring of queueing systems.

The relevant research on this topic can be divided into two groups, depending on whether the data autocorrelation is considered. The first group of monitoring methods ignored the autocorrelation and focused on the partial sampling scheme, and they observed queue length $Q_n$ immediately after the $n^{th}$ departure epoch (e.g., Bhat and Rao (1972), Bhat (1987), Shore (2006)). To evaluate the efficacy of the existing monitoring methods, Chen et al. (2011) proposed an analytical method based on Markov Chain model, and investigated the performance of the WZ chart and the nL chart, where the WZ chart is the control chart proposed by Bhat and Rao (1972), and the nL chart is a simple extension of the chart proposed by Shore (2006). The second group of monitoring method not only considered the data autocorrelation but also investigated both the partial sampling scheme and the complete sampling scheme. In the complete sampling scheme, they observed both queue length and system times at arrival and departure epochs. Considering the data autocorrelation, Chen and Zhou (2013) proposed the cumulative sum (CUSUM) scheme to monitor typical queueing systems, in particular the M/M/1 queueing system. However, the performance of the CUSUM chart might be deteriorated if the real out-of-control (OC) parameters were far from the designated region. Therefore, Chen and Zhou (2013) suggested using the multiple CUSUM charts with different design parameters or the generalized likelihood ratio (GLR) chart to mitigate the problem. However, the GLR charts cannot be updated recursively and had to be computed by maximizing the likelihood ratio with respect to all possible change locations which may lead to the computational load increase significantly. And the GLR charts perform worse than the CUSUM charts when detecting changes that the CUSUM charts are designed for.

All of the afore-mentioned research did not consider the fact that recent data may carry more up-to-date information. Many researchers have shown that giving higher weights to recent data can lead to better monitoring performance. That is why the weighted CUSUM (WCUSM) charts (e.g., Yashchin (1989), Shu et al. (2008a, b), Jiang et al. (2011)) and the exponentially weighted moving average (EWMA) charts (e.g., Robert (1959), Grigg (2007), Qiu et al. (2010)) have been widely applied. On the other hand, the control charts via a likelihood-ratio test (LRT) often outperform other methods and have been effectively used in phase I or phase II control chart (e.g., Paynabar et al. (2012), Zhou et al. (2010, 2012)).

In this paper, motivated by giving higher weights to recent data, we focus on the partial sampling scheme in the M/M/1 queueing system. Since the queue length $Q_{n-1}$ and $Q_n$ are dependent due to the queueing dynamic, we observe $A_n = Q_n - Q_{n-1} + 1 - Z_{n-1}$, which is the number of arrivals during the $n^{th}$ service period, where $Z_{n-1}$ is an indicator variable which equals 1 if $Q_{n-1} = 0$ and equals 0 otherwise. According the queueing theory, $A_n$'s
have the independently identically distribution. Then, we develop a control chart via the weighted LRT (WLRT) (Zhou et al. (2012)), and propose a Markov Chain method similar to that described by Lucas and Saccucci (1990) to compute the average run length (ARL). Henceforth, the proposed control chart is simply called WLRT chart. As for the complete sampling scheme, we can follow the similar procedure. By the way, compared with the complete sampling scheme, the partial sampling scheme is easy and/or inexpensive.

We would like to stress that our approach can be naturally extended to more general types of queues if we can express the likelihood function according the queueing theory. Based on our simulation, for WLRT control chart, a small value of the smoothing parameter leads to optimal detection of small shifts. If practitioners do not have any particular preference, we suggest choosing smoothing parameter as 0.05 or using the multiple WLRT charts with different smoothing parameters. Alternatively, we present a control chart based on the maximum weighted likelihood estimate (MWLE) for the M/M/1 queueing system. By simulation, we find that the performance of the proposed WLRT chart is better then other competitors.

