Computers & Industrial Engineering 72 (2014) 198–205

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

LASSO-based diagnosis scheme for multistage processes with binary data ☆

Yanfen Shang [a,*], Xuemin Zi [b], Fugee Tsung [c], Zhen He [a]

[a] College of Management and Economics, Tianjin University, China
[b] School of Science, Tianjin University of Technology and Education, China
[c] Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 17 October 2013
Received in revised form 13 January 2014
Accepted 11 March 2014
Available online 21 March 2014

Keywords:
BIC
Binary state-space model
Fault isolation
LASSO
Variable selection

ABSTRACT

Process monitoring using multistage processes with binary data remains an important and challenging problem in statistical process control (SPC). Although the multistage processes has been extensively studied in the literature, the challenges associated with designing diagnostic schemes when only binary responses are observed are yet to be addressed well. This paper develops a practical LASSO-based diagnostic procedure which combines BIC with the popular adaptive LASSO variable selection method. Given the oracle property of LASSO and its algorithm, the diagnostic result can be obtained easily and quickly. More importantly, the proposed method does not require making any extra tests that are necessary in existing diagnosis methods. Numerical and real-data examples demonstrate the effectiveness of our method.

1. Introduction

Statistical process control (SPC) has been widely used to monitor process or product quality in the manufacturing and service industries. One of primary tasks in SPC is to detect the shift in process or product quality characteristics. Apart from the detection of the shift presence, another primary task of SPC is to identify the change point when the shift occurred and to detect what variables or parameters shifted after detecting the shift by using control charts, which can be terms as diagnosis problem (Mason & Young, 2002). Once the out-of-control signal is triggered by the control charts, the diagnosis tools can help the engineers to identify abnormal variables or the root cause of the problem and to restore the process to the normal condition quickly. Especially for the product or process quality with a complex data structure, diagnosis becomes very critical. For instance, if the quality is characterized by several correlated variables, the diagnostic aid is needed to isolate the shifting variables; if the process comprises multiple stages, such as print circuit board (PCB) manufacturing process, after detecting the shift, the faulty stages need to be diagnosed to eliminate the root cause of the change.

A lot of the research on the diagnosis problem has been widely conducted in the literature. Previous diagnostic work has focused on identifying the change point when the shift occurred (Nedumaran & Pignatiello, 1998; Nedumaran, Pignatiello, & Calvin, 2000; Sullivan & Woodall, 2000; Sullivan, 2002; Zou, Tsung, & Liu, 2008). Besides the change point estimation, some effort has also been devoted to shifted variables or parameters. One of the methods for accomplishing this task is based on decomposition, such as the decomposition of the T^2 statistic, the decomposition based on principal components and factor analysis (cf. Apley & Shi, 2001; Jackson, 1985; Li, Jin, & Shi, 2008; Mason, Tracy, & Young, 1995, 1997; Sun, Tsung, & Qu, 2007). Various step-down diagnostic procedures are also popularly studied in the literature, such as Hawkins (1991), Mason, Chou, and Young (2001), Marvelakis, Bersimis, Panaretos, and Psarakis (2002), Chen and Wang (2004), Cheng and Cheng (2011) and Sullivan, Stoumbos, Mason, and Young (2007). Moreover, simultaneous testing procedures are also used to diagnose multiple changed variables (cf. Doganaksoy, Faltin, & Tucker, 1991; Du & Xi, 2011; Hayter & Tsui, 1994; Zhu & Jiang, 2009). Although these conventional methods are basically sound, they are mainly for the multivariate diagnosis problems and also computationally expensive.

A few of research on multistage process diagnosis methods can be found in the literature. Ding, Shi, and Ceglarek (2002) and Zhou, Ding, Chen, and Shi (2003) investigated the diagnosability of a multistage process. Zou and Tsung (2008) proposed the diagnostic scheme using directional information. However, their method failed when more than one faults occurred in the process. Li and Tsung (2009) proposed the scheme to identify multiple faulty

☆ This manuscript was processed by Area Editor Min Xie.

