Received: 25 May 2020 | Revised: 15 July 2020 | Accepted: 6 August 2020

DOI: 10.1002/qre.2745

RESEARCH ARTICLE

WILEY

Control charts for monitoring a Poisson hidden Markov process

Sebastian Ottenstreuer | **Christian H. Weiß** | **Sven Knoth**

Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany

Correspondence
Christian H. Weiß, Helmut Schmidt University, Department of Mathematics and Statistics, PO box 700822, 22008 Hamburg, Germany.
Email: weissc@hsu-hh.de

Abstract

Monitoring stochastic processes with control charts is the main field of application in statistical process control. For a Poisson hidden Markov model (HMM) as the underlying process, we investigate a Shewhart individuals chart, an ordinary Cumulative Sum (CUSUM) chart, and two different types of log-likelihood ratio (log-LR) CUSUM charts. We evaluate and compare the charts' performance by their average run length, computed either by utilizing the Markov chain approach or by simulations. Our performance evaluation includes various out-of-control scenarios as well as different levels of dependence within the HMM. It turns out that the ordinary CUSUM chart shows the best overall performance, whereas the other charts' performance strongly depend on the particular out-of-control scenario and autocorrelation level, respectively. For illustration, we apply the HMM and the considered charts to a data set about weekly sales counts.

KEYWORDS

count time series, CUSUM charts, hidden Markov model, log-likelihood ratio charts, run length performance, statistical process control

1 | INTRODUCTION

Hidden Markov models (HMMs) are commonly used in practice when being concerned with serially dependent processes $(X_t)_{\mathbb{N}_0}$, where $\mathbb{N}_0 = \{0,1,...\}$, especially when the process is discrete-valued, such as a count process or a categorical process, see the literature.[1,2] The basic idea of HMMs is the presence of a sequence $(Q_t)_{\mathbb{N}_0}$ of latent states (out of finitely many ones), which themselves follow a Markov chain (MC) model; the observations X_t are then emitted depending on the current state Q_t. More details on HMMs are presented in Section 2. In the present work, we solely focus on the case of $(X_t)_{\mathbb{N}_0}$ being a count process; so the range of X_t is equal to the set $\mathbb{N}_0$ of nonnegative integers or a subset thereof. But it would also be possible to adapt our methods to HMMs with a different type of range for the observations X_t. More precisely, we assume that the process is stationary if running in control and follows a certain HMM for counts. But it may happen that the process leaves this model after some time (it thus runs out of control), and the aim is to detect such a process change as early as possible.

In order to monitor such stochastic processes, the discipline of statistical process control (SPC) provides tools for detecting changes in the process compared to the given in-control model, namely, control charts. The most basic types of control charts are *Shewhart control charts*, named after Walter A. Shewhart (1891–1967) in view of his pioneering

work in this area. They declare the monitored process to be in control as long as the values of its plotted statistics, for example, the individual observations X_t themselves, lie between two predefined control limits. Otherwise, one concludes that the process might be out of control. Shewhart control charts are very useful to detect assignable causes resulting, for example, in large shifts in the process mean. At the same time, Shewhart charts exhibit a considerable disadvantage: if small sustained shifts are of interest, these charts are known to be very ineffective, because they ignore the information embodied by earlier data (i. e., they lack memory). When more sensitivity to smaller shifts is preferred, *Cumulative Sum (CUSUM) control charts* constitute a better choice, as they accumulate information from previous samples, see Montgomery[3] for details. For ease of presentation, we concentrate on the case of using upper-sided charts for monitoring the considered count process $(X_t)_{\mathbb{N}_0}$, but our approaches presented below are easily adapted to lower-sided or two-sided charts as well, by adapting the idea of Yontay et al.[4]

As mentioned before, HMMs have been widely applied in practice, see Zucchini et al[1] for references, but there seems to be only little work on SPC methods applied to HMMs. Alshraideh and Runger[5] propose to monitor the residuals obtained from the considered HMM by using a Shewhart chart. Fuh[6] derives asymptotic optimality for some CUSUM-type sequential procedures, and Fuh and Mei[7] study approaches related to the generalized likelihood ratio (GLR) scheme for a two-state HMM. Actually, one more frequently finds applications of HMMs in a surveillance context, see the references in Sparks.[8] Such an example is given by Rafei et al,[9] who monitor weekly counts of tuberculosis cases. In contrast to typical SPC applications, where an in-control model (and a corresponding control chart design) is developed for the "common phase" in order to detect any out-of-control situation that violates the in-control assumptions, Rafei et al[9] propose a two-state HMM with two states representing the common phase (in-control) and a particular type of uncommon phase (out-of-control), that is, in-control and out-of-control are part of one model. An analogous definition of the considered HMM can be found in Simões et al,[10] where diesel engines are monitored and five hidden states are defined that correspond to a degradation of the system. Two of these hidden states are interpreted as fault states which require an immediate maintenance action. Similar HMM designs are commonly used in machining and maintenance applications, where the hidden states express increasing tool wear, see, for example, other works.[11-13] In summary, an HMM could be used within control charting in two ways. Either it serves as an explicit change point model or it is the appropriate statistical model for the in-control process. In the first case, however, we assume to know the probability law for reaching the change point while utilizing very simple probability models for the in-control and one or a few more out-of-control phases. In the second case, we make use of the full flexibility of the HMM to describe in-control processes exhibiting particular patterns. Here, we utilize the HMM in its primary sense (or as a parsimonious time series model) for characterizing the in-control process and its potential deviations.

As an illustrative data example, we consider a time series of counts of weekly sales of a soap product,[14] where the hidden states can be interpreted as different (in-control) levels of demand. The monitoring of demand is important, for example, within an inventory management system. The aim is to detect changes compared to the in-control HMM such as shifts within the demand states or distributional changes between the demand states, to be able to act accordingly in presence of changed sale activities. Further details on these data are provided in Section 2, where we also review the definition and important properties of HMMs. Section 3 then presents the control charts for monitoring HMMs considered in this work, namely, the ordinary c and CUSUM chart as well as two different types of log-LR CUSUM charts. The ARL performances of these charts, for diverse out-of-control scenarios, are analyzed in Section 4. Section 5 then picks up the data example from Section 2 and illustrates the application of our control charts in practice. Section 6 discusses the topic of state-dependent control charts, and Section 7 concludes the article.

2 | HIDDEN MARKOV MODELS AND APPLICATIONS

An HMM is defined to be a bivariate process $(X_t, Q_t)_{\mathbb{N}_0}$, where the X_t are the *observable random variables* (counts in our work), and the Q_t are the *hidden states* (latent states) with a finite qualitative range $\mathcal{Q}$, see the literature[1,2] for comprehensive treatments. For simplicity, we label these states by integer numbers, that is, $\mathcal{Q} = \{0, \ldots, d_{\mathcal{Q}}\}$ with some $d_{\mathcal{Q}} \in \mathbb{N}$, where $d_{\mathcal{Q}}$ is typically a small number. The hidden states are assumed to follow a homogeneous MC. Given the state process, the observation process is generated serially independently, with its probability mass function (PMF) depending solely on the current state Q_t.

To put it in a nutshell, a (basic) HMM for $(X_t, Q_t)_{\mathbb{N}_0}$ satisfies the following three conditions: the *observation equation*

$$P(X_t|X_{t-1},...,Q_t,...) = P(X_t|Q_t),\ \forall t \in \mathbb{N}_0; \tag{2.1}$$

the *state equation*

$$P(Q_t|X_{t-1},...,Q_{t-1},...) = P(Q_t|Q_{t-1},...),\ \forall t \in \mathbb{N}_0; \tag{2.2}$$

and the (homogeneous) Markov assumption with state transition probabilities $a_{q|r}$:

$$P(Q_t = q|Q_{t-1} = r,...) = P(Q_t = q|Q_{t-1} = r) = a_{q|r},\ \forall q,r \in \mathcal{Q}. \tag{2.3}$$

In addition, the initial distribution of Q_0 is hereafter assumed to satisfy the invariance equation $\mathbf{A}\boldsymbol{\pi} = \boldsymbol{\pi} = (\pi_0,...,\pi_{d_\mathcal{Q}})^\top$, where $\mathbf{A} = (a_{q|r})_{q,r}$ denotes the transition matrix of the hidden states, and $\boldsymbol{\pi}$ a vector of marginal probabilities. In other words, the MC $(Q_t)_{\mathbb{N}_0}$ given by (2.3), and thus, the whole HMM $(X_t, Q_t)_{\mathbb{N}_0}$, is assumed to be stationary with $P(Q_t = q) = \pi_q$ for all $t \in \mathbb{N}_0$ and all $q \in \mathcal{Q}$. For a vivid description of an HMM's data-generating mechanism, see Alshraideh and Runger.[5, Section 2]

In our performance investigations in Section 4, we ensure this stationarity by defining $(Q_t)_{\mathbb{N}_0}$ as a *first-order discrete autoregressive (DAR(1)) process*,[2] that is, it follows the recursion:

$$Q_t = \alpha_t \cdot Q_{t-1} + (1-\alpha_t) \cdot \epsilon_t, \tag{2.4}$$

where the α_t are independent and identically distributed (i. i. d.) Bernoulli random variables with "success probability" ϕ, and where the innovations ϵ_t are i. i. d. according to $\boldsymbol{\pi}$ with range $\mathcal{Q}$. Both processes, $(\alpha_t)_{\mathbb{N}}$ and $(\epsilon_t)_{\mathbb{N}}$, are assumed to be independent of each other and of $(Q_s)_{s<t}$. This DAR(1) model leads to a parsimoniously parametrized MC model with marginal distribution $\boldsymbol{\pi}$ and transition matrix

$$\mathbf{A} = \begin{pmatrix} \pi_0(1-\phi)+\phi & \pi_0(1-\phi) & \dots & \pi_0(1-\phi) \\ \pi_1(1-\phi) & \pi_1(1-\phi)+\phi & \dots & \pi_1(1-\phi) \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{d_\mathcal{Q}}(1-\phi) & \pi_{d_\mathcal{Q}}(1-\phi) & \dots & \pi_{d_\mathcal{Q}}(1-\phi)+\phi \end{pmatrix}. \tag{2.5}$$

The parameter ϕ controls the extent of serial dependence, with larger values corresponding to stronger dependence.