The rest of the paper is organized as follows. In the next section, the statistical model and the WLRT chart for M/M/1 queueing system are introduced. A MWLE chart is developed as an alternative method. The following section is devoted to comparing, through Monte Carlo simulations, the performance of the proposed WLRT chart with the WZ (Bhat and Rao (1972)), nL (Chen et al. (2011)), CUSUM and GLR (Chen and Zhou (2013)) charts. Finally, an illustrative example and our conclusions are given. The proof of properties of our proposed control chart and derivation of Markov chain method for computing ARL are deferred in Appendix A and B, respectively.

2 Our proposed WLRT chart

The M/M/1 queueing system is a classical Poisson-input, exponential-service, single-server queue (Gross and Harris (1998)), and the arrival rate, the service rate, the utilization is denoted by $\lambda, \mu, \rho$ respectively, where $\rho = \frac{\lambda}{\mu}$. Our statistical model supposes $\rho$ changes from $\rho_0$ to another unknown value $\rho_1$ immediately after an unknown departure epoch $\tau$. It suffices to test the following hypotheses

\[
\begin{align*}
H_0 : \rho &= \rho_0, \\
H_1 : \rho &\neq \rho_0,
\end{align*}
\]

after each departure epoch.

According to the queueing theory, the number of arrivals during the $n^{th}$ service period,
$A_n$ has the following distribution (Chen and Zhou (2013))

$$Pr\{A_n = k\} = \frac{1}{1 + \rho} \cdot \left(\frac{\rho}{1 + \rho}\right)^k, \ k = 0, 1 \ldots$$

Apparently, after any departure epoch $N$, the weighted-log-likelihood function can be written as

$$l_N(\rho) = \ln \rho \cdot \sum_{n=0}^{N} w_n A_n - \ln(1 + \rho) \cdot \sum_{n=0}^{N} w_n (A_n + 1), \quad (2.1)$$

where the weight $w_0 = (1 - \theta)^N$, $w_n = \theta(1 - \theta)^{N-n}, n = 1, \ldots N$, such that $\sum_{n=0}^{N} w_n = 1$, and $\theta \in (0, 1)$ is a smoothing parameter, which is similar to Qiu et al. (2010) and Zhou et al. (2012), but with a few differences. Including $w_0$ and $A_0$ in equation (2.1) has its own merit, because $A_0$ can be viewed as a pseudo “sample”, and is chosen as $\rho_0$ here, as $E(A_n) = \rho_0$ under null hypothesis. Given the value of $\theta$, we can obtain the MWLE of $\rho$

$$\hat{\rho}_N = \arg \max_{\rho} l_N(\rho) = \sum_{n=0}^{N} w_n A_n. \quad (2.2)$$

Furthermore, we can express the WLRT statistic as,

$$W_N = 2[l_N(\hat{\rho}_N) - l_N(\rho_0)] = 2[\hat{\rho}_N \cdot \ln \frac{\hat{\rho}_N (1 + \rho_0)}{\rho_0 (1 + \hat{\rho}_N)} - \ln \frac{1 + \hat{\rho}_N}{1 + \rho_0}]. \quad (2.3)$$

When the WLRT statistic in (2.3) is larger than the upper control limit (UCL), we can declare the system utilization $\rho$ has deviated from the nominal value, which means the system is OC.

By some simple algebra (see Appendix A), we get the following properties immediately.

**P1.** $\hat{\rho}_N$ can be updated recursively,

$$\hat{\rho}_N = \hat{\rho}_{N-1} \cdot (1 - \theta) + \theta \cdot A_N,$$

where the initial values are $\hat{\rho}_0 = \rho_0$ based on $A_0$ and $w_0$ defined above. Hence, the computational load will decrease significantly.

