A LASSO-Based Diagnostic Framework for Multivariate Statistical Process Control

Changliang Zou[1], Wei Jiang[2], and Fugee Tsung[2]

[1] *LPMC and Department of Statistics, School of Mathematical Sciences, Nankai University, Tianjin, China*

[2] *Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*

Abstract

In monitoring complex systems, apart from quick detection of abnormal changes of system performance and key parameters, accurate fault diagnosis of responsible factors has become increasingly critical in a variety of applications that involve rich process data. Conventional statistical process control (SPC) methods, such as interpretation and decomposition of Hotelling's T^2-type statistic, suffer from such high-dimensional problems because they are often computationally expensive. In this paper, we frame fault isolation as a two-sample variable selection problem to provide a unified diagnosis framework based on Bayesian information criterion (BIC). We propose 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 with a similar computational effort as least-squares regressions. More importantly, the proposed method does not require making any extra tests that are necessary in existing diagnosis methods. Under some mild conditions, the diagnostic consistency of the proposed method is established. Finally, we present several specific SPC examples, including multistage process control and profile monitoring, to demonstrate the effectiveness of our method.

Keywords: BIC; Consistency; Fault Isolation; High-Dimensional; Least-Squares Approximation; Variable Selection

1 Introduction

In modern manufacturing and service industries, one frequently monitors several quality characteristics of a process simultaneously. Recently a great deal of research interest has arisen in *multivariate statistical process control* (MSPC) due to the availability of huge amount of data in such processes. For instance, under the intensive competitions in business and industry, besides the mean levels of process characteristics, monitoring and diagnosis of covariance matrix and even higher-order moments for a multivariate process have received more and more attention (see Huwang et al. 2007 and the references therein). Instead of monitoring a set of quality characteristics, profile monitoring focuses on the relationship between the response and explanatory variables assuming a univariate or multivariate multiple linear regression model (Woodall et al. 2004). In multistage processes involving serial operations, both downstream and upstream stages have to be monitored to detect departures from the prescribed relationships among subsequent stages (Shi 2007). In these applications, the data structure can often be represented by some parametric models which contain more than one parameter and the above SPC problems can all be characterized as MSPC in a certain sense. One of the major challenges in these applications is that the models are typically complex in nature, and in some applications the parameter dimension may be very high and results in the "curse of dimensionality" problem.

The fundamental tasks of MSPC include determining whether the process has changed, identifying when a detected shift has occurred, and isolating the shifted components or factors that are responsible for the change. The focus of this paper is the last one, which is also called *diagnosis* or *fault identification* in the MSPC literature (Mason and Young 2002). In monitoring complex systems, apart from quick detection of abnormal changes in system performance and key parameters, accurate fault diagnosis of responsible factors has become increasingly critical in a variety of applications that involve rich process data (Sullivan et al. 2007). A diagnostic aid to isolate the type of parameter changes will help

business managers and engineers identify and eliminate root causes of a problem quickly and accurately so that quality and productivity can be improved. Traditionally, statistical methods for accomplishing this task are usually based on interpretation and decomposition of Hotelling's T^2-type statistic, which essentially captures the relationships among different process parameters, e.g., Mason et al. (1995; 1997) and Li et al. (2008). Various step-down test procedures are also discussed based on certain *ad hoc* rules or assumptions, e.g., Hawkins (1991), Mason et al. (2001), Maravelakis et al. (2002), Sullivan et al. (2007), and Zhu and Jiang (2009). These conventional approaches are intuitively sound, but have the following practical shortcomings: (1) The decomposition of the T^2 statistic considers $p!$ different decompositions of the T^2 statistic (assuming the data dimension is p). Even the standard step-down testing procedures still require computing C_k^p test statistics of quadratic-form in the k-th step (Sullivan et al. 2007), which may be very inefficient when p is large; (2) More importantly, certain parameters in these approaches (e.g., the threshold values or significance levels) have considerable effects on the diagnostic ability, but are generally difficult to determine *a priori* in practice. See Sullivan et al. (2007) for a related discussion.

Recently, by applying variable selection methods in MSPC monitoring applications, Wang and Jiang (2009) and Zou and Qiu (2009) independently proposed new multivariate process monitoring and diagnosis schemes. Both of these schemes consider a penalized likelihood function approach based on the conventional multi-normality assumption. By taking advantage of recent developments in the L_1-penalized regression, the method in Zou and Qiu (2009) enjoys more computational efficiency in implementation. While their method primarily focuses on the monitoring task, our interest is to develop a unified framework for the diagnosis problems in many MSPC applications. We follow the settings in Sullivan et al. (2007) 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. In such circumstances, we are faced with a two-sample problem of change estimation, that is, the change-point partitions the observations into two subsets with different values for (some of) the parameters. An implicit but important assumption we make here is that, in a high-dimensional process, the probability that all parameters 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 one or a small set of parameters (see Wang and Jiang 2009, Zou and Qiu 2009).

We propose a unified treatment of parametric fault diagnosis based on two-sample Bayesian information criterion (BIC) to assist in the isolation of variables responsible for the signal. Although the proposed BIC framework is heuristic, it roots in the likelihood paradigm and does not rely on any *a priori* domain knowledge of potential shift parameters as Jiang and Tsui (2008) assumed or arbitrary sensitivity thresholds for testing the significance levels of parameter changes. Under the sparsity assumption of parameter changes, we further combine BIC with a penalized technique to facilitate the fault tracking process and suggest a practical LASSO-based (Tibshirani 1996) diagnostic procedure. Among other good properties, Fan and Li (2001) have shown that, with the proper choice of the penalty functions and regularization parameters, the estimator based on penalized techniques would perform asymptotically as well as if the correct model was known, which is referred to as the *oracle property* in the literature. Given the oracle property of the LASSO method (Zou 2006) and its connection with LARS algorithm (Efron et al. 2004), the diagnostic result can be obtained easily and quickly with a similar computational effort to least-squares regressions in implementation. The combination of BIC and the LASSO-based algorithm frees our method from having to make any extra statistical tests that are necessary in existing diagnostic methods. Under mild conditions, we establish the diagnostic consistency of the proposed method. Further numerical examples also demonstrate the effectiveness of the proposed method in various applications.