We restrict ourselves to time-homogeneous state-dependent distributions; therefore (2.1) extends to $P(X_t = x \mid Q_t = q) =: p(x|q)$ for all $t \in \mathbb{N}_0$. In applications, mainly parametric distributions are assumed for $p(\cdot\,|q)$. For illustration, we only consider the *Poisson HMM* in the sequel, which means that the state-dependent distributions of X_t are assumed to be Poisson distributions. So being in state $q \in \mathcal{Q}$ at time t, X_t follows Poi(λ_q) with some $\lambda_q > 0$:

$$P(X_t = x|Q_t = q) = e^{-\lambda_q} \cdot \frac{\lambda_q^x}{x!}. \tag{2.6}$$

Without loss of generality but to simplify the interpretation, we assume that $\lambda_0 < ... < \lambda_{d_\mathcal{Q}}$. From (2.6), we have conditional equidispersion, that is, $E[X_t \mid Q_t = q] = \lambda_q = V[X_t \mid Q_t = q]$, but the unconditional distribution of X_t is overdispersed as a Poisson mixture. It should be noted that any other count model could be used for the HMM as well, also different models for different states. Recently, Adam et al[15] even developed an estimation approach for a nonparametric count HMM. A brief summary of further stochastic properties as well as approaches for the parameter estimation, forecasting, and decoding of the hidden states is provided in Appendix A. More detailed discussions can be found in previous studies.[1,2] These books also provide R codes for doing these computations. In addition, ready software implementations are available for common types of computational software, for example, the "HiddenMarkov" package in R,[16] the command `HiddenMarkovProcess` in Wolfram Mathematica™, or the `hmm...` commands in MATLAB's Statistics and Machine Learning Toolbox™.

At this point, a practitioner might ask how to recognize at all that an HMM could be used for count time series modeling. In our opinion, one should recall the famous words "All models are wrong but some are useful"[17] here. HMMs are very flexible and can be adapted to many types of data and dependence structure. In fact, Zucchini et al[1]

even recommend them as "general-purpose models for time series" (p. 5). In this spirit, Alshraideh and Runger[5] use HMMs as a competitor to the ARIMA models because of their simplicity. Besides their flexibility in capturing quite different time series properties, two further advantages should be stressed: the hidden states are often well interpretable, and a lot of software implementations for HMMs are readily available, also see the previous discussion. So there are several pragmatic reasons for using HMMs in the context of count time series modeling. Nevertheless, any application of an HMM should be complemented by a careful analysis of model adequacy, see, for example, in Weiß.[2, Sections 2.4 and 5.2]

2.1 | A numerical example

To illustrate the HMM's structure as well as common types of statistical inference for HMMs, we first present a simple numerical example (later in Section 2.2, we also discuss a real-data example). The numerical example is motivated by a manufacturing situation with two hidden states (i. e., $d_{\mathcal{Q}} = 1$). For instance, we may count the number of tool changes at a machine within specified time intervals (the HMM's observations, X_t). The two possible hidden states are caused by the incoming raw material (two quality levels, good vs. medium), and the state-dependent Poisson distribution (2.6) has a lower mean for the good-quality batches than for the medium-quality ones. This setup is similar to the coal mill example in Kisić et al,[13] where the wear of the grinding table depends on the quality of the processed coal. Another example for a two-state HMM would be a production environment with two load conditions (normal load vs. overload). The number X_t of produced items in the tth time interval has a larger mean under normal-load operation, because overload operation leads to a throttling of the machine output. Similar to the material example, we cannot directly observe the actual load condition, so this constitutes the HMM's hidden state Q_t.

We assume the following parametrization of the two-state Poisson HMM:

$$\mathbf{A} = \begin{pmatrix} 0.8 & 0.1 \\ 0.2 & 0.9 \end{pmatrix}, \ \pi = \begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix}, \ \lambda = \begin{pmatrix} \lambda_0 \\ \lambda_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}.$$

Note that the transition matrix $\mathbf{A}$ corresponds to a DAR(1) model with $\phi = 0.7$, recall (2.5). The hidden MC features inertia in the sense that it tends to stay in its present state with a probability of 0.8 and 0.9, respectively. The overall probability for being in state "0" is 1/3, and the corresponding state-dependent Poisson distribution has the mean 2 (whereas state "1" goes along with mean 5). Using the formulae given in Appendix A, we compute the marginal mean as 4 and the variance as 6, so the HMM exhibits 50% overdispersion. The autocorrelation function (ACF) takes the values 0.233, 0.163, 0.114,... at lags 1, 2, 3, ...

For illustration, we simulated a time series of length $T = 100$ from this HMM, see the left plot in Figure 1. There, we also show the true values of the hidden states as empty circles, plotted at the value λ_q if the respective state equals q. Then, we fitted a two-state Poisson HMM to the data via maximum likelihood (ML) estimation (see Appendix A for details), leading to the estimates

$$\hat{\mathbf{A}} \approx \begin{pmatrix} 0.868 & 0.086 \\ 0.132 & 0.914 \end{pmatrix}, \ \hat{\pi} \approx \begin{pmatrix} 0.395 \\ 0.605 \end{pmatrix}, \ \hat{\lambda} \approx \begin{pmatrix} 1.97 \\ 4.86 \end{pmatrix}.$$

Because the sample size $T = 100$ is quite small, there are some deviations from the true parameter values. Nevertheless, the fitted model has quite similar properties as the true model, for example, a mean about 3.72, a variance about 5.73 (so about 54 % overdispersion), and the ACF values 0.274, 0.214, 0.167, ...

Next, the fitted model was applied to uncover the hidden states. Here, one distinguishes between *global decoding* of all hidden states, $q_1, \ldots, q_{100}$, and *local decoding*, where only a single hidden state Q_t is determined. The result of a global decoding (using the Viterbi algorithm as described in Appendix A) is shown by grey squares in Figure 1 (plotted as resulting mean $\hat{\lambda}_{\hat{q}_t}$). Nine of the hidden states were decoded incorrectly, most of them between $t = 54, \ldots, 60$.

Finally, the fitted model was used for computing the h-step-ahead forecasting distributions $P(X_{100+h} = x \mid x_1, \ldots, x_{100})$; the ones for the forecast horizons $h = 1,5$ are plotted in the right part of Figure 1. Defining the respective point forecast as the count value maximizing the forecast PMF (mode forecast value), the point forecast for time 101 is 4, and the one for time 105 is 3.

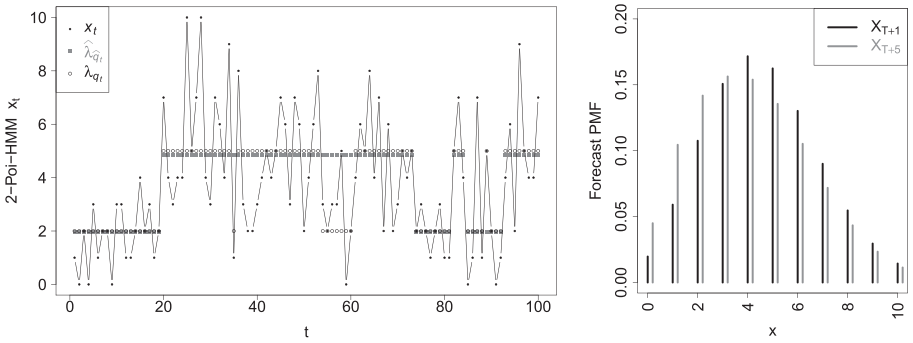

FIGURE 1 Left: Plot of simulated time series $x_1, \ldots, x_{100}$ from 2-state Poisson HMM, together with true hidden states (represented by λ_{q_t}) as well as with globally decoded states (represented by $\hat{\lambda}_{\hat{q}_t}$). Right: Plot of forecast PMFs $P(X_{100+h} = x \mid x_1, \ldots, x_{100})$ with forecast horizons $h = 1,5$

2.2 | Application to sales counts time series

Although manufacturing is probably their most well-known field of application, control charts are also used in many other areas such health surveillance or service industries.[18] As already indicated in Section 1, HMMs are frequently applied in these fields. A health-related example is described by Sebastian et al,[19] where a time series of monthly *Vibrio cholerae* (VC) counts is modeled by a three-state Poisson HMM. There, the three states express "mild," "moderate," and "severe" VC epidemic, and they are related to climate conditions. In what follows, we consider a data example that can be attributed to the field of inventory management. There has been a lot of research activity on the use of control charts for inventory management systems during the last 25 years, see, for example, other works.[20-22] There, control charts are applied, among others, to demand time series to detect changes in the operating environment, which, in turn, would affect the inventory system's performance. Related to such applications, we analyze a count time series regarding the demand for a certain soap product in a supermarket.[14] It contains 242 data points, $x_1, \ldots, x_{242}$, each observation representing the weekly number of sales of the considered soap product. For this time series, it has been established that a stationary three-state Poisson HMM (i. e., $d_{\mathcal{Q}} = 2$) provides an excellent fit, see MacDonald and Zucchini.[14] Therefore, we shall use this data set to further illustrate the interpretation and model fitting of HMMs and to design and apply appropriate control charts for process monitoring, where the latter is done in Section 5.

The nine model parameters (transition matrix $\mathbf{A}$ and state-dependent Poisson parameters λ_q, $q = 0,1,2$) are again estimated by ML estimation, that is, by a direct numerical maximization of the likelihood function

$$\begin{aligned} L(\theta) = &\ P(X_{242} = x_{242}, \ldots, X_1 = x_1 \mid \theta) \\ = &\ \mathbf{1}^\top \mathbf{P}(x_{242}) \mathbf{A} \mathbf{P}(x_{241}) \mathbf{A} \ldots \mathbf{P}(x_1) \pi, \end{aligned} \tag{2.7}$$

with $\boldsymbol{\theta}$ and $\mathbf{P}(x)$ denoting the parameter vector and the diagonal matrix $\mathbf{P}(x) := \mathrm{diag}(p(x|0), \ldots, p(x|d_{\mathcal{Q}})) \in [0;1]^{(d_{\mathcal{Q}}+1)\times(d_{\mathcal{Q}}+1)}$, respectively (also see Appendix A). The corresponding implementation in R can be found in the supplemental materials (together with all other codes used in this article). The state-dependent means of the fitted Poisson HMM (see (2.6)) are $\hat{\lambda} = (\hat{\lambda}_0, \hat{\lambda}_1, \hat{\lambda}_2) \approx (3.74, 8.44, 14.93)$, which – from a practical point of view – can be interpreted as a low (state 0), medium (state 1) and high (state 2) level of demand for the particular soap product. Together with the estimates for the states' PMF, $\hat{\pi} = (\hat{\pi}_0, \hat{\pi}_1, \hat{\pi}_2)^\top \approx (0.722, 0.220, 0.058)^\top$, they yield a model mean of 5.42 and a model variance of 14.72, both being very close to the corresponding sample values 5.44 and 15.40, respectively. As before, the fitted model was also used for global decoding of the complete series of hidden states, $q_1, \ldots, q_{242}$, by using the Viterbi algorithm (see Appendix A).