**P2.**

$$E(\hat{\rho}_N) = \rho, Var(\hat{\rho}_N) = (\rho + \rho^2) \sum_{n=1}^{N} w_n^2. \quad (2.4)$$
P3. Under null hypothesis,
\[
\hat{\rho}_N - \rho_0 \over \sqrt{(\rho_0 + \rho_0^2) \sum_{n=1}^{N} w_n^2} \rightarrow^d N(0, 1),
\]  
(2.5)
as \theta N \rightarrow \infty \text{ and } \theta \rightarrow 0.

P4. WLRT statistic $W_N < UCL$ (the system is in-control (IC)) is essentially equivalent to
\[
a < \hat{\rho}_N < b,
\]  
(2.6)where $a, b$ ($a < \rho_0 < b$) are the real roots of the equation $W_N = UCL$. This property makes our proposed WLRT chart look like the traditional EWMA chart noting that $\hat{\rho}_N$ admits the classical EWMA updating formulas. So the ARL results of the WLRT chart may be closely approximated using the Markov chain method (see the Appendix B). It is worth pointing out that, we use the WLRT chart in this paper rather than the EWMA chart because the WLRT chart can be extended to complete sampling scheme easily.

Similar to the traditional EWMA chart, for our WLRT chart, a small value of $\theta$ leads to optimal detection of small shifts. Based on our simulation results, we suggest choosing $\theta$ smaller than 0.05 when detect small downward shift. If the practitioners are not certain which patterns of changes they are most interested in detecting, the multiple WLRT charts with different smooth parameters can be deployed. Alternatively, we present a control chart based on MWLE. Based on equation (2.2), (2.4) and $\sum_{n=1}^{\infty} w_n^2 = \theta/(2 - \theta)$, we propose the following monitoring statistic
\[
M_N = \max_{\theta} \left| \frac{\hat{\rho}_N - \rho_0}{\sqrt{(\rho_0 + \rho_0^2)\theta/(2 - \theta)}} \right|,
\]for all $\theta \in [c, d]$, where $c$ and $d$ are some critical values. The corresponding control chart will trigger OC signal when $M_N$ is larger than UCL. In order to simplify the calculation, considering the performance when detect small downward shift, we only choose $\theta = 0.005, 0.01, 0.02$ to calculate $M_N$ in Section 3.

In practice, we may be only interested in decrease of the service rate and/or increase of the arrival rate, so we develop a one-sided chart for the hypotheses
\[
\begin{align*}
H_0 &: \rho = \rho_0, \\
H_1 &: \rho > \rho_0.
\end{align*}
\]Following Zhou et al. (2012), by substituting $\hat{\rho}_N = \hat{\rho}_N I(\hat{\rho}_N > \rho_0) + \rho_0 I(\hat{\rho}_N \leq \rho_0)$ (Shu et al. 2012) into (2.3), the monitoring statistic can be modified by
\[
W_N^+ = W_N I(\hat{\rho}_N > \rho_0),
\]when $W_N^+$ is larger than UCL, the corresponding control chart generates OC signal.
3 Performance Comparisons

In this section, the performance of the proposed WLRT$^+$ and MWLE charts for the partial sampling scheme in M/M/1 systems are evaluated through Monte Carlo simulations. We find that the ARL results have slightly different between using the Markov chain method and the Monte Carlo simulation. For instance, by the Markov chain method ($a=0$, $m=50$), we get $\text{ARL}_0 = 375$ (in control ARL) when $\theta = 0.1$, $\rho_0 = 0.7$ and UCL=0.3146. While by the Monte Carlo simulation (20,000 replications), we obtain $\text{ARL}_0 = 370$. The OC ARL (termed $\text{ARL}_1$) and the OC average number of samples (termed $\text{ANOS}_1$) are two criteria used for the performance comparison. To detect the same amount of utilization shifts, a smaller value of $\text{ARL}_1$ or $\text{ANOS}_1$ often suggests a quicker response to the changes. To assess how long it takes the control charts to signal alarms, two scenarios with the IC utilization $\rho_0 = 0.5$ and $\rho_0 = 0.7$ are tested. When $\rho_0 = 0.5$, we only compare the $\text{ARL}_1$ of the WLRT$^+$ chart and the CUSUM chart because Chen and Zhou (2013) have revealed that the CUSUM chart outperforms the $nL$ chart and WZ chart in this scenario; when $\rho_0 = 0.7$, we compare the $\text{ANOS}_1$ of the WLRT$^+$ chart with $nL$ chart, WZ chart and CUSUM chart (the calculated $\text{ARL}_1$ has been converted to the corresponding $\text{ANOS}_1$ for fair comparison between these four charts). To compare the MWLE chart with GLR chart, the system utilization in both scenarios increases from 0.02 to 1.96 and the corresponding $\text{ARL}_1$ are computed and compared.