The proposed method can be applied to retrospective analysis of a historical data set (Phase I) as Sullivan and Woodall (2000) discussed or to the post-signal diagnostic analysis following prospective on-line monitoring (Phase II). However, it should be emphasized that in the prospective analysis, including self-starting or change-point methods (Zamba and Hawkins 2006) that lack a formal Phase I analysis, a major objective is the quick detection of a change, and if that objective is met then there may be only a few observations from the shifted models (the process is always stopped after a signal). So, the diagnostic results would be not accurate. If the objective of diagnosis is as important as detection, one may

consider to balance the performance of detection and diagnosis by either increase the in-control ARL or allowing the process operate for a while after a signal and collect more data from the shifted model. Of course, there should be a tradeoff between the cost of production in presence of faults and the gain of fault diagnosis.

The rest of the paper is organized as follows. Section 2 provides a heuristic derivation of BIC model selections in fault diagnosis and isolation. Section 3 develops the LASSO-based diagnosis procedure and derives its consistency property. Section 4 presents several industrial examples including monitoring mean and variance-covariance simultaneously, profile monitoring, and multistage process control. Statistical performance of the proposed diagnostic procedure is discussed by comparing with other existing diagnosis methods. In Section 5, we demonstrate the method using a real-data example from a white wine production process. Section 6 concludes the paper by summarizing our contributions and suggesting some future research issues. The necessary derivations and proofs of the theoretical results are detailed in the Appendix.

2 BIC for MSPC Fault Diagnosis

Zou and Qiu (2009) propose the use of LASSO-based variable selection techniques for MSPC monitoring when the data dimension is high. In this section, we first frame fault diagnosis as a model selection and estimation problem using the likelihood criteria. Unlike Zou and Qiu (2009), we do not assume that the baseline model is known *a priori*, but that it can be estimated from a sample of observations as discussed below. This allows our method to be applicable in a variety of SPC applications, including those discussed in Section 4.

2.1 Fault diagnosis and variable selection

Let $\mathbf{Z}_1 = \{\mathbf{z}_{11}, \ldots, \mathbf{z}_{1n_1}\}$ and $\mathbf{Z}_2 = \{\mathbf{z}_{21}, \ldots, \mathbf{z}_{2n_2}\}$ be two sets of independent random observations of size n_1 and n_2 respectively before and after a parameter change. Assume that $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1d})^T$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2d})^T$ are d-dimensional parameter vectors of interest for the two sets of observations, $\mathbf{z}_{1j} \sim f(\cdot|\boldsymbol{\beta}_1)$ and $\mathbf{z}_{2j} \sim f(\cdot|\boldsymbol{\beta}_2)$. Let $L_1(\boldsymbol{\beta}_1)$ and $L_2(\boldsymbol{\beta}_2)$

be the plausible loss functions for $\mathbf{Z}_1$ and $\mathbf{Z}_2$ respectively, whose global minimizers $\tilde{\boldsymbol{\beta}}_i$ are natural estimates of $\boldsymbol{\beta}_i$ for $i = 1, 2$.

For example, we are frequently concerned with a multivariate location problem in MSPC, that is, $\mathbf{z}_{ij} = \mathbf{x}_{ij}$ is multivariate normally distributed with $\mathbf{x}_{ij} \sim N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $\boldsymbol{\beta}_i = \boldsymbol{\mu}_i$ $(i = 1, 2)$ are the mean vectors of the in-control and out-of-control states. Then $L_i(\boldsymbol{\beta}_i)$ can be chosen as the $-2\times$log-likelihood function by ignoring some constants with respect to (w.r.t.) $\boldsymbol{\mu}_i$,

$$n_i(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i),$$

and $\tilde{\boldsymbol{\beta}}_i = \bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$. Consider another regression example. Assume we have n_i observations $\{(y_{ij}, \mathbf{x}_{ij})\}_{j=1}^{n_i}$ for each set $i = 1, 2$ collected from the model,

$$y_{ij} = \mathbf{x}_{ij}^T \mathbf{b}_i + \epsilon_{ij},$$

where y_{ij}'s are the response observations, $\mathbf{x}_{ij}$'s are the p-dimensional explanatory variables and ϵ_{ij}'s are the random errors. We are concerned about the difference between $\mathbf{b}_1$ and $\mathbf{b}_2$. In this case, $\mathbf{z}_{ij} = (y_{ij}, \mathbf{x}_{ij})$, $\boldsymbol{\beta}_i = \mathbf{b}_i$ and $d = p$. The $L_i(\cdot)$ can be chosen as the conventional least-squares loss function (equivalent to likelihood under normal error distributions) or some other robust regression loss functions, such as least absolute deviation or rank-based loss function (c.f., Hettmansperger and McKean 1998). We will discuss this type of problem later in Section 4.3.

Suppose that $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 + \boldsymbol{\delta}_n$. In many applications, it is rare that all variables shift at the same time, and the number of simultaneously changed variables is usually rather small as outlined by Wang and Jiang (2009). Hence, we also assume here that some components of the vector $\boldsymbol{\delta}_n = (\delta_{n1}, \ldots, \delta_{nd})$ are zero. Note that we use subscript "n" in $\boldsymbol{\delta}_n$ to allow the difference between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ to approach zero with a certain rate as the sample sizes become large. In other words, $\boldsymbol{\delta}_n$ is not necessarily a fixed change magnitude in an asymptotic viewpoint (see Section 3 and Assumption 2 for details). The MSPC fault diagnosis is interested in determining which components, δ_{nk}, are not zero. A natural way to accomplish this task is to obtain an appropriate estimate of $\boldsymbol{\delta}$ and to find out which components are non-zero. A simple estimation can be achieved by minimizing the joint loss function

$$L(\boldsymbol{\beta}_1, \boldsymbol{\delta}) = L_1(\boldsymbol{\beta}_1) + L_2(\boldsymbol{\beta}_1 + \boldsymbol{\delta}), \tag{1}$$

in which $\boldsymbol{\beta}_1$ can be regarded as a nuisance parameter that needs to be profiled out in the estimation procedure. However, this estimation suffers from inadequacy because its components usually take non-zero values, although a large portion of the values are quite close to zero.