Figure 2 shows the data and the sequence of states (left-hand side) as well as the sample and model ACF against increasing time lags (right-hand side). Clearly, the sample ACF indicates a significant degree of positive serial dependence, which is almost exactly captured for the first four lags by the fitted HMM. As the time series in Figure 2 nicely shows, state 0 is predominant with an overall probability of 72.2% for being in this state, whereas the occurrence of state 2 seems to be most unlikely (5.8%). Also when looking at the estimates $\hat{a}_{3|1}, \hat{a}_{3|2}$ in

$$\hat{\mathbf{A}} \approx \begin{pmatrix} 0.864 & 0.445 & 0.000 \\ 0.117 & 0.538 & 0.298 \\ 0.019 & 0.017 & 0.702 \end{pmatrix},$$

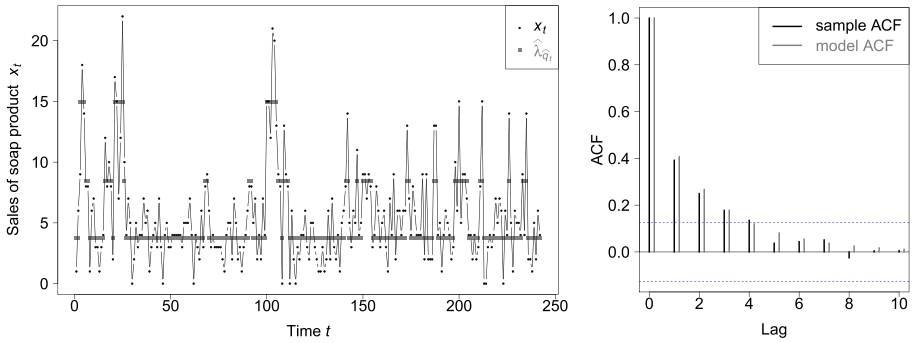

FIGURE 2 Left: Plot of sales of soap product together with globally decoded states (fitted model means $\hat{\lambda} = (3.74, 8.44, 14.93)$ are displayed in gray if the respective decoded state equals q). Right: Sample and model ACF [Colour figure can be viewed at wileyonlinelibrary.com]

one can conclude that a period of high demand seems to start very rarely. However, once the demand for soap reaches state 2, it tends to remain at this level instead of changing directly. Furthermore, it is interesting to note that the fitted model nearly excludes the transition from state 2 to state 0 (the displayed values are roundings), which means that we would not expect a decline in demand to be that radical.

3 | MONITORING HMMS WITH C CHART, CUSUM CHART, AND LOG-LR CUSUM CHART

To detect changes in, for example, the sales activity as early as possible, we shall from now on concentrate on control chart approaches for count processes following an HMM. Because in most applications, the detection of increases in the counts' mean is particularly relevant (e. g., to request supply of sales products, or to detect an outbreak of an epidemic disease), we focus on upper-sided control charts; so the investigated charts require only one (that is, an upper) control limit. But it would easily be possible to adapt our monitoring concepts to the lower-sided or two-sided case. As the monitoring approaches, we consider the c chart as a common benchmark chart on the one hand, and different types of CUSUM charts on the other hand, providing an improved sensitivity towards small process changes thanks to an inherent memory.

The c chart is a basic Shewhart-type control chart,[3] which simply plots the counts X_t as they arrive in time. It triggers an alarm if its specified control limit $u > 0$ is violated, that is, if $X_t > u$. Because the monitored data are autocorrelated, we cannot use simple 3-σ or quantile limits but have to adapt the control limits according to the actual serial dependence structure (so a "modified Shewhart chart" in the sense of Schmid[23]). Like for the remaining charts, this is done based on average run length (ARL) considerations. The c chart is often well suited for detecting large shifts, and it also serves as a benchmark for performance analysis in this work, analogously to, for example, the literature.[24,25]

The statistics of the (standard) upper-sided CUSUM chart are calculated as

$$C_0 = c_0, \quad C_t = \max\{0, X_t - k + C_{t-1}\} \text{ for } t = 1, 2, \ldots \tag{3.1}$$

with $k > 0$ and $c_0 \geq 0$ as the reference and starting value, respectively. The parameter k is commonly chosen larger than but close to the in-control mean μ_0. A value $c_0 > 0$ is referred to as a Fast Initial Response (FIR) feature. (Generally, employing a FIR feature may help to detect an initial out-of-control situation more quickly; at the same time, it may also cause more false alarms when the process is initially in-control. Fixing this leads to slightly delayed detection of late changes.) If C_t exceeds the CUSUM chart's control limit $h > 0$, an alarm indicates that the process might be out of control. In view of an exact performance evaluation, we choose rational or even integer design parameters k, h, c_0.

The ordinary CUSUM recursion (3.1) can be motivated by the log-likelihood ratio test applied to i. i. d. Poisson counts. If explicitly considering the log-likelihood ratio for HMMs, a more complex type of CUSUM chart is obtained, the *log-LR CUSUM control chart*. To prevent a possible numerical underflow in later computations, the following recursive scheme (similar to the one recommended by Zucchini et al[1]) is utilized for calculating the likelihood function given the count time series $x_1, \ldots, x_T$ (also see Appendix A):

$$\begin{aligned} &\boldsymbol{v}_1 := \mathbf{P}(x_1)\pi,\ w_1 := \mathbf{1}^\top \boldsymbol{v}_1; \\ &\boldsymbol{v}_t := \mathbf{P}(x_t)\mathbf{A}\boldsymbol{v}_{t-1}/w_{t-1},\ w_t := \mathbf{1}^\top \boldsymbol{v}_t \text{ for } t = 2, \ldots, T. \end{aligned} \tag{3.2}$$

The log-likelihood function then equals

$$\ell(\theta) := \ln(w_T \ldots w_1) = \sum_{t=1}^{T} \ln(w_t).$$

The log-likelihood ratio requires another parameter vector which targets a certain out-of-control scenario, say $\boldsymbol{\theta}_1$, whereas $\boldsymbol{\theta}_0$ denotes the in-control parameter vector. The log-likelihood ratio reads as

$$\ell R(\theta_0, \theta_1) := \ln LR(\theta_0, \theta_1) := \ln\left(\frac{L(\theta_1)}{L(\theta_0)}\right) = \sum_{t=1}^{T} \underbrace{(\ln(w_{t,1}) - \ln(w_{t,0}))}_{=:\ell R_t}, \tag{3.3}$$

where "0" and "1" in the index of $w_{t,\cdot}$ refer to the in-control and anticipated out-of-control parameters, respectively, used for the computations in (3.2). Note that the HMM allows for different $\boldsymbol{\theta}$ yielding the same process mean μ. Therefore, in Section 4, we investigate two log-LR CUSUM charts with different $\boldsymbol{\theta}_1$: one, where only the state-dependent means λ_q change compared to $\boldsymbol{\theta}_0$ (thus $\mathbf{P}(x)$ in (3.2)); and another one, where the PMF $\boldsymbol{\pi}$ of the hidden states Q_t changes (thus the transition probabilities $\mathbf{A}$ in (3.2)). We refer to these specific charts as the *log-LR$_\lambda$* and *log-LR$_{\boldsymbol{\pi}}$ CUSUM chart*, respectively. However, both parameter vectors $\boldsymbol{\theta}_1$ result in the same out-of-control mean μ_1; further explanations are provided in Section 4. The log-LR CUSUM chart is now defined as

$$\ell C_0 = 0,\ \ell C_t = \max\{0, \ell R_t + \ell C_{t-1}\} \text{ for } t = 1, 2, \ldots \tag{3.4}$$

with $\ell R_t = \ln(w_{t,1}) - \ln(w_{t,0})$. Again, if $\ell C_t > \ell h > 0$ with ℓh as the control limit, this indicates that the process might be out of control. It is worth mentioning that Fuh and Mei[7] also consider a log-LR CUSUM statistic but applied to a two-state HMM with normal state-dependent distributions (thus variables data).

4 | PERFORMANCE EVALUATION OF C CHART, CUSUM CHART, AND LOG-LR CUSUM CHART

We assess the effectiveness of all investigated charts by their ARL. In the field of SPC, the ARL is a common performance measure, which indicates the expected number of points plotted on a control chart before a possible out-of-control condition is signaled. Thus, a control chart's ARL should be as large as possible while the monitored process is in control, whereas the out-of-control ARL should be preferably small. In the following, we restrict our investigations to the most popular ARL concept, the *zero-state* ARL, for evaluating both the in-control and out-of-control performance. (Although there exist a few more ARL concepts, see Knoth,[26] the zero-state ARL is particularly meaningful in the CUSUM case as it expresses some kind of worst-case behavior.) In order to compute the ARL values of the c and the CUSUM chart, the MC approach of Brook and Evans[27] is utilized. This is possible, because $(X_t, Q_t)_{\mathbb{N}_0}$ and $(X_t, Q_t, C_t)_{\mathbb{N}_0}$ constitute a bivariate and trivariate MC, respectively, see Appendix B for further details. Because the processes $(X_t)_{\mathbb{N}_0}$ and $(C_t)_{\mathbb{N}}$ only take discrete values, applying the MC approach yields exact numerical results. ARL values of the log-LR charts are obtained by simulations (with 10^6 replications per ARL value); so the results for the log-LR charts are only approximations, nevertheless sufficiently accurate ones.

The charts' design parameters are chosen such that the in-control ARL (hereafter signified by adding the subscript "0") is between 200 and 300. In order to be able to compare the investigated charts, the differences between their ARL_0 values should be as small as possible. Because the c chart has only one parameter affecting its ARL, that is, its upper control limit $u \in \mathbb{N}$, the chart is most inflexible in terms of adjusting the ARL_0 to a given target value. More flexibility is shown by the CUSUM chart with two design parameters (neglecting a FIR feature, i. e., setting $c_0 := 0$), which — in contrast to the c chart's control limit u — can even take noninteger values for a further fine-tuning of the ARL performance. For practical reasons concerning software implementations (see the corresponding R-code in the supplemental materials) and without loss of generality, the CUSUM parameters are chosen to be rational, that is, $k, h \in \mathbb{Q}^+$, and to have the same denominator. So as the first step of our performance study, an appropriate value for u is chosen. Then the CUSUM chart's parameters are determined with respect to the c chart's ARL_0, and finally, the control limits of both log-LR charts are selected such that their ARL_0 values locate between the ones of the c and the CUSUM chart. Usually,

one can find more than just one relevant tuple (k, h) as CUSUM chart design having nearly the same ARL_0; the same holds for both log-LR charts. For designing the CUSUM chart, we first have to set the reference value k, which we decided to be close to the in-control mean μ_0 (analogous to the recommendation of Weiß,[24] who investigated CUSUM charts for an alternative integer-valued model with autocorrelation, namely, first-order integer-valued autoregressive (INAR(1)) processes). Subsequently, an appropriate value for h was determined. For the log-LR charts, a certain out-of-control mean μ_1 must be predefined, then a corresponding control limit ℓh is identified. Now, for comparison purposes, the log-LR charts should aim for the detection of smaller (positive) shifts as well, that is, μ_1 should be reasonably close to μ_0 (one may use the formula $k = (\mu_1 - \mu_0)/(\ln\mu_1 - \ln\mu_0)$ in Lucas,[28] which was derived for i. i. d. Poisson counts, as a rough guide for adjusting the different CUSUM concepts).