We use the following Monte Carlo simulation to determine the control limit of WLRT$^+$ chart. To generate the simulated data, we also suppose $Q_0 = 0$, $Z_0 = 1$, which indicates the monitored system starts from an empty queue. The monitored system starts from an empty queue is common practice (Chen et al. (2011)) and an empty queue is a regeneration point in M/M/1 systems. When $\rho_0 = 0.5$, given the UCL, let the arrival rate $\lambda = 0.5$, the service rate $\mu = 1.0$ to generate 2000 random observations and calculate the monitor statistic $W_N^+$'s, from which the run length (RL) is obtained. If $W_N^+ \leq \text{UCL}$, $N = 1, \ldots, 2000$, then suppose $\text{RL}=2000$. This procedure is repeated 20000 times and the values of RL are recorded in each repetition to compute the ARL. In addition, we use the bisection searching algorithms to find the control limit such that $\text{ARL}_0$ is around 370. The control limit of other charts in both scenarios follow the similar way. Henceforth, we use the notation $L$ to denote the control limit coefficient. Please note that the control limits of the WZ charts are similar to Chen et al. (2011), but different from that reported in Bhat and Rao (1972) because, following Chen et al. (2011), we used $\text{ANOS}_1$ as the design criterion rather than the $\alpha$ risk defined in Bhat and Rao (1972). For a relatively fair comparison, we adjust the control limits of different charts to make their $\text{ARL}_0$ or $\text{ANOS}_0$ (IC ANOS) as close as possible, and obtain all results in this section from 20,000 replications.

As mentioned by some researchers (e.g., Borror et al. (1998), Chen et al. (2011), Zhou et
a critical issue is whether it is possible and straightforward to find design parameters that ensure the specified IC performance when the data are discrete. By simulation, we find the WLRT$^+$ chart’s IC ARL can always be attained quite closely if $\theta \leq 0.2$. Figure 1 shows the IC performance of the WLRT$^+$ and CUSUM charts when $\rho_0 = 0.5$. For the illustration purpose, we only present the IC cumulative distribution function (CDF) curves of CUSUM0.6, CUSUM0.99, WLRT0.05$^+$ and WLRT0.2$^+$ charts when $N \leq 100$. Here, the IC run length distribution can be considered satisfactory if it is close to the geometric distribution (Hawkins and Olwell 1998, Zhou et al. (2012)). It is obvious from Figure 1 that the IC run length distribution of our proposed WLRT$^+$ chart is satisfactory compared with the CUSUM chart.

We then evaluate the ARL$_1$ or ANOS$_1$ under the assumption that the process change occurs at the same time as the monitoring starts. Table 1 reports the ARL$_1$ comparison of the WLRT$^+$ and CUSUM charts when $\rho_0 = 0.5$. The exact values of ARL$_0$ are listed in the first row in Table 1, and the corresponding ARL$_1$ for different shifts in the utilization are summarized in the rest of Table 1. To save space, we only list the results when the design parameter of CUSUM $\rho_d = 0.6, 0.7, 0.99$, and the smoothing parameter of WLRT$^+ \theta = 0.05, 0.1, 0.2$, but we also evaluated the case when $\rho_d = 0.51, 0.525, 0.55, 0.75, 0.8, 0.9, 0.95, 1.5$ in practice.