In high-dimensional cases, we often assume that only a few components are non-zero, i.e., the so-called sparsity characteristic (Zou and Qiu 2009). Thus fault diagnosis is essentially analogous to the variable or model selection problem, that is to say, one wishes to select those parameters (δ_{nk}'s) that deviate significantly from zero. To be more specific, consider $\{1, \ldots, d\}$ as the full set of the indices of the parameters. Let s be a candidate subset of $\{1, \ldots, d\}$ which contains all the indices corresponding to the parameters that have changed (we assumed). The true fault isolation model is indexed by s_T which contains all the indices corresponding to the parameters that have really changed. Our objective is to select an optimal s under some specified criteria for fault isolation so that s could be as close to s_T as possible. In the literature, it is well demonstrated that Akaike information criterion (AIC) tends to select the model with the optimal predication performance, while Schwarz's Bayesian information criterion (BIC; Schwarz 1978) tends to identify the true sparse model well if the true model is included in the candidate set (Yang 2005). As we want to identify the non-zero components in $\boldsymbol{\delta}$ rather than obtaining an estimate, BIC is more relevant and appealing here. We will provide a heuristic derivation of BIC in our two-sample diagnosis problem. The following discussions are based on the assumptions presented in Appendix A.

2.2 The diagnostic aid with BIC

In the Bayesian framework, model comparison is based on *posterior* probabilities. Consider a candidate model $s \in \mathcal{S}$ where $\mathcal{S}$ is the model space under consideration. Here we use the generic notation "s" to denote a model or a subset for simplicity, which should not cause any confusion. When used for a model, it implies a model with the joint loss function (1) in which the indices of non-zero parameters in $\boldsymbol{\delta}$ all belong to the candidate subset s as defined above. Without any prior information, it is typically assumed that $\mathcal{S}$ is the full model space, containing totally $2^d - 1$ candidate models and $\mathcal{S} = \bigcup_{j=1}^{d} \mathcal{S}_j$, where $\mathcal{S}_j$ is the collection of all models with j non-zero components in $\boldsymbol{\delta}$. If some prior knowledge is available, then we could

conveniently incorporate it into the proposed framework by re-defining $\mathcal{S}$ as some collection of the full model space.

Assume that model s has *a priori* probability $\pi(s)$, and the prior density of its parameter α_s is $\pi(\alpha|s)$. Then the posterior probability of model s given data D satisfies

$$p(s|D) \propto \pi(s) \int p(D|\alpha_s, s)\pi(\alpha_s|s)d\alpha_s.$$

Under the Bayes paradigm, a model s^* that maximizes the posterior probability is selected, say $s^* = \arg\max_{s\in\mathcal{S}} \pi(s) \int p(D|\alpha_s, s)\pi(\alpha_s|s)d\alpha_s$. In practice, if we do not have any prior knowledge, an implicit underlying assumption is typically that the candidate models are equally likely so that $\pi(s)$ is constant over $\mathcal{S}$. Consequently, model assessment mainly depends on the integral term $\int p(D|\alpha_s, s)\pi(\alpha_s|s)d\alpha_s$ which is usually referred to as the marginal likelihood for model s.

The classical Schwarz's BIC is an approximation to the logarithm of the marginal likelihood, and there is a similar heuristic derivation for the two-sample BIC in the present diagnosis problem. We consider a pseudo-likelihood

$$L(D|\boldsymbol{\delta}_s, s) \propto \exp\{-L(\boldsymbol{\beta}_1, \boldsymbol{\delta}_s)/2\}, \tag{2}$$

where all the functions and parameters with subscript "s" are the analogs corresponding to model s. The main motivation for using Eq.(2) as a pseudo-likelihood is two-fold: on one hand, minimizing $L(\boldsymbol{\beta}_1, \boldsymbol{\delta}_s)$ gives the estimate of $\boldsymbol{\delta}$ which works by maximizing a log-likelihood function; on the other hand, $L(\boldsymbol{\beta}_1, \boldsymbol{\delta}_s)/2$ happens to be the log-likelihood as it ignores some constants w.r.t. $\boldsymbol{\delta}_s$, under the multi-normality assumption, which is common in many industrial applications.

However, it is not convenient to work with $L(\boldsymbol{\beta}_1, \boldsymbol{\delta})$ in many applications. Through a quadratic approximation in (A.2) as illustrated in Appendix B, minimizing $L(\boldsymbol{\beta}_1, \boldsymbol{\delta})$ w.r.t. $\boldsymbol{\delta}$ is asymptotically equivalent to minimizing

$$g(\boldsymbol{\delta}) = (\tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\delta})^T (\ddot{L}_1^{-1} + \ddot{L}_2^{-1})^{-1} (\tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\delta}), \tag{3}$$

where $\ddot{L}_i$ is the second-order derivative of the loss function $L_i(\boldsymbol{\beta}_i)$ w.r.t. $\boldsymbol{\beta}_i$. Very often the quantity $\ddot{L}_i^{-1}(\tilde{\boldsymbol{\beta}}_i)$ is closely related to the asymptotic covariance of $\tilde{\boldsymbol{\beta}}_i$ (denoted as $n_i^{-1}\boldsymbol{\Omega}_i$).