In the following, sustained (positive) shifts in the marginal mean μ of the observable count process are considered. These shifts are obtained in three different ways:

- all state-dependent means λ_q are multiplied by the same factor simultaneously (see the two upper graphs in Figure 3);
- only one of the state-dependent means λ_q is multiplied by a certain factor (see Figure 4);
- the probability mass within the states' PMF $\boldsymbol{\pi}$ is shifted towards the state with the largest mean (i. e., towards state $d_{\mathcal{Q}}$), whereas the remaining probabilities are reduced by the same relative amount (see the lower graphs in Figure 3).

All scenarios are adjusted to give the same increase $\delta = \mu - \mu_0$ in the marginal mean μ (in the last scenario, μ is bounded by the value $\lambda_{d_{\mathcal{Q}}}$ of the largest state-dependent mean).

Our presented results specifically refer to an in-control three-state Poisson HMM with $\boldsymbol{\lambda}=(\lambda_0, \lambda_1, \lambda_2) = (1, 2, 5)$ and $\boldsymbol{\pi}=(\pi_0, \pi_1, \pi_2)^\top = (0.5, 0.35, 0.15)^\top$, leading to $\mu_0 = 1.95$. To easily control the dependence level while keeping $\boldsymbol{\pi}$ fixed, we use the DAR(1) model from Section 2, with the parameter ϕ taking the three different values 0.2, 0.5 and 0.8. But neither the charts nor the schemes for ARL computation are limited to such DAR(1) hidden states, which is illustrated later in Section 5. The respective parameter values of the investigated charts, corresponding to the ARL performances in Figures 3–5, are displayed in Table 1. Here, the target value $\mu_1 = 3.0225$ for the log-$LR_{\boldsymbol{\lambda}}$ chart results from multiplying $\boldsymbol{\lambda}$ with 1.55, and the one for the log-$LR_{\boldsymbol{\pi}}$ chart from the "shifted" marginal distribution $\boldsymbol{\pi}_1 = (0.324, 0.227, 0.449)^\top$. Note that the above shift scenarios used for evaluating the charts' performance include both of these anticipated out-of-control situations. For the c chart design, we always ended up with the same value $u = 9$ for the control limit because of discreteness. As a consequence, the ARL_0 values in Table 1 slightly increase with increasing dependence parameter ϕ, a

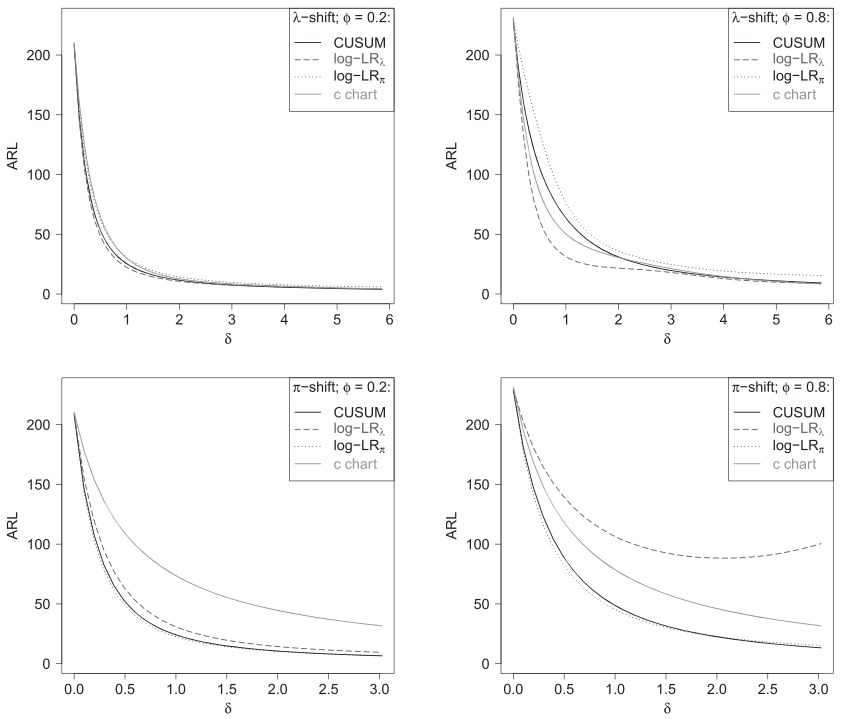

FIGURE 3 ARL performances against mean shifts $\delta = \mu - \mu_0$, $\mu_0 = 1.95$, for two different degrees of dependence: $\phi = 0.2$ (left side) and $\phi = 0.8$ (right side). Mean shift δ caused by uniform shift in $\boldsymbol{\lambda}$ in upper panel, and by change in $\boldsymbol{\pi}$ in lower panel

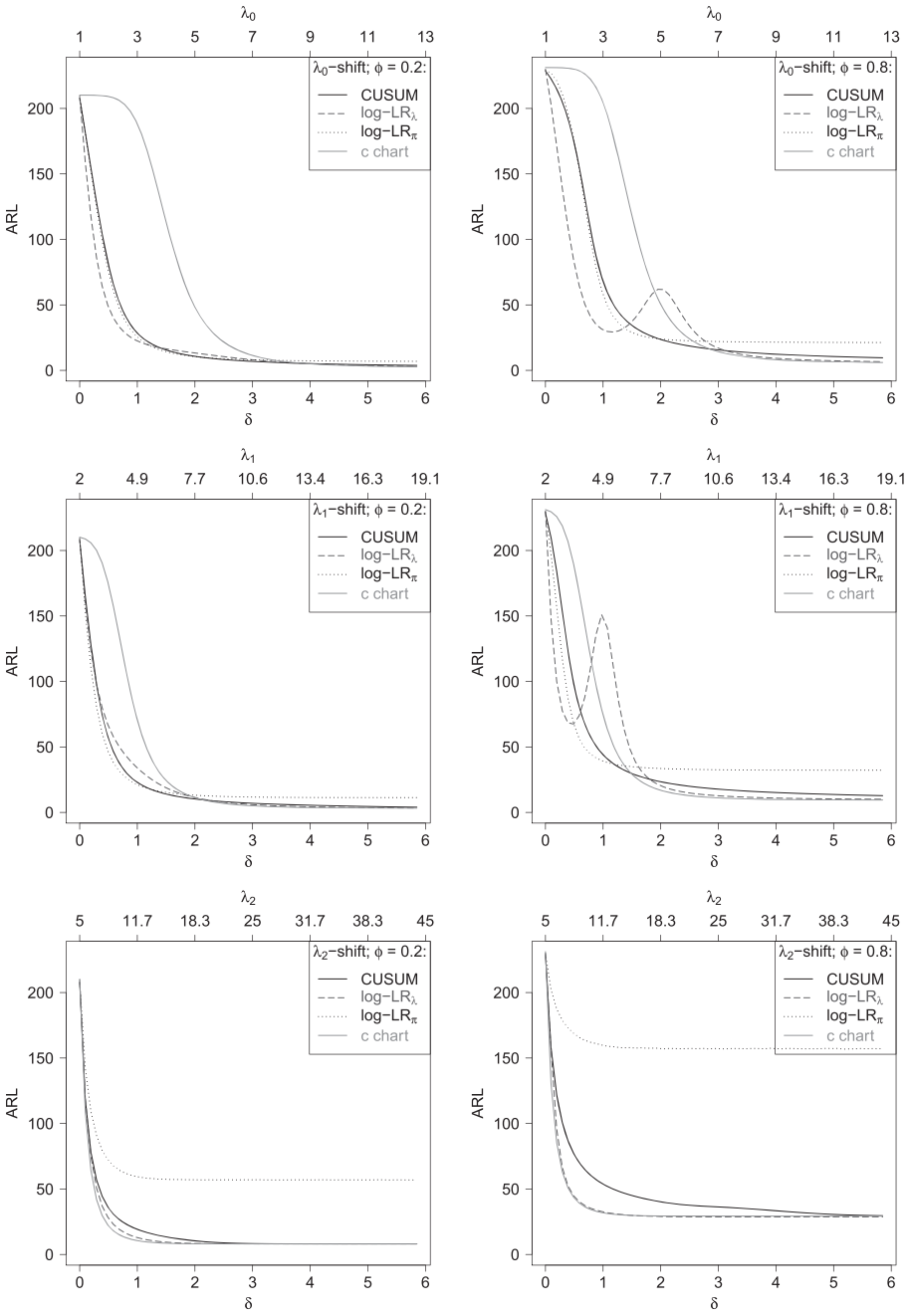

FIGURE 4 ARL performances against mean shifts $\delta = \mu - \mu_0$, $\mu_0 = 1.95$, for two different degrees of dependence: $\phi = 0.2$ (left side) and $\phi = 0.8$ (right side). Mean shift δ caused by shift in λ_0 (upper panel), by shift in λ_1 (middle panel) and by shift in λ_2 (lower panel)

well-known phenomenon for Shewhart-type charts.[23,25] To ensure a fair comparison, also, the different CUSUM charts have been designed to (roughly) meet these c charts' ARL_0 values.

The two upper graphs in Figure 3 show the ARL curves where all state-dependent means are shifted uniformly. For the low dependence level 0.2 (left part), all charts show almost the same performance. At least for smaller shifts, say $\delta \leq 2$, there are some minor differences in the charts' performances, such as the log-LR_{λ} chart tends to exhibit the best performance. This becomes clearer when ϕ rises up to 0.8 (right part), now the log-LR_{λ} chart performs clearly best concerning smaller shifts. Actually, this superior performance is not surprising as the log-LR_{λ} chart is specifically designed for such a uniform shift in $\boldsymbol{\lambda}$. Moreover, it can be concluded that the log-$LR_{\boldsymbol{\pi}}$ chart is least suited for this scenario as its ARL curve locates above the other ones throughout all shift sizes. This can be explained by the fact that the true out-of-control situation strongly differs from the anticipated one.