To evaluate overall performance, we also compute the relative mean index (RMI) (Han and Tsung (2006)) values. The RMI can be considered as the average of all relative efficiency values, and a control chart with a smaller RMI value is considered better in its overall
Table 1: Comparisons of ARL when $\rho_0 = 0.5$

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>CUSUM</th>
<th>WLRT$^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho_d = 0.6$</td>
<td>$\rho_d = 0.7$</td>
</tr>
<tr>
<td></td>
<td>$L = 1.387$</td>
<td>$L = 2.090$</td>
</tr>
<tr>
<td>0.50</td>
<td>369(335)</td>
<td>369(343)</td>
</tr>
<tr>
<td>0.51</td>
<td>322(289)</td>
<td>324(300)</td>
</tr>
<tr>
<td>0.55</td>
<td>193(167)</td>
<td>198(183)</td>
</tr>
<tr>
<td>0.58</td>
<td>145(119)</td>
<td>150(132)</td>
</tr>
<tr>
<td>0.63</td>
<td>97.7(73.2)</td>
<td>98.2(81.5)</td>
</tr>
<tr>
<td>0.71</td>
<td>63.0(41.8)</td>
<td>61.5(45.7)</td>
</tr>
<tr>
<td>0.91</td>
<td>32.9(18.7)</td>
<td>30.4(19.4)</td>
</tr>
<tr>
<td>1.70</td>
<td>11.7(5.90)</td>
<td>10.4(5.64)</td>
</tr>
<tr>
<td>3.00</td>
<td>6.10(3.15)</td>
<td>5.47(2.96)</td>
</tr>
<tr>
<td>10.0</td>
<td>2.34(1.29)</td>
<td>2.17(1.21)</td>
</tr>
<tr>
<td>30.0</td>
<td>1.44(0.68)</td>
<td>1.38(0.65)</td>
</tr>
<tr>
<td>RMI</td>
<td>0.200</td>
<td>0.148</td>
</tr>
</tbody>
</table>

1 NOTE: Standard deviations are in parentheses.

performance (Zhou et al. (2012)). From Table 1, we can observe that the performance of the CUSUM chart is satisfactory when the real OC parameter is consistent with the design parameter, but might be deteriorated if the real OC parameter is far from the designated regions. For instance, when $\rho = 0.71$, the ARL of the CUSUM0.7 chart is 61.5, while it is 65.3 for the CUSUM0.99 chart. We can also find that the performance of WLRT$^+$ charts depend on the smoothing parameter, smaller parameter $\theta$ performs better at detecting small shifts, while larger parameter $\theta$ performs better at detecting larger shifts. The cases when $\rho = 0.55$ and $\rho = 3.0$ make this point evidently. Additionally, the WLRT$^+$ charts perform slightly better at detecting large shifts compared with the CUSUM charts. Considering the overall performance, WLRT0.05$^+$ outperforms other competitors.

Considering the effect of the difference of IC run-length distributions, we compare $\gamma_N = Pr_{OC}(RL \leq N) - Pr_{IC}(RL \leq N)$, and show the values for $N \leq 100$ which correspond to early detection in Figure 2. The legend in the first plot is applicable for all the others. The quantity $\gamma_N$ reflects the “true” detection capability of a chart and a control chart with a larger value of $\gamma_N$ is considered better (Zhou et al. (2012)). Figure 2 reveals that WLRT0.05$^+$ stands out at early detection. In other words, giving higher weights to recent data is indeed helpful, which can help detect changes faster.

The ANOS$^1$ comparison of the nL, WZ, CUSUM and WLRT$^+$ charts when $\rho_0 = 0.7$ are given in Table 2. The first row in Table 2 are the exact values of ANOS$^0$. From Table 2, it
Figure 2: The “true” detection capability between CUSUM and WLRT.