This further motivates us to consider the least-squares function

$$g(\boldsymbol{\delta}) = (\tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\delta})^T (n_1^{-1}\widehat{\boldsymbol{\Omega}}_1 + n_2^{-1}\widehat{\boldsymbol{\Omega}}_2)^{-1} (\tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\delta}), \tag{4}$$

as a simple approximation to the original loss function in Eq.(1). Here $\widehat{\boldsymbol{\Omega}}_i$ is an appropriate estimate of $\boldsymbol{\Omega}_i$. Naturally, we may use $\widehat{\boldsymbol{\Omega}}_i = n_i \ddot{L}_i^{-1}(\tilde{\boldsymbol{\beta}}_i)$ and Eq.(4) reduces to Eq.(3). Generally, Eq.(4) is a two-sample version of the least-squares approximation in Wang and Leng (2007). From Eq.(4), the integral pseudo-likelihood can be written as

$$IL(D|\boldsymbol{\delta}_s, s) \propto \int \exp\{-g(\boldsymbol{\delta}_s)/2\} \pi(\boldsymbol{\delta}_s|s) d\boldsymbol{\delta}_s. \tag{5}$$

Next, we consider the approximation of $IL(D|\boldsymbol{\delta}_s, s)$ using the Laplace method (Tierney and Kadane 1986). As we know, the basic idea of Laplace approximation is that in large samples, the integral is largely determined by the value of the integrand in a region close to $\widehat{\boldsymbol{\delta}}_\pi$, the value of $\boldsymbol{\delta}_s$ that maximizes $\tilde{g}(\boldsymbol{\delta}_s) = -g(\boldsymbol{\delta}_s) + \log(\pi(\boldsymbol{\delta}_s|s))$. As mentioned before, the prior used for model s is typically the unit information prior (Raftery 1996). Hence, in such cases, we have $\widehat{\boldsymbol{\delta}}_\pi \approx \tilde{\boldsymbol{\delta}}_s$, the minimizer of (4) under model s. As a result, a second-order Taylor expansion leads to

$$g(\boldsymbol{\delta}_s) \approx g(\tilde{\boldsymbol{\delta}}_s) + (\boldsymbol{\delta}_s - \tilde{\boldsymbol{\delta}}_s)^T (n_1^{-1}\widehat{\boldsymbol{\Omega}}_{1s} + n_2^{-1}\widehat{\boldsymbol{\Omega}}_{2s})^{-1} (\boldsymbol{\delta}_s - \tilde{\boldsymbol{\delta}}_s). \tag{6}$$

Now, by applying Laplace approximation, we obtain

$$\begin{aligned}
&\int \exp\{-g(\boldsymbol{\delta}_s)/2\} \pi(\boldsymbol{\delta}_s|s) d\boldsymbol{\delta}_s \\
&\qquad \approx \exp\{-g(\tilde{\boldsymbol{\delta}}_s)/2 + \log(\pi(\boldsymbol{\delta}_s|s))\} \int \exp\{-(\boldsymbol{\delta}_s - \tilde{\boldsymbol{\delta}}_s)^T \tilde{\boldsymbol{\Lambda}}_s (\boldsymbol{\delta}_s - \tilde{\boldsymbol{\delta}}_s)/2\} d\boldsymbol{\delta}_s \\
&\qquad = \exp\{-g(\tilde{\boldsymbol{\delta}}_s)/2 + \log(\pi(\boldsymbol{\delta}_s|s))\} (2\pi)^{d_s} |\tilde{\boldsymbol{\Lambda}}_s|^{-1/2},
\end{aligned}$$

where $\tilde{\boldsymbol{\Lambda}}_s = (n_1^{-1}\widehat{\boldsymbol{\Omega}}_{1s} + n_2^{-1}\widehat{\boldsymbol{\Omega}}_{2s})^{-1}$ and d_s is the dimension of the model s (say, the number of elements in the subset s). By Assumption 3 in Appendix A, we have $\tilde{\boldsymbol{\Lambda}}_s \xrightarrow{p} \boldsymbol{\Lambda}_s = (n_1^{-1}\boldsymbol{\Omega}_{1s} + n_2^{-1}\boldsymbol{\Omega}_{2s})^{-1}$. Following from the well known Weyl-Theorem and Sturm-Theorem (c.f., Theorems 4.4.5 and 4.4.14 on pages 118–120 in Marcus and Minc 1992), we have

$$[(n_1^{-1} + n_2^{-1})^{-1} \min\{\lambda_{1d}, \lambda_{2d}\}]^{d_s} \leq |\boldsymbol{\Lambda}_s| \leq [(n_1^{-1} + n_2^{-1})^{-1} \max\{\lambda_{11}, \lambda_{21}\}]^{d_s}, \tag{7}$$

where $\lambda_{i1} \geq \ldots \geq \lambda_{id}$ are the eigenvalues of $\boldsymbol{\Omega}_i$ for $i = 1, 2$. Hence, we can immediately obtain

$$\begin{aligned}
&\log\Big[\int \exp\{-g(\boldsymbol{\delta}_s)/2\}\pi(\boldsymbol{\delta}_s|s)d\boldsymbol{\delta}_s\Big] \\
&\qquad \approx -g(\tilde{\boldsymbol{\delta}}_s)/2 + \log(\pi(\boldsymbol{\delta}_s|s)) + d_s\log(2\pi) - (1/2)\log|\tilde{\boldsymbol{\Lambda}}_s| \\
&\qquad = -g(\tilde{\boldsymbol{\delta}}_s)/2 - \frac{1}{2}d_s \cdot \log\frac{n_1 n_2}{n_1+n_2} + O(1).
\end{aligned}$$

If we ignore the terms of $O(1)$, finding the model that gives the highest posterior probability based on the pseudo-likelihood (2) leads to minimizing BIC, which is defined by

$$\mathrm{BIC}_s = g(\tilde{\boldsymbol{\delta}}_s) + d_s \cdot \log\frac{n_1 n_2}{n_1+n_2}. \tag{8}$$

In cases where n_1 is sufficiently large, i.e., the baseline model can be estimated accurately as assumed in most Phase II applications, the second term can be reduced to $d_s \log n_2$, which strongly depends on n_2.