The opposite situation is considered in the two lower graphs in Figure 3, where the mean shifts are caused by changes within the states' PMF $\boldsymbol{\pi}$. By an analogous reasoning as before, it is not surprising that the log-$LR_{\boldsymbol{\pi}}$ chart performs very well for both dependence levels. However, it is interesting to observe that the ordinary CUSUM chart has

TABLE 1 Chart designs and ARL_0 values used for the ARL performance analyses in Figures 3–5

	Control Limit (ARL_0), $\phi =$					
Control Chart	**0.2**		**0.5**		**0.8**	
c chart	9	(210.15)	9	(214.37)	9	(231.22)
CUSUM ($k = 2.5$)	14	(207.97)	19	(217.33)	30.5	(228.66)
log-LR_λ ($\mu_1 = 3.0225$)	2.465	(208.71)	2.295	(216.63)	2.25	(229.92)
log-LR_π ($\mu_1 = 3.0225$)	2.73	(209.43)	2.5075	(216.37)	2.025	(229.51)

nearly the same sensitivity towards small shifts, and even a slightly better sensitivity towards medium-sized shifts for $\phi = 0.8$ (note that the largest possible shift size in this scenario is $\delta = \lambda_2 - \mu_0 = 3.05$). In contrast, the c chart exhibits a rather bad ARL performance, irrespective of ϕ. The log-LR_λ chart is even worse for $\phi = 0.8$ (even a nonmonotonic behavior in δ). So together with our analyses of the upper panel in Figure 3, it becomes clear already at this point that the log-LR charts are very sensitive regarding deviations from the anticipated out-of-control scenario.

Figure 4 shows the remaining shift scenarios of our investigations (again left column $\phi = 0.2$ and right column $\phi = 0.8$), that is, only one of the three state-dependent means λ_q is shifted such that the increase in the (overall) mean μ is still the same as in the two preceding scenarios. Starting with shifts of the lowest state-dependent mean λ_0, we can observe that the c chart performs very poorly concerning smaller shifts for both ϕ values. Keeping in mind the c chart's simple structure (i. e., the observed counts X_t are plotted against the upper limit u), it seems plausible that its ARL does not significantly decline in this case: If the hidden state equals 1 or 2, counts are emitted with the same state-dependent mean as before, and if the hidden state equals 0, the corresponding mean λ_0 is somewhat larger but still far away from the control limit u. For large shifts, however, λ_0 exceeds u and the c chart consequently shows its well-known advantage over other control charts. A similar pattern appears in case of λ_1-shifts, though with an improved performance for small shifts, since already the in-control value of λ_1 is closer to u than λ_0. Finally, for the λ_2-shifts (with the in-control value of λ_2 being closest to u), the c chart has even one of the best performances throughout all shift sizes. This time, its simple structure has a positive effect on the ARL, even for smaller shifts, because λ_2 moves even closer to u, and exceeds the control limit much earlier (in terms of δ-shifts) than λ_0 and λ_1 do in the previous shift scenarios.

The most striking performances in Figure 4 are caused by the log-LR_λ chart. While the chart performs pretty well for $\phi = 0.2$, its ARL curves become nonmonotonic for $\phi = 0.8$, note the (local) maxima in both curves for λ_0- and for λ_1-shifts. To understand this behavior, it is important to know that these local maxima are attained if λ_0 and λ_1, respectively, are approximately equal to $\lambda_2 = 5$. In the boundary cases $\lambda_0 = \lambda_2$ and $\lambda_1 = \lambda_2$, respectively, the three-state HMM actually reduces to a two-state HMM with increased probability mass in the uppermost state – so these λ_q-shifts are equivalent to the pure $\boldsymbol{\pi}$-changes towards $(0, 0.35, 0.65)$ and $(0.5, 0, 0.5)$, respectively. (Although we omitted to print these numbers in the manuscript, it is worth mentioning that for $\phi = 0.9$, a further local maximum appears in the ARL curve of the λ_0-shifts, located where λ_0 approximately reaches the value of λ_1.) Generally, the out-of-control states are not uniquely characterized in the sense that, for example, the sole shift $\lambda_0 \mapsto 3$ cannot be distinguished from the joint shift $\lambda_0 \mapsto 2$, $\lambda_1 \mapsto 3$, $\pi_0 \mapsto 0.35$, $\pi_1 \mapsto 0.5$. So corroborating the results in the lower panel of Figure 3, the log-LR_λ chart seems to be severely affected by deviations from its anticipated out-of-control scenario (though becoming apparent only if the dependence level is sufficiently large). For the λ_2-shifts, in contrast, we have a strictly monotone ARL curve as λ_2 is the largest among the state-dependent means, that is, the ordering among the means does not change by such shifts. To sum up, the log-LR_λ chart exhibits serious weaknesses when a shifted λ_q approaches another state-dependent mean λ_r, $r \neq q$, which is more pronounced for larger ϕ.

Also the log-LR_π chart shows a clear pattern: the ARL converges to some limiting level for increasing shifts, which is lowest in the upper panel, somewhat larger in the middle panel and largest in the lower panel of Figure 4. Again, this behavior is considerably more pronounced for $\phi = 0.8$. Our explanation is as follows: In the upper panel, once λ_0 has exceeded λ_2, the states' PMF $\boldsymbol{\pi}$ (corresponding to the ordered state-dependent means λ_q) is $(0.35, 0.15, 0.5)$, which is pretty close to the anticipated out-of-control $\boldsymbol{\pi}$ given by $(0.324, 0.227, 0.449)$. Analogously, in the middle panel, we end up with an ordered $\boldsymbol{\pi}$ of $(0.5, 0.15, 0.35)$ again close to $(0.324, 0.227, 0.449)$. For the λ_2-shifts in the lower panel, in contrast, the order within $\boldsymbol{\pi}$ remains $(0.5, 0.35, 0.15)$ throughout all shift sizes and thus exhibits the least fit with $(0.324, 0.227, 0.449)$.

Let us complete the analysis of the single λ_q-shifts by briefly discussing the ordinary CUSUM chart's performance. Obviously, the CUSUM chart yields an almost equally good performance in all panels of Figure 4 (this impression is

also confirmed for $\phi = 0.5$ in Figure 5 below), that is, the CUSUM chart's ARL performance exhibits the best overall robustness against both the actual autocorrelation level and the specific shift scenario.

Figure 5 displays the resulting performances for the moderate dependence level $\phi = 0.5$, now sorted by the control charts instead of the different shift scenarios. As already indicated above, the CUSUM chart shows the most homogeneous performance regarding the different types of shifts across all investigated charts. Any other chart turns out to be very sensitive to the actual shift type: The c chart has strongly varying ARL values for small shifts, the log-LR$_{\boldsymbol{\pi}}$ chart performs very badly for λ_2-shifts, and the log-LR$_{\boldsymbol{\lambda}}$ chart has nonmonotonic ARL curves for the λ_0-, λ_1- and $\boldsymbol{\pi}$-shifts. So if we cannot be sure about the expected type of out-of-control scenario, only the ordinary CUSUM chart appears to be a reliable choice.

5 | APPLICATION TO SALES COUNTS DATA

To illustrate the application of the c and CUSUM chart as well as the log-LR approach, we continue the example about sales counts as introduced in Section 2. There, it was argued that a stationary three-state Poisson HMM with transition matrix

$$\mathbf{A}_0 = \begin{pmatrix} 0.864 & 0.445 & 0.000 \\ 0.117 & 0.538 & 0.298 \\ 0.019 & 0.017 & 0.702 \end{pmatrix} \text{ and thus PMF } \boldsymbol{\pi}_0 = \begin{pmatrix} 0.722 \\ 0.220 \\ 0.058 \end{pmatrix},$$

and state-dependent means $\boldsymbol{\lambda}_0 = (3.74, 8.44, 14.93)$, constitutes an adequate fit to the time series of weekly sales counts (Phase I analysis). We interpreted the three hidden states as demand states for the considered soap product.

This model is considered as the in-control model. For Phase II monitoring, we simulate a time series $(\tilde{x}_1, \tilde{q}_1), \ldots, (\tilde{x}_{100}, \tilde{q}_{100})$ of length 100, which is initialized by the last hidden state of our Phase I data according to the global decoding shown in Figure 2, that is, by $\tilde{q}_0 = 0$. The first 20 observations are simulated according to the in-control model. But then, there is a change point (marked by the vertical dotted lines in Figure 6), at which the two lower

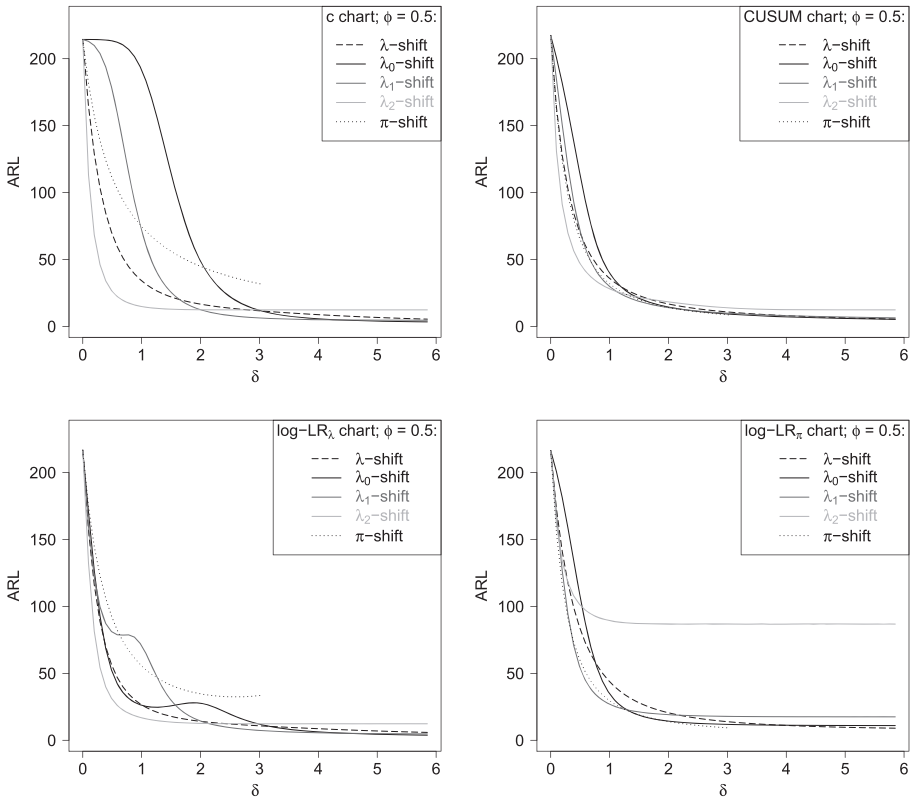

FIGURE 5 ARL performances against mean shifts $\delta = \mu - \mu_0$, $\mu_0 = 1.95$, of c, CUSUM, log-LR$_{\boldsymbol{\lambda}}$ and log-LR$_{\boldsymbol{\pi}}$ chart for $\phi = 0.5$