Figure 3: The ARL₁ comparison between GLR and MWLE. The legend in the plot (a) is applicable for the plot (b).
can be seen that the performance of CUSUM chart is better than the nL and WZ chart. This result is consistent with the finding by Chen and Zhou (2013). Moreover, the performance of WLRT$^+$0.05 is satisfactory compared with other alternative method.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$n = 5$</th>
<th>$n = 10$</th>
<th>$\delta_u = 4$</th>
<th>$\delta_u = 7$</th>
<th>$\theta = 0.7$</th>
<th>$\theta = 0.98$</th>
<th>$\theta = 0.05$</th>
<th>$\theta = 0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.70</td>
<td>366</td>
<td>370</td>
<td>362</td>
<td>370</td>
<td>370</td>
<td>370</td>
<td>370</td>
<td>370</td>
</tr>
<tr>
<td>0.72</td>
<td>304</td>
<td>307</td>
<td>301</td>
<td>309</td>
<td>293</td>
<td>297</td>
<td>298</td>
<td>307</td>
</tr>
<tr>
<td>0.76</td>
<td>221</td>
<td>234</td>
<td>220</td>
<td>224</td>
<td>197</td>
<td>206</td>
<td>203</td>
<td>218</td>
</tr>
<tr>
<td>0.80</td>
<td>168</td>
<td>170</td>
<td>166</td>
<td>170</td>
<td>149</td>
<td>152</td>
<td>147</td>
<td>164</td>
</tr>
<tr>
<td>0.86</td>
<td>113</td>
<td>116</td>
<td>123</td>
<td>115</td>
<td>103</td>
<td>101</td>
<td>96.1</td>
<td>109</td>
</tr>
<tr>
<td>0.91</td>
<td>87.8</td>
<td>90.4</td>
<td>87.0</td>
<td>89.5</td>
<td>81.9</td>
<td>78.4</td>
<td>73.7</td>
<td>84.0</td>
</tr>
<tr>
<td>0.99</td>
<td>62.1</td>
<td>65.0</td>
<td>61.7</td>
<td>63.8</td>
<td>61.3</td>
<td>55.9</td>
<td>51.3</td>
<td>58.3</td>
</tr>
<tr>
<td>1.31</td>
<td>27.0</td>
<td>30.5</td>
<td>26.8</td>
<td>28.9</td>
<td>30.6</td>
<td>25.3</td>
<td>21.6</td>
<td>22.8</td>
</tr>
<tr>
<td>1.64</td>
<td>17.5</td>
<td>21.1</td>
<td>17.4</td>
<td>19.6</td>
<td>20.3</td>
<td>16.2</td>
<td>13.3</td>
<td>13.5</td>
</tr>
<tr>
<td>1.99</td>
<td>13.4</td>
<td>16.9</td>
<td>13.2</td>
<td>15.6</td>
<td>15.0</td>
<td>11.8</td>
<td>9.51</td>
<td>9.33</td>
</tr>
</tbody>
</table>

Finally, we compare the MWLE chart with the GLR chart and prefer to detect both the upward and downward shift. When $\rho_0 = 0.5$, the control limit of MWLE chart ($\theta = 0.005, 0.01, 0.02$) is chosen as 2.108 such that $\text{ARL}_0 = 370$, the control limit of GLR chart is chosen as 4.863 such that $\text{ARL}_0 = 369$; when $\rho_0 = 0.7$, the control limit of MWLE chart is chosen as 2.113 such that $\text{ARL}_0 = 370$, the control limit of GLR chart is chosen as 5.010 such that $\text{ARL}_0 = 368$. Here, we modify the GLR chart (Chen and Zhou (2013)) with $\tilde{\rho}_k^j = \frac{1}{j-k+2}$ when $\tilde{\rho}_k^j = 0$, where $\tilde{\rho}_k^j$ is the MLE estimator of the utilization given the observations from the $j$th departure to the $k$th departure. From Figure 3, we find there is no evident difference between these two charts in their ability to detect medium downward shifts in the utilization. Furthermore, the MWLE charts perform worse at detecting large downward shift, but perform better at detecting small upward shift.