* Corresponding author. Tel.: +86 15102266607.
E-mail address: syf8110@gmail.com (Y. Shang).

http://dx.doi.org/10.1016/j.cie.2014.03.013
0360-8352/

stages occurred in the multistage process. Their method is to firstly obtain the one-step forecast errors (OSFE) or residuals based on the process model and then apply the false discovery rate (FDR) control approach to the residuals. It is applicable for multiple-faults cases. However, as shown by Zou and Tsung (2008), the accuracy of OSFE procedure fails in certain cases and hardly improves as the sample sizes become larger. This is because the effectiveness of OSFE relies on the implicit assumption that the OSFE's at the true faulty-stages significantly deviate away from zero but those at the normal-stages are close to zero. This assumption is not always satisfied as shown by the theoretical and numerical analysis in Zou and Tsung (2008). All of these methods are only applicable for the multistage processes with numerical data. However, in many modern manufacturing and service environments, due to some restrictions, for instance the time, cost or intrinsic character of the variables, quantitative observations cannot be measured directly or collected promptly for on-line monitoring use. Instead, only qualitative or categorical observations can be obtained. Shang, Tsung, and Zou (2013) described some motivating examples from manufacturing and service processes and proposed the monitoring and diagnosing schemes for the multistage process with binary data, but their diagnostic method can only deal with the single-fault case as well. Designing accurate and convenient diagnostic procedures for multi-fault isolation in binary-multistage processes is challenging but particularly useful because they will help business managers and engineers identify and eliminate root causes quickly and accurately so that quality and productivity can be improved. Therefore, this paper focuses on how to identify multiple faulty stages occurred in the multistage process with binary data.

Some recent research integrated the variable selection method with the multivariate SPC (cf. Jiang, Wang, & Tsung, 2012; Li, Wang, & Yeh, 2013; Maboudou-Tchao & Agboto, 2013; Maboudou-Tchao & Diawara, 2013; Yeh, Li, & Wang, 2012; Zou & Qiu, 2009; Zou, Jiang, & Tsung, 2011). For example, Zou and Qiu (2009) proposed multivariate monitoring and diagnosing schemes based on the least absolute shrinkage and selection operator (LASSO). Zou et al. (2011) developed a unified SPC diagnosis framework based on LASSO and demonstrated that it can be used for various multivariate SPC problems. Motivated by their methods, we develop a LASSO-based diagnosis scheme which can identify multiple faulty stages for multistage processes with binary data. We follow the settings in Sullivan et al. (2007) and Zou et al. (2011) and focus on the diagnostic process under the assumption that other MSPC methods have been used to detect and estimate a change point *a priori*. Assuming that the estimation of change point is sufficiently accurate, our objective is to determine the parameters that are responsible for the change. An implicit but important assumption we make here is that, in a high-dimensional multistage process, the probability that all stages shift simultaneously is rather low. It is believed that a fault is more likely to be caused by a hidden source, which is reflected in unknown changes of a small portion of stages.

The remainder of this paper is organized as follows: in Section 2 the modeling of multistage processes with binary data is briefly introduced, and the new proposed diagnosis scheme is elaborated. Following that, the performance of the proposed scheme is studied by simulation. One real example is used to illustrate the implementation of the proposed approach step by step in Section 4. Finally the concluding remarks and future studies are presented.

2. Methodology

2.1. Modeling of multistage processes with binomial data

Consider a common manufacturing process comprised of d stages. Without loss of generality, the stages are numbered in ascending order, such that if stage k precedes stage l, then $k < l$. Assume that a multistage process data set (or a batch) contains m identically and independently distributed (i.i.d) vector observations in the form

$$\mathbf{y}_j = \{y_{1j}, y_{2j}, \ldots, y_{dj}\},$$

where y_{ij} is a quality measurement observed from the jth product produced in the ith stage and is assumed to be drawn from a binomial (or Bernoulli) distribution, say $y_{ij} \sim \mathrm{BIN}(n_i, p_{ij})$. Here, p_{ij} represents the defect rate of the product sample at stage i. As shown in Shang et al. (2013), the two-level univariate binary state space model (BSSM) to link y_{ij} is the following:

