Computers & Industrial Engineering 113 (2017) 502–511

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Directional monitoring of categorical processes with serial dependence

Dong Ding[a], Dongdong Xiang[b], Jian Li[c,*]

[a] School of Management, Xi'an Polytechnic University, No. 19, Jinhua South Road, Xi'an, Shaanxi 710048, China
[b] School of Statistics, East China Normal University, No. 500, Dongchuan Road, Shanghai 200241, China
[c] School of Management and State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, No. 28, Xianning West Road, Xi'an, Shaanxi 710049, China

ARTICLE INFO

Keywords:
Autocorrelation coefficient
Conditional probability
Contingency table
Log-linear model
Statistical process control

ABSTRACT

The monitoring of a categorical process with serial dependence in which the current observation depends on its past values is of great importance in many applications, including manufacturing and service management. However, a great majority of existing research works are restricted to the cases where data are binary and of first-order dependency, based on the assumption of a two-state first-order Markov chain. In this article, a general categorical process with serial dependence that can have more than two attribute levels and higher-order dependency structure is under consideration. We adopt the multivariate representation of the categorical variables and integrate directional shift information into an adjusted log-linear model. Based on this, a novel control chart is proposed for detecting shifts in the marginal distribution and in the dependence structure of serially dependent categorical processes. Simulations have demonstrated its efficiency and robustness. The implementation of the proposed control chart through a real example is provided as the guidance for practitioners.

1. Introduction

To maintain and improve the quality of processes, statistical process control (SPC) proves its strength and effectiveness. For various types of data different control charts have been developed accordingly. Among all these methods, it is often assumed that observations are independent of each other. In many applications however, this assumption is invalid for some reason such as high frequency or short intervals of sampling. For example in the textile industry, in the production line of cotton yarn especially, quality inspection to see whether the yarn is conforming or not is in fact conducted at very high speed, making serial dependence very common.

For such autocorrelated processes, conventional control charts with the independence assumption cannot work well, or even be misleading sometimes. See Montgomery (2009) and references therein. Specifically for autocorrelated continuous data, many works have been done to deal with their monitoring. One may refer to Montgomery and Mastrangelo (1991), Lu and Reynolds (1999), Apley and Shi (1999), Jiang, Tsui, and Woodall (2000), Jiang (2004), Psarakis and Papaleonida (2007), Zou, Wang, and Tsung (2008), Huang, Bisgaard, and Xu (2014) and so on.

However, for serially dependent categorical data, monitoring techniques still remain rare. Recent research has made progress in monitoring serially dependent binary processes that have two attribute levels, such as good or bad. The main idea is to use a two-state Markov chain. See Bhat and Lal (1990), Shepherd, Champ, Ridgon, and Fuller (2007), Mousavi and Reynolds (2009) for example. Only first-order dependency is considered by the above mentioned methods because of the assumption of a first-order Markov chain. Recently, He, Wang, Tsung, and Shang (2016) proposes to use the bivariate binomial AR(1) model for the monitoring of autocorrelated bivariate binomial processes, still accounting for only first-order dependency and binary data.

Furthermore, if the categorical process is multinary, i.e., it has more than two attribute levels, such as good, marginal, and bad, its monitoring will become even more complex. For independent multinary processes, Marcucci (1985), Duran and Albin (2010), Ryan, Wells, and Woodall (2011) and Weiß (2012) proposed corresponding control charts. For multinary categorical processes with serial dependence, Weiß (2016) applied the Pearson chi-square statistic and the Gini index to their monitoring, and found that their distributions are different from those obtained in the independent cases.

To handle a general serially dependent categorical process with h levels ($h \geqslant 2$) and dth-order dependency ($d \geqslant 1$), there exist many challenges. One major problem is how to model such processes. Though a dth-order Markov chain with h states is a natural candidate, it requires $h^d(h-1)$ parameters which may lead to overparametrization (Kedem & Fokianos, 2002). Dimension reduction techniques are needed if higher-order Markov chains are used. Another observation is that current methods derived from Markov chains mainly focus on first-

* Corresponding author.
E-mail addresses: laceyding@foxmail.com (D. Ding), terryxdd@163.com (D. Xiang), jianli@xjtu.edu.cn (J. Li).

http://dx.doi.org/10.1016/j.cie.2017.09.035
Received 26 May 2017; Received in revised form 17 September 2017; Accepted 20 September 2017
Available online 28 September 2017
0360-8352/

order dependency, while marginal probabilities and higher-order dependency are ignored. To make the monitoring scheme more effective, it is necessary to encompass the information of the marginal distribution and the dependence structure simultaneously.