### 4 Illustrative Example

In this section, we change the M/G/1 make-to-order production plant model proposed by Chen and Zhou (2013) into M/M/1 as an illustrative example. The orders arrive according to a Poisson process with rate $\lambda = 2.0$ per day. Each order needs to be transact in an integrated machine center, and only one order can be processed at one time, the service times are independent and exponentially distributed random variables with rate $\mu = 3.0$, the utilization rate when the operation is normal is therefore $\rho = 0.67$. As mentioned earlier, in
the partial sampling scheme, we only observe queue length after departure epoch, and do not
know the system times. So, we consider the service rate changes from $1/8$ per hour to $1/10.8$
per hour after departure epoch $\tau$, which means $\rho$ changes from 0.67 to 0.9 correspondingly.

Table 3: The CEDs of the control charts

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>5</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>ARL0</th>
<th>UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUSUM</td>
<td>66.2</td>
<td>64.8</td>
<td>60.5</td>
<td>61.4</td>
<td>61.3</td>
<td>369</td>
<td>2.054</td>
</tr>
<tr>
<td>WLRT+</td>
<td>59.0</td>
<td>59.3</td>
<td>59.9</td>
<td>61.7</td>
<td>61.8</td>
<td>370</td>
<td>0.055</td>
</tr>
</tbody>
</table>

1 \text{NOTE: } \rho_0 = 0.67.

We compare the conditional expected delay (CED) (Kenett and Zacks (1998)) due to the
detection ability depends on the time point of the change (Sonesson and Bock (2003)). The
CED is known to be more effective for the evaluation of control schemes such as CUSUM and
EWMA charts (Zacks (2008), Lee and Jun (2012)). For the illustration purpose, only the
CUSUM0.9 and WLRT0.03+ are considered, and any series in which a signal occurs before
the $(\tau +1)$-th observation is discarded. For CUSUM0.9, the real OC parameter is consistent
with the design parameter, which has the best performance when the process change occurs
at the same time as the monitoring starts compared with CUSUM chart with other design
parameters. The CED comparison results of the CUSUM chart and the WLRT+ chart are
given in Table 3. Note a control chart with a smaller CED value is considered performing
better. It is clear that the performance of the WLRT+ chart is satisfactory.

5 Concluding Remarks

This paper proposed a new Phase II control charting scheme based on a weighted likelihood
ratio test (WLRT) for monitoring queue length data in M/M/1 queueing system. The weights
make the control chart like the traditional EWMA scheme, compared with the CUSUM chart
and GLR chart, which can be an alternative method for queueing system.

Although this paper focuses on the partial sampling scheme in the M/M/1 queueing
system, we would like to stress that our approach can be naturally extended to more general
types of queueing system if we know the likelihood function according the queueing theory.
The partial sampling scheme implies we can observe queue length only after departure epoch,
while the complete sampling scheme implies we should observe both queue length and system
times at arrival and departure epochs. Compared with the complete sampling scheme, the
partial sampling scheme is easy and/or inexpensive. We focus on the partial sampling scheme
not only because the observations from queueing systems are only partially available in many
situations, but also we can follow the similar procedure in the complete sampling scheme.
Acknowledgements

The authors would like to thank Dr. Nan Chen of National University of Singapore for discussions. This paper is supported by the National Natural Science Foundation of China Grants 11071128, 11131002, 11001138, 11101306, 11201246 and the RFDP of China Grant 20110031110002.