Since the ordinary BIC is somewhat liberal for model selection when the model space is large with a moderate sample, it tends to select a model with many spurious parameters (Broman and Speed 2002). This situation often occurs in MSPC diagnosis problems, especially in today's service industries where the datasets always contain many variables. For instance, when diagnosing both mean vector and variance-covariance matrix of a p-dimensional vector, we would face a situation with $d = [p(p+3)/2]$. For $p = 8$, d will be 44 which is relatively large in the usual circumstances where the sample sizes (n_1, n_2) are several hundreds at most. By re-examining the Bayesian paradigm for model selection, Chen and Chen (2008) proposed an extended family of BIC (EBIC), which takes into account both the number of unknown parameters and the complexity of the model space. It is demonstrated that EBIC incurs a small loss in the positive selection rate but tightly controls the false discovery rate. To be specific, in the present problem, BIC in Eq.(8) can be similarly extended by the following EBIC,

$$\mathrm{EBIC}_s = g(\tilde{\boldsymbol{\delta}}_s) + \left[\log\frac{n_1 n_2}{n_1+n_2} + 2\log d\right] d_s. \tag{9}$$

That is, using an additional term "$2\log d$" gives a further "correction" to the ordinary BIC. Due to its attractive properties (Chen and Chen 2008) and appealing numerical performance

in our problems, we recommended using this EBIC which will be further combined with a LASSO procedure in the next section. We will also show the asymptotic consistency of the EBIC selection.

3 A LASSO-Based Diagnosis Procedure

BIC and EBIC are both consistent in the sense that they select the true model with probability approaching one when sample size increases if such a true model is in the class of candidate models. In practice, when d is large, we cannot afford to calculate the EBIC values (9) for all possible s. Instead, we prefer to combine this criterion with some penalized techniques (see Zou and Qiu 2009 and the references therein). Motivated by Eq.(4), we will consider the penalized loss function

$$PL(\boldsymbol{\delta};\boldsymbol{\theta}) = g(\boldsymbol{\delta}) + \sum_{k=1}^{d} h_{\theta_k}(|\delta_k|), \tag{10}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$ are the penalty parameters (also called the regularization parameters), and $h_{\theta_k}(\cdot)$ are the penalty functions. If the adaptive LASSO (ALASSO) penalty function (cf., Zou 2006) is used, then the loss function becomes

$$AL(\boldsymbol{\delta};\boldsymbol{\theta}) = g(\boldsymbol{\delta}) + \sum_{k=1}^{d} \theta_k|\delta_k|. \tag{11}$$

Given $\boldsymbol{\theta}$, the minimizer $\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}} = \arg\min AL(\boldsymbol{\delta};\boldsymbol{\theta})$ naturally defines a candidate model $s_{\boldsymbol{\theta}} = \{i : \widehat{\delta}_{\boldsymbol{\theta}i} \neq 0\}$. Now, substituting the estimator $\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}}$ into (9) leads to

$$\text{EBIC}_{\boldsymbol{\theta}} = g(\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}}) + \left[\log\frac{n_1 n_2}{n_1+n_2} + 2\log d\right] \cdot d_{\boldsymbol{\theta}}, \tag{12}$$

where $d_{\boldsymbol{\theta}}$ denotes the number of non-zero elements of $\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}}$.

The following result establishes the diagnostic consistency of LASSO-based EBIC in Eq.(12) under some mild conditions. Although this result requires the sample size approaches infinity, which can never be met in reality, we believe it is one of the basic requirements for any diagnosis methods since a quality engineer may not trust and use a practical diagnosis

method if it can't guarantee to find the correct subset of shifted parameters even when sample size is very large. The consistency result essentially warrants the users the confidence of using the diagnosis method when more information are collected.

Theorem 1 *Let* $\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} \mathrm{EBIC}_{\boldsymbol{\theta}}$. *If Assumptions 1-3 in Appendix A hold, as* $\min\{n_1, n_2\} \to \infty$, $\Pr(s_{\widehat{\boldsymbol{\theta}}_n} = s_T) \to 1$.

The proof of this theorem is given in Appendix C. The main idea of the proof is similar to Nishii (1984), i.e., comparing the values between the considered model and the true model in two different cases according to whether the model is underfitted or overfitted. In the post-signal diagnosis of Phase II monitoring, it seems that the requirement of large sample sizes is not likely fulfilled. In asymptotic studies of average run length (ARL) in the literature (e.g., Han and Tsung 2006), it is typically assumed that the control limit goes to infinity and both of the asymptotic expressions of the IC and OC ARLs approach to infinity. The consistency result given by Theorem 1 can therefore be regarded as a sequel to asymptotic analysis of control charts and guarantees that our diagnosis result is correct following the fault detection from a control chart in the asymptotic point of view. In addition, as indicated by Assumption 2, the diagnostic consistency requires $\sqrt{n/\log n}\liminf_{n\to\infty}(\min_{i\in s_T}|\delta_{nj}|) \to \infty$. Roughly speaking, as long as the smallest non-zero element in $\boldsymbol{\delta}_n$ is of lower order than $(n/\log n)^{-1/2}$, the procedure could give a correct identification in an asymptotic view. Such an asymptotic result also sheds lights on the statistical property of the identified subset $s_{\widehat{\boldsymbol{\theta}}_n}$ in finite-sample cases. That is, it reflects the relative magnitude of $\boldsymbol{\delta}_n$ to the sample size with which we can obtain a reasonable diagnosis result in small-sample cases (although this may not be easily quantified in practice).

Following Zou's (2006) recommendation, we will set $\theta_k = \theta|\tilde{\delta}_k|^{-r}$, where $r > 0$ is some specified constant, such as 0.5 or 1, and $\tilde{\boldsymbol{\delta}} = \tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1 = (\tilde{\delta}_1, \ldots, \tilde{\delta}_d)^T$. By Assumption 2 in Appendix A, $\tilde{\delta}_k$ is $\sqrt{n}$-consistent. Thus, it is easy to verify that such tuning parameter definition satisfies the conditions on a_n and b_n needed by Lemma 1 in the Appendix as long as $\sqrt{n}\theta \to 0$ and $n^{(1+r)/2}\theta \to \infty$. Hence, Theorem 1 is still valid after applying this specific penalty setting. Such a strategy not only effectively transforms the original d-dimensional tuning parameter selection problem into a univariate one, but also greatly facilitates the

search procedure based on the following generalization of Theorem 3 in Zou et al. (2007). We rewrite the ALASSO-type penalized likelihood (11) with $\theta_k = \theta|\tilde{\delta}_k|^{-r}$ as