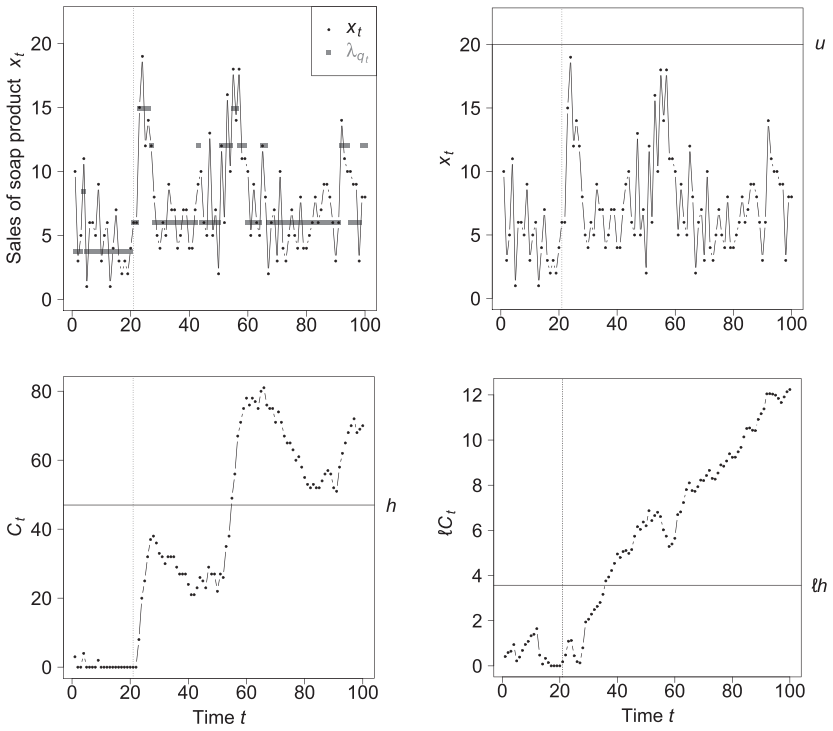

FIGURE 6 Top left: Plot of simulated Phase II data with state-dependent means (displayed in gray if the respective state equals q) $\boldsymbol{\lambda}_0 = (3.74, 8.44, 14.93)$ for $t \leq 20$, and $\boldsymbol{\lambda}_1 = (6,12, 14.93)$ for $t \geq 21$. Top right: c chart ($u = 20$). Bottom left: CUSUM chart $(k,h) = (7,47)$. Bottom right: log-LR$_{\boldsymbol{\lambda}}$ chart $(\mu_1, \ell h) = (8.4057, 3.57)$

state-dependent means increase from $\boldsymbol{\lambda}_0 = (3.74, 8.44, 14.93)$ to $\boldsymbol{\lambda}_1 = (6, 12, 14.93)$, whereas all other parameters remain the same. Consequently, the model mean (variance) rises from 5.42 (14.72) to 7.84 (17.01). The remaining 80 observations are simulated according to this out-of-control model. The resulting Phase II data, that is, 100 weekly sales counts together with the true hidden states (or their means, respectively), are plotted in the first graph of Figure 6. Note that there are three possible hidden states before and after the dotted line, but their actual mean levels (the grey squares) change from $\boldsymbol{\lambda}_0$ to $\boldsymbol{\lambda}_1$. In terms of a practical meaning of this shift, the out-of-control model assumes the same mean number of sales in the high-demand state, whereas in weeks with a low or medium demand, the supermarket is confronted with an increased sales activity.

As the next step, we have to select and design the control charts to be used for Phase II monitoring. As before, we consider the c, the CUSUM and the log-LR$_{\boldsymbol{\lambda}}$ chart for process monitoring, but the log-LR$_{\boldsymbol{\pi}}$ chart shall not be applied in this section. The reason for this is as follows: For our performance analyses in Section 4, we assumed a DAR(1) structure for the hidden states' MC, because this model allowed us to separate the marginal PMF $\boldsymbol{\pi}$ from the extent of serial dependence as expressed by the parameter ϕ. The hidden MC of our sales counts data, however, turned out to be more complex than just DAR(1), and the relation between $\boldsymbol{\pi}$ and $\mathbf{A}$ is not obvious. Thus, it would be difficult to change $(\mathbf{A}, \boldsymbol{\pi})$ for receiving an anticipated μ_1. So we concentrate on c, CUSUM and log-LR$_{\boldsymbol{\lambda}}$ chart this time, and the charts' design parameters are chosen as

- $u = 20$ for the c chart, yielding (the in-control) ARL$_0$ 245.35,
- $(k, h) = (7, 47)$ for the CUSUM chart, yielding ARL$_0$ 244.37, and
- $(\mu_1, \ell h) = (8.4057, 3.57)$ for the log-LR$_{\boldsymbol{\lambda}}$ chart, yielding ARL$_0$ 245.15.

These charts are now applied to the Phase II data, see Figure 6 (the displayed results can be replicated by applying the corresponding R code in the supplemental materials). After the shift, the ARLs reduce to 152.38 (c chart), 53.37 (CUSUM chart) and 23.84 (log-LR$_{\boldsymbol{\lambda}}$ chart). In line with these values, the log-LR$_{\boldsymbol{\lambda}}$ chart is the first chart triggering an alarm for the considered simulation, that is at $t = 36$ (delay 15). It is followed by the CUSUM chart, whose control limit is first exceeded at $t = 55$ (delay 34). Before the change point though, the statistics of both charts are always almost 0 such that no false alarms are triggered, see the corresponding graphs in Figure 6. In contrast, the c chart does not signal any alarm within the considered time interval.

The bad performance of the c chart is mainly due to the shift-type that does not affect the value of λ_2. We thus have a situation similar to what has been previously discussed in Section 4 regarding the two upper panels of Figure 4, where

the value of λ_2 remains unchanged likewise. In the present case, the chosen shift-type only concerns those state-dependent means, which are farthest below the control limit u and which additionally even remain below the value of λ_2 after the shift. The deficiency of the c chart is also confirmed by the large out-of-control ARL of 152.38. At the same time, it seems reasonable that the log-LR$_\lambda$ chart outperforms both other charts, as its anticipated out-of-control scenario, $1.55 \cdot \boldsymbol{\lambda}_0 = (5.797, 13.082, 23.1415)$, is quite close to $\boldsymbol{\lambda}_1$. But also the simple CUSUM chart does quite well for these data, which confirms our conclusion in Section 4 that the CUSUM appears to be a good compromise: It is well suited for detecting mean shifts independent of the actual shift structure.

6 | ON STATE-DEPENDENT CONTROL CHARTS

The control charts discussed so far consider the HMM for chart design, but the HMM's structure does not become visible to the operator during process monitoring. However, one may also think about a state-dependent control chart, where the action taken at time t tries to make use of the current state q_t. For example, one might define a state-dependent c chart, where still the observed counts x_t are plotted on the chart, but the control limit u_{q_t} varies according to the underlying state q_t. A related approach was proposed by Alshraideh and Runger,[5] who define a residuals-based Shewhart chart with 3-σ limits. Here, the residual r_t at time t considers the previous state q_{t-1} in the following way: $r_t := x_t - \sum_{j=0}^{d_{\mathcal{Q}}} \lambda_j \cdot a_{j|q_{t-1}}$. In both cases, however, the true hidden states q_t cannot be observed, so we have to rely on a (locally) decoded state $\hat{q}_t$. This is possible by using (3.2): the most probable state at time t, given the observed counts $x_1, \ldots, x_t$, equals $\hat{q}_t = \text{argmax}_q v_{t,q}$, see Appendix A. So the state-dependent c chart actually plots x_t against $u_{\hat{q}_t}$, and the residuals chart plots $\hat{r}_t := x_t - \sum_{j=0}^{d_{\mathcal{Q}}} \lambda_j \cdot a_{j|\hat{q}_{t-1}}$.

Because we do not know the true data-generating HMM, the local decoding has to be based on the supposed in-control model, whereas the generated data might be out-of-control. This, however, may lead to systematic misidentifications of the hidden states, that is, we permanently have $\hat{q}_t \neq q_t$, which, in turn, deteriorates the charts' out-of-control performance. This can be seen in Figure 7, where we applied both types of investigated state-dependent charts to the Phase II data from Section 5. Comparing the decoded states with the true ones, see the table at the bottom of Figure 7, we observe that during the out-of-control period, the decoded states are often to high. This, however, causes the state-dependent c chart to choose the "wrong" control limit, and the residuals are not sufficiently increased in value. In simulations, we also observed that the ARL performance often shows strong breaks of monotonicity similar to the performance of the above log-LR$_\lambda$ chart, but much more pronounced at the corresponding shift sizes.

In view of the previous results, it appears to be more promising to develop such control charts, which make use of a possible misidentification of the hidden states (in order to detect the out-of-control scenario), but do not suffer from it. A possible approach could be to run multiple (two-sided) Bernoulli CUSUM control charts as introduced by Reynolds and Stoumbos[29] in parallel, to directly monitor the hidden states. For example, we may use one chart for each type of hidden state, see the analogous application by Ryan et al[30] for the monitoring of a categorical process (but Bernoulli CUSUM charts for certain combinations of states might also turn out to be useful). In case of a three-state HMM, this means that three two-sided Bernoulli CUSUM charts with individual reference values and control limits constitute one large scheme (thus, 12 design parameters—in contrast to just 2 parameters for the ordinary CUSUM chart from Section 4). For each $q \in \mathcal{Q}$, the corresponding qth Bernoulli CUSUM accumulates the binary indicators $\mathbb{1}(\hat{q}_t \neq q)$,

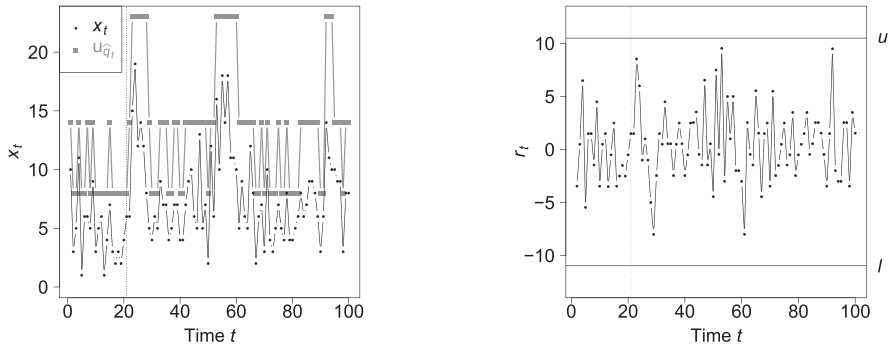

FIGURE 7 Left: State-dependent c chart ($u_0 = 8$, $u_1 = 14$, $u_2 = 23$). Right: Residuals chart ($u = 10.50349$, $l = -10.96609$). Bottom: True versus decoded hidden states