$$\begin{aligned} &\mathrm{logit}(p_{ij}) = \alpha_i + c_i x_{ij}, \\ &x_{ij} = a_i x_{i-1j} + \omega_{ij}, \end{aligned} \quad \text{for} \quad i = 1, 2, \ldots, d, \tag{1}$$

where a_i and c_i are univariate parameters. The generalization to the multivariate case is trivial but it involves more complicated matrix notation. Similar to the assumptions commonly used in the linear state-space model (Zou & Tsung, 2008), we assume the initial value, $x_{0j} \sim N(\alpha_0, \sigma_0^2)$, and the error, $\omega_{ij} \sim N(0, \sigma_{\omega_i}^2)$.

Now, for the purpose of diagnosis, we consider a BSSM-based change model. Suppose that faults may happen in some stages in the process. In light of (1), we consider the following faulty model:

$$\begin{aligned} &\mathrm{logit}(p_{ij}) = \alpha_i + c_i x_{ij}, \\ &x_{ij} = a_i x_{i-1j} + \omega_{ij} + I_{\{k \geqslant \tau\}}\delta_i, \end{aligned} \quad \text{for} \quad i = 1, 2, \ldots, d, \quad j = 1, \ldots, m, \tag{2}$$

where $I_{\{\cdot\}}$ is the indicator function and δ_i is an unknown mean shift magnitude occurring in stage i. Based on this model, the monitoring scheme has been discussed in Shang et al. (2013), so the major task for such a multistage process here is to diagnose in which stage the shift occurs, so that proper adjustments to the process can be made in a timely fashion.

2.2. Model transformation

Recall model (1) and the associated notation. Moreover, denote d-dimensional vectors, $\boldsymbol{v}_j = (v_{1j}, \ldots, v_{dj})^T$, and a matrix, $\boldsymbol{\Sigma} = (\rho_{ij}^2)_{d \times d}$. Firstly, we rewrite the model (1) as the following equivalent form by replacing x_{ij} with the corresponding transformation of x_{0j} plus errors,

$$\begin{aligned} &y_{ij} | p_{ij} \sim \mathrm{BIN}(n_i, p_{ij}) \\ &\mathrm{logit}(p_{ij}) = \mu_{0i} + v_{ij}, \quad \text{for} \quad i = 1, \ldots, d, \end{aligned} \tag{3}$$

where $\boldsymbol{v}_j \overset{\text{i.i.d}}{\sim} N_d(\mathbf{0}, \boldsymbol{\Sigma})$, and

$$\mu_{0i} = \alpha_i + c_i \prod_{t=1}^{i} a_t \alpha_0,$$

$$\rho_{ii}^2 = c_i^2 \left(\prod_{t=1}^{i} a_t^2 \sigma_0^2 + \sum_{l=1}^{i} \sigma_{\omega_l}^2 \prod_{t=l+1}^{i} a_t^2 \right),$$

$$\rho_{ir}^2 = \rho_{ri}^2 = c_i c_r \left(\prod_{t=1}^{i} a_t \prod_{t=1}^{r} a_t \sigma_0^2 + \sum_{l=1}^{i} \sigma_{\omega_l}^2 \prod_{t=l+1}^{i} a_t \prod_{t=l+1}^{r} a_t \right) \quad (i < r).$$

Similarly, the model (2) can be written as follows: for $i = 1, \ldots, d$, $j = 1, \ldots, m$,

$$\begin{aligned} &y_{ij} | p_{ij} \sim \mathrm{BIN}(n_i, p_{ij}) \\ &\mathrm{logit}(p_{ij}) = \mu_{1i} + v_{ij}, \quad \mu_{1i} = \mu_{0i} + c_i \left(\sum_{l=2}^{i} \prod_{t=l}^{i} a_t \delta_{l-1} + \delta_i \right). \end{aligned} \tag{4}$$

As discussed in Shang et al. (2013), the BSSM is essentially equivalent to a special case of the generalized linear mixed-effects model with multivariate responses, called multivariate GLMM (MGLMM)