$$PL(\boldsymbol{\alpha};\theta) = (\tilde{\boldsymbol{\delta}} - \boldsymbol{\Delta}\boldsymbol{\alpha})^T \tilde{\boldsymbol{\Lambda}} (\tilde{\boldsymbol{\delta}} - \boldsymbol{\Delta}\boldsymbol{\alpha}) + \theta \sum_{i=1}^{d} |\alpha_i|, \tag{13}$$

where $\alpha_i = \delta_i/|\tilde{\delta}_i|^r$ and $\boldsymbol{\Delta} = \mathrm{diag}(|\tilde{\delta}_1|^r, \ldots, |\tilde{\delta}_d|^r)$. This is exactly a LASSO-type penalized likelihood function. According to Zou et al. (2007), there is a finite sequence

$$\tilde{\theta}_0 > \tilde{\theta}_1 > \ldots > \tilde{\theta}_K = 0, \tag{14}$$

such that (i) for all $\theta > \tilde{\theta}_0$, $\widehat{\boldsymbol{\alpha}}_\theta = 0$, and (ii) in the interval $(\tilde{\theta}_{m+1}, \tilde{\theta}_m)$, the active set $\mathfrak{B}(\theta) = \{j : \mathrm{sgn}[\alpha_{\theta j}] \neq 0\}$ and the sign vector $\boldsymbol{S}(\theta) = \{\mathrm{sgn}[\alpha_{\theta 1}], \ldots, \mathrm{sgn}[\alpha_{\theta d}]\}$ are unchanged with θ. These $\tilde{\theta}_m$'s are called transition points because the active set changes at each $\tilde{\theta}_m$. Let $\widehat{\theta}_n = \arg\min_\theta \mathrm{EBIC}_\theta$. By generalizing the proof of Theorem 3 in Zou et al. (2007), we can obtain results in the following proposition without much difficulty.

Proposition 1 *$\widehat{\boldsymbol{\delta}}_{\widehat{\theta}_n}$ is one of the LASSO solutions at transition points, i.e, $\widehat{\boldsymbol{\delta}}_{\widehat{\theta}_n} \in \{\widehat{\boldsymbol{\delta}}_{\tilde{\theta}_1}, \ldots, \widehat{\boldsymbol{\delta}}_{\tilde{\theta}_K}\}$.*

It is worth noting that this proposition avoids the use of some numerical search algorithms for finding the solution of the optimization problem, $\arg\min_\theta \mathrm{EBIC}_\theta$. With the help of this proposition and the LARS algorithm (see Efron et al. 2004), we can obtain the diagnostic result easily and quickly because the LARS algorithm produces these transition points with similar computational cost to a least-squares regression with d covariates.

To end this section, we summarize the detailed steps for implementing the proposed LASSO-based EBIC diagnosis method (called LEB procedure for abbreviation) as follows. A Fortran package and R interface for implementing the proposed procedure are available from the authors upon request.

LEB diagnosis procedure:

1. Specify the loss functions L_i and obtain the corresponding estimators $\tilde{\boldsymbol{\beta}}_i$ for $i = 1, 2$.

2. Find appropriate estimators of covariance matrices $\widehat{\boldsymbol{\Omega}}_i$ for $i = 1, 2$.

3. Use the LARS algorithm to solve (13) and obtain the K solutions of ALASSO at all the transition points.

4. Substitute these K solutions into (12) and find the one $\widehat{\boldsymbol{\delta}}_{\widehat{\theta}_n}$ whose EBIC value is the smallest. Then the corresponding diagnostic result is $s_{\widehat{\theta}_n} = \{i : \widehat{\delta}_{\theta_{ni}} \neq 0\}$.

4 Industrial Examples

In this section, we present three industrial examples in SPC practice to investigate the performance of the proposed fault diagnosis framework and compare it with some existing methods. Assume $\widehat{s}$ is a subset of $\{1, \ldots, d\}$ determined by a diagnosis method. To evaluate the statistical performance of a specific diagnosis procedure, we consider three accuracy measures. Correctness, which is represented as "C" in the following tables, present the relative frequencies of the cases when the diagnostic procedure identifies faulty parameters *fully correctly* (i.e., $\widehat{s} = s_T$). In addition, to represent the overall quality of a diagnosis procedure, we also consider an index - parameter selection score (PSS) - defined as,

$$E(\sum_{i=1}^{d} |I_{\{i \in s_T\}} - I_{\{i \in \widehat{s}\}}|),$$

where $\mathrm{I}_{\{\cdot\}}$ is the indicator function. This index provides certain indication of the precision of the diagnostic results, and the expectation is of course approximated by the average of many repetitions. So, for a given diagnostic procedure, it performs better in a given case if its value in column "C" is comparatively larger and its value in column "PSS" is comparatively smaller. The following simulation results are obtained from 10,000 replications.

4.1 Diagnosis in multivariate mean vector and covariance matrix

Fault diagnosis in multivariate location parameters may be one of the problems we are most commonly faced with in SPC practice (Mason and Young 2002). In this situation, $\mathbf{z}_{ij} = \mathbf{x}_{ij} \in \mathbb{R}^p$ $(i = 1, 2)$ are assumed to come from some distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. A standard choice is $\boldsymbol{\beta}_i = \boldsymbol{\mu}_i$. Naturally,

$$g(\boldsymbol{\delta}) = (\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1 - \boldsymbol{\delta})^T (n_1^{-1}\widehat{\boldsymbol{\Sigma}}_1 + n_2^{-1}\widehat{\boldsymbol{\Sigma}}_2)^{-1} (\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1 - \boldsymbol{\delta}),$$

where $\bar{\mathbf{x}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$ are the estimated mean vector and covariance matrix for the i-th dataset, respectively. Furthermore, diagnosis in both the mean vector and the covariance matrix (or equivalently the correlation matrix) is also of practical interest (e.g., Huwang et al. 2007; Sullivan et al. 2007). In this case, we are concerned with $\boldsymbol{\beta}_i = [\boldsymbol{\mu}_i^T, \mathrm{Sur}^T(\boldsymbol{\Sigma}_i)]^T$ and $d = p(p+3)/2$, where $\mathrm{Sur}(\boldsymbol{\Sigma})$ denotes the vector formed by stacking the upper triangular elements of $\boldsymbol{\Sigma}$. Certainly, we could set $\tilde{\boldsymbol{\beta}}_i = [\bar{\mathbf{x}}_i^T, \mathrm{Sur}^T(\widehat{\boldsymbol{\Sigma}}_i)]^T$. The estimate of the asymptotic covariance matrix of $\tilde{\boldsymbol{\beta}}_i$, $\widehat{\boldsymbol{\Omega}}_i$, can be found in Neudecker and Wesselman (1990). It should be worth noting that in this case the number of important parameter changes may differ depending on how to parameterize the model, say using the covariance matrix or alternatively using the correlations and standard deviations. If a shift involves only a single standard deviation but none of correlations, it would change several components if the covariance matrix is used. It is therefore advised that one should always choose the representation which is more likely to result in some physical interpretation so that the engineers could use the results directly. Of course, this requires the engineers to take the engineering knowledge about the specific problem into consideration in practical applications.