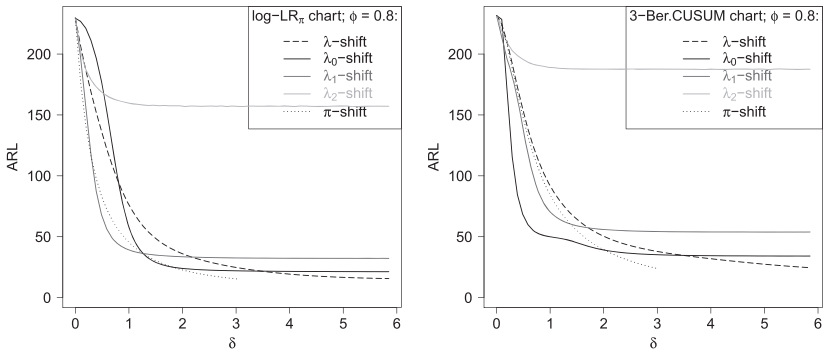

FIGURE 8 ARL performances against mean shifts $\delta = \mu - \mu_0$, $\mu_0 = 1.95$, of log-LR$_{\boldsymbol{\pi}}$ and joint trivariate Bernoulli CUSUM chart for $\phi = 0.8$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Thus, this chart has the potential to detect any type of out-of-control scenario that effects the identification of the hidden states. Figure 8 presents some first results from a simulation experiment (again with 10^6 replications). These results, however, do not show a clear advantage of the joint Bernoulli CUSUM. Its ARL graphs are close to the ones of the log-LR$_{\boldsymbol{\pi}}$ chart (which is the natural counterpart among our HMM CUSUM schemes), but with the latter having less design parameters. In the case of λ_2-shifts, the joint Bernoulli CUSUM chart performs considerably worse than the log-LR$_{\boldsymbol{\pi}}$ chart, whereas it seems to be more sensitive towards λ_0-shifts. Altogether, it appears worth to be analyzed in a future research if the design of the joint Bernoulli CUSUM chart can be optimized to perform superiorly in further out-of-control scenarios.

7 | CONCLUSIONS

In this work, we considered the task of monitoring counts being generated by a Poisson HMM. Besides the well-established c and CUSUM chart, we also proposed types of log-LR CUSUM charts for this purpose. In comparing their ARL performance with regard to diverse out-of-control scenarios, it turned out that both the c chart and the log-LR charts may show a rather bad ARL performance, depending on the actual type of mean shift. The log-LR charts can only be recommended if one can be sure that the type of out-of-control scenario to be expected agrees well with the anticipated one. The ordinary CUSUM chart, in contrast, showed nearly the same ARL performance for any scenario and any dependence level. Therefore, it can generally be recommended as a reliable chart for monitoring mean shifts in a Poisson HMM.

Although the ordinary CUSUM chart turned out to be a good "all-purpose chart", it is still a relevant question to ask if the log-LR approach could be modified to make it an attractive alternative for practice. A solution could be to consider a GLR chart instead, also see previous studies,[7,31,32] where the fixed target value $\boldsymbol{\theta}_1$ in (3.3) is omitted by considering the statistic

$$G^t = \max_{\tau,\,\theta} \ln\left(\frac{P(X_t,\ldots,X_{\tau+1}\,|X_\tau,\ldots,X_1;\theta)}{P(X_t,\ldots,X_{\tau+1}\,|X_\tau,\ldots,X_1;\theta_0)}\right)$$

at time t. This statistic, however, does not only require to maximize in τ (i. e., to find the most probable position of the change point), but for each value of $\tau \in \{1,\ldots,t\}$, a maximization also has to be done in $\boldsymbol{\theta}$ (i. e., altogether t maximum likelihood estimations). Because ML estimation is computationally very demanding for HMMs, this does not seem to be feasible in practice, even if using a moving-window technique like in Wang and Reynolds.[31] Therefore, it would be an interesting task for future research to find out if the computational burden could be substantially reduced by employing appropriate constraints on the out-of-control parameter vector $\boldsymbol{\theta}$, for example, by keeping the parameters related to $(Q_t)_{\mathbb{N}}$ fixed but by jointly maximizing about all state-dependent means λ_q. Certainly, it then has to be checked if such a constrained GLR chart does show a more stable ARL behavior.

Another relevant topic for future research was already presented in Section 6, namely, trying to find a successful way of state-based process monitoring. Although the state-dependent Shewhart charts suffer from the problem of decoding the hidden states, the idea of a joint Bernoulli CUSUM chart for state monitoring appears to be quite promising. Finally, the presented performance analyses rely on specified model parameters; future research should also investigate the effect of parameter estimation on the charts' performance.

ACKNOWLEDGEMENT

The authors thank the referee for very useful comments on an earlier draft of this article.

ORCID

Christian H. Weiß https://orcid.org/0000-0001-8739-6631
Sven Knoth https://orcid.org/0000-0002-9666-5554

REFERENCES

1. Zucchini W, MacDonald IL, Langrock R. *Hidden Markov Models for Time Series: An Introduction Using R*. 2nd ed. London: Chapman & Hall/CRC; 2016.
2. Weiß CH. *An Introduction to Discrete-Valued Time Series Chichester*: John Wiley & Sons, Inc; 2018.
3. Montgomery DC. *Introduction to Statistical Quality Control*. 6th ed. New York: John Wiley & Sons, Inc; 2009.
4. Yontay P, Weiß CH, Testik MC, Bayindir ZP. A two-sided CUSUM chart for first-order integer-valued autoregressive processes of Poisson counts. *Qual Reliab Eng Int*. 2013;29(1):33–42.
5. Alshraideh H, Runger G. Process monitoring using hidden Markov models. *Qual Reliab Eng Int*. 2014;30(8):1379–87.
6. Fuh C-D. SPRT and CUSUM in hidden Markov models. *Annals of Statistics*. 2003;31(3):942–77.
7. Fuh C-D, Mei Y. Quickest change detection and Kullback-Leibler divergence for two-state hidden Markov models. *IEEE Trans Signal Process*. 2015;63(18):4866–78.
8. Sparks R. Challenges in designing a disease surveillance plan: what we have and what we need? *IIE Trans on Healthc Systems Eng*. 2013;3(3):181–92.
9. Rafei A, Pasha E, Jamshidi Orak R. Tuberculosis surveillance using a hidden Markov model. *Iranian J Public Health*. 2012;41(10):87–96.
10. Simões A, Viegas JM, Farinha JT, Fonseca I. The state of the art of hidden Markov models for predictive maintenance of Diesel engines. *Qual Reliab Eng Int*. 2017;33(8):2765–79.
11. Heck LP, McClellan JH. Mechanical system monitoring using hidden Markov models. In: Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 91); 1991:1697–700.
12. Boutros T, Liang M. Detection and diagnosis of bearing and cutting tool faults using hidden Markov models. *Mech Syst Signal Process*. 2011;25(6):2102–24.
13. Kisić E, Durović Ž., Kovačević B., Petrović V.. Application of T^2 control charts and hidden Markov models in condition-dased maintenance at thermoelectric power plants. *Shock Vib*. 2015;2015:960349.
14. MacDonald IL, Zucchini W. Hidden Markov models for discrete-valued time series. In: R. A. Davis, S. H. Holan, R. Lund, N. Ravishanker, eds. *Handbook of Discrete-Valued Time Series*. Boca Raton: Chapman & Hall/CRC Press; 2016:267–286.
15. Adam T, Langrock R, Weiß CH. Penalized estimation of flexible hidden Markov models for time series of counts. *Metron*. 2019;77(2):87–104.
16. Harte D. *HiddenMarkov: Hidden Markov Models*, R package version 1.8-11. Wellington: Statistics Research Associates; 2017.
17. Box GEP. Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GH, eds. *Robustness in Statistics*. New York: Academic Press; 1979:201–236.
18. MacCarthy B, Wasusri T. A review of non-standard applications of statistical process control (SPC) charts. *Int J Qual Reliab Manag*. 2002;19(3):295–320.
19. Sebastian T, Jeyaseelan V, Jeyaseelan L, Anandan S, George S, Bangdiwala SI. Decoding and modelling of time series count data using Poisson hidden Markov model and Markov ordinal logistic regression models. *Stat Methods Med Res*. 2019;28(5):1552–63.
20. Watts C, Hahn C, Sohn B. Monitoring the performance of a reorder point system: a control chart approach. *Int J Oper Prod Manag*. 1994;14(2):51–61.
21. Pfohl H-C, Cullmann O, Stölzle W. Inventory management with statistical process control: simulation and evaluation. *J Bus Logist*. 1999;20(1):101–20.
22. Cheng J-C, Chou C-Y. A real-time inventory decision system using Western Electric run rules and ARMA control chart. *Expert Syst Appl*. 2008;35(3):755–61.
23. Schmid W. On the run length of a Shewhart chart for correlated data. *Stat Pap*. 1995;36(1):111–30.
24. Weiß CH, Testik MC. CUSUM monitoring of first-order integer-valued autoregressive processes of Poisson counts. *J Qual Technol*. 2009;41(4):389–400.
25. Ottenstreuer S, Weiß CH, Knoth S.. A combined Shewhart-CUSUM chart with switching limit. *Qual Eng*. 2019;31(2):255–68.
26. Knoth S. The art of evaluating monitoring schemes—How to measure the performance of control charts? In: Lenz H-J, ed. *Frontiers in Statistical Quality Control 8*. Heidelberg: Physica-Verlag; 2006:74–99.
27. Brook D, Evans DA. An approach to the probability distribution of CUSUM run length. *Biometrika*. 1972;59(3):539–49.
28. Lucas JM. Counted Data CUSUM's. *Technometrics*. 1985;27(2):129–44.
29. Reynolds JMR, Stoumbos ZG. A CUSUM chart for monitoring a proportion when inspecting continuously. *J Qual Technol*. 1999;31(1):87–108.
30. Ryan AG, Wells LJ, Woodall WH. Methods for monitoring multiple proportions when inspecting continuously. *J Qual Technol*. 2011;43(3):237–248.