Appendix A

In this Appendix, we give the proof of the properties in Section 2.1. Apparently, \( \hat{\rho}_N \) can be updated recursively. To begin with, according to the queueing theory, we have the following facts

\[
E(A_n) = E(E(A_n|T_n)) = \rho, \\
Var(A_n) = Var(E(A_n|T_n)) + E(Var(A_n|T_n)) = \rho^2 + \rho,
\]

where \( T_n \) is the service time corresponding to the \( n \) departure. Thus, we obtain the property (2.4) immediately.

In addition, it is not difficult to see that

\[
\frac{\hat{\rho}_N - \rho_0}{\sqrt{(\rho_0 + \rho_0^2) \sum_{n=1}^{N} w_n^2}} = \frac{\sum_{n=1}^{N} w_n (A_n - \rho_0)}{\sigma \sqrt{\sum_{n=1}^{N} w_n^2}},
\]

where \( \sigma^2 = Var(A_n) = \rho_0 + \rho_0^2 \). This, together with

\[
\max_{1 \leq n \leq N} \frac{w_n^2}{\sum_{n=1}^{N} w_n^2} = \frac{\theta^2}{\sum_{n=1}^{N} w_n^2} \to 0,
\]

gives the property (2.5) by the Hajek-Sidak’s Theorem.

At last, by some simple algebra we get

\[
\frac{\partial WLR_T}{\partial \hat{\rho}_N} = 2 \cdot \ln \frac{\hat{\rho}_N (1 + \rho_0)}{\rho_0 (1 + \hat{\rho}_N)} = 2 \cdot \ln \frac{\hat{\rho}_N + \rho_0 \hat{\rho}_N}{\rho_0 + \rho_0 \hat{\rho}_N}.
\]

It is easy to see that \( WLRT_N \) is monotonically increasing (decreasing) on the right (left) side of \( \rho_0 \). This complete the proof (2.6).
Appendix B

In this Appendix, Markov chain methodology similar to that described by Lucas and Saccucci (1990) is used to derive approximate run length properties of the WLRT chart defined by (2.3). From inequality (2.6), we can see that the domain of $\hat{\rho}_N$ is $[a, b]$. Then we can divide the interval into $2m + 1$ subintervals of width $2\delta$, and denote the midpoint of the $j$th subinterval by $S_j$. Thus the one step transition probability can be written as

$$P_{jk} = Pr\{\hat{\rho}_N \text{ in state } k | \hat{\rho}_{N-1} \text{ in state } j\}$$

$$\approx Pr\{S_k - \delta < \hat{\rho}_N \leq S_k + \delta | \hat{\rho}_{N-1} = S_j\}$$

$$= Pr\{S_k - \delta < (1 - \theta)\hat{\rho}_{N-1} + \theta A_N \leq S_k + \delta | \hat{\rho}_{N-1} = S_j\}$$

$$= Pr\{\theta^{-1}[(S_k - \delta) - (1 - \theta)S_j] < A_N \leq \theta^{-1}[(S_k + \delta) - (1 - \theta)S_j]\},$$

where $j = -m, -m + 1, \ldots, m$. This, together with

$$Pr\{A_N = i\} = \frac{1}{1 + \rho} \cdot (\frac{\rho}{1 + \rho})^i, \ i = 0, 1 \ldots$$

we can obtain the transition probability matrix, represented in partitioned matrix form, as

$$P = \begin{pmatrix}
R & (I - R)1 \\
0^T & 1
\end{pmatrix}$$

where the submatrix $R$ is the transition matrix for IC states, $I$ is the identity matrix and $1$ is a column vector of ones. The probabilities of going from an IC state to OC state can be found by subtraction because the rows of the transition probability matrix must add to 1. The zero-state ARL of the WLRT chart can be given by

$$ARL = 1_0(I - R)^{-1}1$$

where $1_0$ is a vector with one in the component corresponding to the initial state and zero in others. The steady-state ARL can be found in a similar manner (see Lucas and Saccucci (1990)).

References