We now compare the proposed LEB procedure with the step-down procedure discussed before (Sullivan et al. 2007 and the references therein). The step-down procedure depends heavily on a pre-specified type I error probability, which is chosen to be 0.05, 0.01, or 0.002 here for comparison use. The number and variety of covariance matrices and change cases are too large to allow a comprehensive, all-encompassing comparison. Our goal is to show the effectiveness, robustness and sensitivity of LEB, and thus we only choose certain representative models for illustration. Specifically, the covariance matrix $\boldsymbol{\Sigma}_1 = (\sigma_{1(ij)})$ is chosen to be $\sigma_{1(ii)} = 1$ and $\sigma_{1(ij)} = 0.5^{|i-j|}$, for $i, j = 1, \ldots, p$. Three comparison scenarios are chosen: (i) $\mu_{2k} = \mu_{1k} + \delta$ for $k = 1, 2$; (ii) $\mu_{2k} = \mu_{1k} + \delta$ for $k = 1, 2$, and $\sigma_{2(11)} = \sigma_{1(11)} + \delta$; (iii) $\mu_{2k} = \mu_{1k} + \delta$ for $k = 1, 2$, $\sigma_{2(11)} = \sigma_{1(11)} + \delta$, and $\sigma_{2(23)} = \sigma_{1(23)} - \delta$. We set $\delta = 1$ in all the cases but vary n_i to obtain results corresponding to various *signal-to-noise ratios* in diagnosis. $\boldsymbol{\mu}_1$ is set to zero vector without loss of generality. The simulation results are summarized in Tables 1 and 2 where $p = 4$ and 6 are considered. In Table 1, we fixed the sample sizes as $n_1 = 2n_2$, whereas in Table 2 n_1 is fixed at 1,000 which may reflect the performance of post-signal diagnosis in Phase II, say the IC dataset (n_1) is fairly large.

Table 1: Diagnostic performance comparisons between the proposed LEB and the step-down procedures for MSPC in various change cases ($n_1 = 2n_2$).

scenarios	sample sizes	step-down procedure						LEB	
		$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.002$			
	(n_1, n_2)	$p = 4$							
		C	PSS	C	PSS	C	PSS	C	PSS
	(50,25)	0.16	1.81	0.33	1.20	0.36	1.07	0.37	1.04
(i)	(100,50)	0.20	1.54	0.44	0.83	0.58	0.57	0.58	0.61
	(200,100)	0.21	1.48	0.48	0.73	0.66	0.43	0.69	0.43
	(50,25)	0.17	1.80	0.25	1.44	0.20	1.52	0.24	1.42
(ii)	(100,50)	0.24	1.39	0.46	0.79	0.55	0.63	0.49	0.76
	(200,100)	0.27	1.27	0.52	0.64	0.69	0.38	0.67	0.46
	(50,25)	0.13	2.07	0.11	1.95	0.06	2.24	0.12	1.95
(iii)	(100,50)	0.24	1.40	0.39	0.96	0.36	0.97	0.37	0.98
	(200,100)	0.30	1.13	0.58	0.57	0.70	0.37	0.63	0.50
	(n_1, n_2)	$p = 6$							
		C	PSS	C	PSS	C	PSS	C	PSS
	(100,50)	0.04	3.33	0.22	1.67	0.41	1.03	0.51	0.85
(i)	(200,100)	0.05	3.05	0.25	1.43	0.50	0.75	0.66	0.51
	(500,250)	0.06	2.90	0.29	1.27	0.52	0.67	0.80	0.28
	(100,50)	0.04	3.08	0.23	1.57	0.39	1.09	0.38	1.07
(ii)	(200,100)	0.07	2.78	0.30	1.27	0.53	0.68	0.60	0.62
	(500.250)	0.07	2.68	0.32	1.18	0.57	0.58	0.77	0.31
	(100,50)	0.05	3.17	0.19	1.83	0.27	1.43	0.23	1.52
(iii)	(200,100)	0.06	2.80	0.31	1.28	0.51	0.76	0.52	0.78
	(500,250)	0.08	2.61	0.35	1.12	0.59	0.56	0.75	0.35

Note: The standard error of the frequency (π) in each entry, $\sqrt{\widehat{\pi}(1-\widehat{\pi})/10000}$, is typically less than 0.01.

Table 2: Diagnostic performance comparisons between the proposed LEB and the step-down procedures for MSPC in various change cases (n_1 is fixed at 1,000).