31. Wang N, Reynolds JR. The generalized likelihood ratio chart for monitoring a proportion with autocorrelation. *Qual Reliab Eng Int*. 2015;31(6):1023–34.
32. Lee J, Woodall WH. A note on GLR charts for monitoring count processes. *Qual Reliab Eng Int*. 2018;34(6):1041–44.
33. Weiß CH. The Markov chain approach for performance evaluation of control charts—a tutorial. In: Werther SP, ed. *Process Control: Problems, Techniques and Applications*. New York: Nova Science Publishers, Inc; 2011:205–228.

AUTHOR BIOGRAPHIES

Sebastian Ottenstreuer is a research assistant in the Department of Mathematics and Statistics at the Helmut Schmidt University in Hamburg, Germany. He received his BSc (2013) and MSc (2016) degrees in Economathematics from the University of Würzburg, Germany.

Christian H. Weiß is a Professor at the Department of Mathematics and Statistics at the Helmut Schmidt University in Hamburg, Germany. His research areas include time series analysis, statistical quality control, and computational statistics.

Sven Knoth is a Professor at the Department of Mathematics and Statistics at the Helmut Schmidt University in Hamburg, Germany. His research areas include statistical process control, computational statistics, and engineering statistics.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Ottenstreuer S, Weiß CH, Knoth S. Control charts for monitoring a Poisson hidden Markov process. *Qual Reliab Engng Int*. 2021;37:484–501. https://doi.org/10.1002/qre.2745

APPENDIX A: STATISTICAL INFERENCE FOR HMMs

This appendix provides a brief summary of relevant stochastic properties of HMMs (beyond those properties already discussed in Section 2) as well as of approaches for the parameter estimation, forecasting, and decoding of the hidden states; more detailed discussions can be found in the books by other studies.[1,2]

Let $\mathbf{P}(x)$ denote the diagonal matrix $\mathbf{P}(x) := \mathrm{diag}(p(x|0),\ldots,p(x|d_{\mathcal{Q}})) \in [0;1]^{(d_{\mathcal{Q}}+1)\times(d_{\mathcal{Q}}+1)}$ of state-dependent probabilities for observing the count $x \in \mathbb{N}_0$. Then, the marginal PMF is computed as $P(X_t = x) = \mathbf{1}^\top \mathbf{P}(x)\pi$, where $\mathbf{1}$ denotes the vector of ones. In particular, marginal mean and variance of the Poisson HMM are given by

$$\mu = \sum\nolimits_{q\in\mathcal{Q}} \pi_q \lambda_q,\ \sigma^2 = \mu + \sum\nolimits_{q\in\mathcal{Q}} \pi_q \lambda_q^2 - \mu^2 \geq \mu.$$

The autocovariance function $\gamma(k) = Cov[X_t, X_{t-k}]$ is computed as

$$\gamma(k) = \sum\nolimits_{q,r\in\mathcal{Q}} (\mathbf{A}^k)_{q,r} \pi_r \lambda_q \lambda_r - \mu^2.$$

For an efficient implementation of ML estimation, forecasting, and decoding schemes, the so-called "forward probabilities" $\boldsymbol{\alpha}_t$ and "backward probabilities" $\boldsymbol{\beta}_t$ at time t are required. They are defined as

$$\alpha_{t,q} = P(X_t = x_t,\ldots, X_1 = x_1, Q_t = q) \text{ and}$$
$$\beta_{t,q} = P(X_{t+1} = x_{t+1},\ldots, X_T = x_T \,|\, Q_t = q),$$

respectively, for all $q \in \mathcal{Q}$, and they can be computed recursively from

$$\alpha_1 = \mathbf{P}(x_1)\pi, \qquad \alpha_t = \mathbf{P}(x_t)\mathbf{A}\alpha_{t-1};$$
$$\beta_T = \mathbf{1}, \qquad \beta_t^\top = \beta_{t+1}^\top \mathbf{P}(x_{t+1})\mathbf{A}.$$

This, in turn, enables a recursive computation of the likelihood function, because $L(\theta) = \mathbf{1}^\top \alpha_T$ holds. It was also used to derive the recursive scheme (3.2), which is computationally more stable because of the scaling $\boldsymbol{\nu}_t = \boldsymbol{\alpha}_t/(w_1 \ldots w_{t-1})$ with $w_t = \mathbf{1}^\top \boldsymbol{\nu}_t$.

For an h-step-ahead forecasting of the HMM's observations, we compute the forecast distribution

$$P(X_{t+h} = x | x_t, \ldots, x_1) = \frac{\mathbf{1}^\top \mathbf{P}(x)\mathbf{A}^h \alpha_t}{\mathbf{1}^\top \alpha_t} = \mathbf{1}^\top \mathbf{P}(x)\mathbf{A}^h \boldsymbol{\nu}_t / w_t,$$

and take the mode or median thereof as the point forecast value. Analogously, we can predict future hidden states from

$$P(Q_{t+h} = q | x_t, \ldots, x_1) = \frac{\boldsymbol{e}_q^\top \mathbf{A}^h \alpha_t}{\mathbf{1}^\top \alpha_t} = \boldsymbol{e}_q^\top \mathbf{A}^h \boldsymbol{\nu}_t / w_t,$$

where $\boldsymbol{e}_q$ is the qth unit vector. Concerning a decoding of the already realized (but invisible) hidden states, we distinguish between the "local decoding" of a single hidden state, and the "global decoding" of all hidden states. To decode q_t with $1 \le t \le T$, given the observations $x_1, \ldots, x_T$, we compute

$$\hat{q}_t := \arg\max_q P(Q_t = q | x_T, \ldots, x_1) = \arg\max_q \frac{\alpha_{t,q} \cdot \beta_{t,q}}{\mathbf{1}^\top \alpha_T}.$$

For an online local decoding, we set $T = t$ in this scheme (or $h = 0$ in the above forecast distribution for Q_{t+h}), leading to $\hat{q}_t = \operatorname{argmax}_q \nu_{t,q}$. A global decoding can be done by using the "Viterbi algorithm" to maximize $P(Q_T = q_T, \ldots, Q_1 = q_1 \mid x_T, \ldots, x_1)$. For each $q \in \mathcal{Q}$, we define the probabilities

$$m_{1,q} := P(Q_1 = q, X_1 = x_1) = p(x_1|q)\pi_q,$$
$$m_{t+1,q} := \max_{q_1,\ldots,q_t} P(Q_{t+1} = q, Q_t = q_t, \ldots, Q_1 = q_1, X_{t+1} = x_{t+1}, \ldots, X_1 = x_1).$$

These are computed recursively as

$$m_{t+1,q} = p(x_{t+1}|q) \cdot \max_r \{m_{t,r} \cdot a_{q|r}\} \text{ for } t \ge 1.$$

Then the globally decoded states are

$$\hat{q}_T := \arg\max_q \{m_{T,q}\}, \; \hat{q}_t := \arg\max_q \{m_{t,q} \cdot a_{\hat{q}_{t+1}|q}\} \text{ for } t = T-1, \ldots, 1.$$

APPENDIX B: IMPLEMENTATION OF MC APPROACH

Let us describe the implementation of the MC approach by Brook and Evans[27] for the c chart and the CUSUM chart applied to a HMM; also see Weiß[33] for general details concerning the MC approach. The approach assumes that the monitored process can be represented as a finite homogeneous MC $(Z_t)_{\mathbb{N}}$ with state space $\mathcal{S} = \mathcal{N} \cup \{\mathrm{a}\}$, where the set $\mathcal{N}$ contains the "no-alarm states" and the set {a} consists of a single "alarm state" "a." We define $\mathbf{Z}^\top$ to be the transition matrix for the states in $\mathcal{N}$, that is, $\mathbf{Z}^\top := (p_{i|j})_{i,j \in \mathcal{N}}$. Furthermore, let $\mathbf{I}$ be the identity matrix and $\mathbf{1}$ the vector of ones. To compute the zero-state ARL, we first have to solve the linear equation $(\mathbf{I} - \mathbf{Z})\boldsymbol{\mu} = \mathbf{1}$ in $\boldsymbol{\mu}$. If $\boldsymbol{z}$ denotes the vector of initial probabilities $P(Z_1 = j)$ for $j \in \mathcal{N}$, then the zero-state ARL is computed as $1 + \boldsymbol{\mu}^\top \boldsymbol{z}$.

Note that the choice of $\boldsymbol{z}$ is not unique if computing out-of-control ARLs, see the discussion of Scheme 2.3 in Weiß.[33] For computational simplicity, we followed the approach in Weiß[33] and defined $\boldsymbol{z}$ as the stationary marginal

distribution of the actual out-of-control model. We also experimented with different initializations $\boldsymbol{z}$, but the effect was very small.

ARL computation for c chart

The bivariate process $(X_t, Q_t)_{\mathbb{N}_0}$ is a MC with transition probabilities

$$P((X_t,Q_t)=(x,q)\,|\,(X_{t-1},Q_{t-1})=(y,r))=P(X_t=x|Q_t=q)P(Q_t=q|Q_{t-1}=r)=p(x|q)a_{q|r}.$$

The set of "no-alarm states" is given by

$$\mathcal{N}=\left\{(x,q)^\top\in\mathbb{N}_0\times\mathcal{Q}\mid x\le u\right\},$$

which is a finite set of size $(u+1)(d_{\mathcal{Q}}+1)$ such that the MC approach is applicable.

ARL computation for CUSUM chart

The trivariate process $(X_t, Q_t, C_t)_{\mathbb{N}_0}$ is an MC with transition probabilities

$$\begin{aligned}&P((X_t,Q_t,C_t)=(x,q,c)\,|\,(X_{t-1},Q_{t-1},C_{t-1})=(y,r,d))\\&=P(C_t=c|X_t=x,C_{t-1}=d)P(X_t=x|Q_t=q)P(Q_t=q|Q_{t-1}=r)\\&=\mathbb{1}(c=\max\{0,x-k+d\})p(x|q)a_{q|r},\end{aligned}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. For ease of implementation, we choose $k,h\in\mathbb{Q}^+$ (with same denominator n), so $C_t\in\mathcal{C}:=\{0,1/n,2/n,...,h\}$. It can then be concluded with same arguments as in Weiß[24] that the set of "no-alarm states" equals

$$\mathcal{N}=\{(x,q,c)\in\mathbb{N}_0\times\mathcal{Q}\times\mathcal{C}\,c\le h,c+k-h\le x\le c+k\}. \tag{B1}$$

Note that not all potential "no-alarm states" can be attained by (X_t, Q_t, C_t) given the triple $(X_{t-1},Q_{t-1},C_{t-1})\in\mathcal{N}$. Therefore, using sparse matrix techniques for the transition matrix is an efficient way of implementing the MC approach in order to compute the ARL.