To settle the aforementioned problems of monitoring a categorical process with a general number h of attribute levels and with serial dependence of a general order d, we first transform the observations of the process to a multi-way contingency table. Then an adjusted log-linear model is applied to model the cell counts in this contingency table. The underlying idea is to use the multivariate representation of an autocorrelated process, which is similar to Apley and Tsung (2002) and Jiang (2004) for monitoring autocorrelated continuous processes.

Besides, some directional shift information is also exploited and incorporated into the log-linear model. Such information considers the most possible shift patterns in practice and facilitates the monitoring. Based on this, a novel control chart is proposed, which can be implemented to a serially dependent categorical process with h ($h \geqslant 2$) levels and dth-order dependency ($d \geqslant 1$). Therefore, this chart is able to efficiently detect changes in the marginal distribution or dependence structure.

The remainder of this article is organized as follows. First, the modeling of a categorical process with serial dependence is introduced in Section 2. The directional monitoring approach is proposed in Section 3. Comparison studies are shown and analyzed in Section 4, followed by a case study in Section 5. Finally Section 6 concludes this article. Some derivations are provided in the appendix.

2. Log-linear modeling

Suppose that a categorical factor X has h attribute levels, and that the sequence $\{X_t\}$ has dth-order Markov dependency, i.e., at time t,X_t depends on $X_{t-1},\ldots,X_{t-d}$. We use $(d+1)$-dimensional multivariate vectors to represent this univariate process. To be specific, let $\mathbf{y} = [Y_1,\ldots,Y_{d+1}]^T$ be a vector of $d+1$ factors. At each time point t, let $\mathbf{y}_t = [Y_{t,1},\ldots,Y_{t,d+1}]^T = [X_{t-d},\ldots,X_t]^T$. As a result, the observations of all the h^{d+1} cross-classifications among the level combinations of $Y_{t,i}$ ($i = 1,\ldots,d+1$) will form a square contingency table of dimension $\underbrace{h\times\ldots\times h}_{d+1}$, with the count of a level combination stored in each cell. Let $p_{a_1\ldots a_{d+1}}$ be the joint probability of the occurrence of level combination $(a_1,\ldots,a_{d+1})$ where $a_i = 1,\ldots,h$ for $i = 1,\ldots,d+1$, i.e., , and let $n_{a_1\ldots a_{d+1}}$ be the count. Then in a sample of size N, we have

$$\sum_{a_1,\ldots,a_{d+1}} p_{a_1\ldots a_{d+1}} = 1\,, \qquad \sum_{a_1,\ldots,a_{d+1}} n_{a_1\ldots a_{d+1}} = N,$$

and the cell counts in the contingency table jointly follow the multinomial distribution $\mathrm{MN}(N;\{p_{a_1\ldots a_{d+1}}\})$.

Based on the multivariate representation, log-linear models can be used to characterize categorical processes with serial dependence, which is similar to Li, Tsung, and Zou (2012). Generally log-linear models relate the expectations of cell counts to main factor effects and interaction effects. By assuming that each cell count follows a Poisson distribution, a generalized linear model with the canonical link function results (McCullagh & Nelder, 1989). In other words, for each cell count, the logarithm of its expectation has a linear relationship with main factor effects and interaction effects. However, given a sample of size N within which all the cell counts sum up to N, then the cell counts jointly follow the multinomial distribution $\mathrm{MN}(N;\mathbf{p})$ where $\mathbf{p}$ is the probability vector. Hence the log-linear model can be rewritten based on the probability vector $\mathbf{p}$ instead of expectations $N\mathbf{p}$. For example, suppose that $\{X_t\}$ has second-order dependency and three factors Y_1,Y_2,Y_3 are used to describe this serially dependent categorical process. Then each cell corresponds to a level combination with probability $p_{a_1a_2a_3}$ where $a_1,a_2,a_3 = 1,\ldots,h$. The log-linear model is written as

$$\ln p_{a_1a_2a_3} = u^{(0)} + u^{(1)}_{a_1} + u^{(2)}_{a_2} + u^{(3)}_{a_3} + u^{(1,2)}_{a_1a_2} + u^{(1,3)}_{a_1a_3} + u^{(2,3)}_{a_2a_3} + u^{(1,2,3)}_{a_1a_2a_3} \tag{1}$$

with $\sum_{a_1,a_2,a_3} p_{a_1a_2a_3} = 1$. Here $u^{(0)}$ is the intercept, $u^{(1)},u^{(2)},u^{(3)}$ are the main effects, $u^{(1,2)},u^{(1,3)},u^{(2,3)}$ are the two-factor interaction effects, and $u^{(1,2,3)}$ is the three-factor interaction effect.

To illustrate how the log-linear model characterizes the dependence structure of a serially dependent categorical process, we still use the above example and consider several cases below. If $\{X_t\}$ is an independent sequence, i.e., X_t does not rely on X_{t-1},X_{t-2}, then the three factors Y_1,Y_2,Y_3 are independent of each other. Hence model (1) would reduce to

$$\ln p_{a_1a_2a_3} = u^{(0)} + u^{(1)}_{a_1} + u^{(2)}_{a_2} + u^{(3)}_{a_3},$$

which contains the main effects only. If the process has only first-order dependency, that is, given X_{t-1},X_t is conditionally independent of X_{t-2}, then model (1) becomes

$$\ln p_{a_1a_2a_3} = u^{(0)} + u^{(1)}_{a_1} + u^{(2)}_{a_2} + u^{(3)}_{a_3} + u^{(1,2)}_{a_1a_2} + u^{(2,3)}_{a_2a_3},$$

which implies that given Y_2,Y_1 and Y_3 are independent of each other. Note that in this case, if the two-factor interaction effect $u^{(1,3)}$ is added, then second-order dependency exists. In a nutshell, the main effects $u^{(1)},u^{(2)}$, and $u^{(3)}$ decide the marginal probability distribution of the serially dependent categorical process, $u^{(1,2)}$ and $u^{(2,3)}$ correspond to first-order dependency, $u^{(1,3)}$ and $u^{(1,2,3)}$ reflects second-order dependency.

In a general categorical process with dth-order dependency, the associated $d+1$ factors $Y_1,\ldots,Y_{d+1}$ lead to the log-linear model

$$\ln\mathbf{p} = \mathbf{1}\beta_0 + \sum_{i=1}^{2^{d+1}-1} \mathbf{A}_i\boldsymbol{\beta}_i, \tag{2}$$

with $\mathbf{1}^T\mathbf{p} = 1$, constrained by some identifiability requirements described in Li et al. (2012). Here $\mathbf{p}$ is the cell probability vector of dimension $h^{d+1}\times 1$,$\mathbf{1}$ is a column vector with 1 as all its entries, $\mathbf{A}_i$ is the known design matrix with elements 1, or -1, or 0 and appropriate dimension. The derivation of design matrices can be found in the appendix of Li et al. (2012). Moreover, β_0 represents the intercept, and other $\boldsymbol{\beta}_i$'s represent main effects or interaction effects, among which $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{d+1}$ correspond to the main effects, $\boldsymbol{\beta}_{d+2},\ldots,\boldsymbol{\beta}_{(d+1)(d+2)/2}$ correspond to the two-factor interaction effects. Furthermore, due to the constraint $\mathbf{1}^T\mathbf{p} = 1$, the parameters $\boldsymbol{\beta}_i$ ($i = 1,\ldots,2^{d+1}-1$) are independent and can vary freely, and the intercept β_0 is determined and expressed by $\boldsymbol{\beta}_i$ ($i = 1,\ldots,2^{d+1}-1$). Another observation is that the well-known logit model can be easily derived from the above log-linear model. However the logit model is not a good candidate for serially dependent categorical processes since it cannot describe the dependency structure.

3. Directional monitoring

Based on the log-linear modeling of the multivariate representation of serially dependent categorical processes, we can now turn to the detection of abnormalities. This work focuses on Phase II monitoring, so it is assumed that the in-control (IC) parameters in model (2) are known or have been estimated in the retrospective analysis of Phase I already, including the dependence order $d = d_0$, the coefficient vectors $\boldsymbol{\beta}_i^{(0)}$ ($i = 0,1,\ldots,2^{d_0+1}-1$), and the cell probability vector $\mathbf{p}^{(0)}$. The dependence structure is supposed to be determined by a d_0th-order Markov process, and let $q^{(0)}_{a_1\ldots a_{d_0+1}}$ denote the conditional probability $\Pr^{(0)}(X_t = a_{d_0+1}|X_{t-d_0} = a_1,\ldots,X_{t-1} = a_{d_0})$ where $a_i = 1,\ldots,h$ for $i = 1,\ldots,d_0+1$, that is, the conditional probability of X_t at level a_{d_0+1} given X_{t-d_0} up to X_{t-1}.

Shifts may occur either in the dependence order or in the conditional probabilities. To express the change-point model, first let $\mathbf{x}_t$ be a column indicator vector of X_t, that is, $\mathbf{x}_t$ is of dimension $h\times 1$ with 1 as its kth element if X_t is at level k and 0's otherwise. Then given the past values $a_1,\ldots,a_d$ of X_{t-d} up to X_{t-1}, the elements of $\mathbf{x}_t$ jointly follow the multinomial distribution $\mathrm{MN}(1;\{q_{a_1\ldots a_{d+1}}\})$ where $a_{d+1} = 1,\ldots,h$. Hence the change-point model can be written as