scenarios	sample sizes	step-down procedure						LEB	
		$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.002$			
	(n_1, n_2)	$p = 4$							
		C	PSS	C	PSS	C	PSS	C	PSS
	(1000,25)	0.11	2.39	0.25	1.67	0.34	1.35	0.36	1.26
(i)	(1000,50)	0.16	1.85	0.38	1.08	0.52	0.76	0.57	0.68
	(1000,100)	0.19	1.59	0.44	0.83	0.62	0.50	0.69	0.43
	(1000,25)	0.13	2.16	0.28	1.53	0.35	1.29	0.36	1.22
(ii)	(1000,50)	0.21	1.57	0.44	0.90	0.58	0.63	0.59	0.63
	(1000,100)	0.23	1.40	0.49	0.72	0.66	0.43	0.71	0.40
	(1000,25)	0.14	2.14	0.23	1.71	0.22	1.64	0.25	1.47
(iii)	(1000,50)	0.23	1.46	0.43	0.90	0.52	0.73	0.49	0.78
	(1000,100)	0.28	1.22	0.53	0.62	0.68	0.39	0.67	0.45
	(n_1, n_2)	$p = 6$							
		C	PSS	C	PSS	C	PSS	C	PSS
	(1000,50)	0.03	3.87	0.16	2.24	0.32	1.54	0.50	0.94
(i)	(1000,100)	0.04	3.23	0.23	1.58	0.46	0.87	0.70	0.47
	(1000,250)	0.08	2.60	0.34	1.12	0.61	0.54	0.79	0.28
	(1000,50)	0.03	3.67	0.18	2.07	0.36	1.38	0.48	0.96
(ii)	(1000,100)	0.06	2.97	0.26	1.46	0.49	0.81	0.68	0.48
	(1000,250)	0.06	2.73	0.31	1.21	0.55	0.61	0.81	0.26
	(1000,50)	0.04	3.45	0.20	2.00	0.36	1.41	0.40	1.14
(iii)	(1000,100)	0.05	2.86	0.29	1.36	0.53	0.75	0.63	0.60
	(1000,250)	0.08	2.60	0.34	1.12	0.61	0.54	0.79	0.28

As expected, the step-down procedure strongly depends on the choice of the type I error probability. On the other hand, the proposed LASSO-based approach has comparable diagnostic ability to that of the step-down procedure with the "optimal" choice of type I error probability. In many situations, especially when the dimension p gets larger ($p = 6$), the LASSO-based approach outperforms the step-down procedure by a considerably large margin. In this case, the step-down procedure fails to perform exact identifications well (in terms of "C") when the sample size is relatively large, and even hardly improve as the sample size becomes larger. For example, when $\alpha = 0.002$, the correctness probability remains close to $0.5 \sim 0.6$ even when sample size increases significantly from (100, 200) to (250, 500). In comparison, the LEB procedure tends to identify all relevant and irrelevant parameters correctly, especially when sample size gets larger. For example, for the fault case (iii) when $p = 6$, its correctness improves from 0.52 to 0.75 when sample size increases from (100, 200) to (250, 500). Of course, the performance of the step-down procedure can be improved by choosing other more appropriate type I error probabilities in a given case. However, such an *ad hoc* approach is not very convenient in practical applications because there is a lack of standard recommendations on how to choose this parameter. Similarly, when comparing the PSS values, the LASSO-based approach is comparable to the step-down procedure when $p = 4$ but significantly better when $p = 6$, especially when sample size is large. In addition, Tables 1 and 2 provide similar comparison information. As n_1 increase, the performance of LEB generally gets better, although the improvements are not very substantial. This is not surprising to us because from Section 2, we can see that the "actual" (roughly speaking) sample size in such a two-sample problem is $n_1 n_2/(n_1 + n_2)$ and thus the diagnostic performance of LEB depends mainly on the value of $\min(n_1, n_2)$ as long as n_1 or n_2 is large enough. Overall, after taking into account its computational advantage, we believe that the LEB approach provides a reasonable diagnosis tool for MSPC applications.

It is worth to point out that the LEWMA method proposed in Zou and Qiu (2009) is different from the LEB method here. The LEWMA method integrates the LASSO algorithm with a multivariate exponential weighted moving average charting scheme for Phase II multivariate process monitoring. While it mainly concerns with mean shift detection, the LEB method is designed for fault diagnosis which includes both mean and variance-covariance

changes. The above cases (ii) and (iii) show that the LEB method can not only identify the mean changes, but also the variance and covariance changes. The diagnosis performance not only depends on the sample sizes, but also the number of shifted parameters, especially when the sample sizes are small.

4.2 Diagnosis in multistage process control

As modern technologies become increasingly sophisticated, most manufacturing operations, such as semiconductor manufacturing and automotive body assembly, comprise multiple stages. Shi and Zhou (2009) provide an extensive review of the multistage process control problems with many industrial examples. In these systems, it is often desirable and necessary to design an effective diagnosis approach for isolating and identifying the sources of a change by linking the current stage signal to information about earlier stages in the serial process. Zhou et al. (2003) and many others discuss sensor allocations and fault diagnosability in multistage processes. Zou and Tsung (2008), Zou et al. (2008), and Li and Tsung (2009) also investigate multistage process monitoring and diagnosis problems in various settings.

Consider a common manufacturing process comprised of p stages. For the j-th product collected, a two-level linear state-space model generated from a practical application is usually used to describe the quality measurement to the k-th stage (Zou et al. 2008): for $k = 1, \ldots, p$, $j = 1, 2, \ldots, n_1 + n_2$,

$$y_{kj} = C_k x_{kj} + v_{kj}$$
$$x_{kj} = A_k x_{k-1j} + w_{kj} + \delta_k \mathrm{I}_{\{k \in s_T, j > n_1\}},$$

where $x_{0j} \sim N(a_0, \sigma_\varepsilon^2)$. The v_{kj} and w_{kj} are assumed to be independent from each other and $v_k \sim (\mu_k, \sigma_{v_k}^2)$, $w_k \sim (0, \sigma_{w_k}^2)$. The first level of the model involves the fitting of the quality measurement to the quality characteristic, and C_k is used to relate the unobservable process quality characteristic, x_k, to the observable quality measurement, y_k. The second level of the model involves modeling the transfer of the quality characteristic from the previous stage to the present stage, in which A_k denotes the transformation coefficient of the quality characteristic from stage $k-1$ to stage k. In multistage process applications, A_k and C_k are known constants (or matrices) that are usually derived or estimated from engineering

knowledge (see Jin and Shi 1999 for details). The unknown magnitudes, δ_k's, reflect the difference between two multistage samples. Typically most of δ_k's are zero.

By the above model assumption, we have $\mathbf{z}_{1j} = (y_{j1}, \ldots, y_{jp})^T$ for $j \leq n_1$ and $\mathbf{z}_{2j} = (y_{j1}, \ldots, y_{jp})^T$ for $j > n_1$. Assume $E(\mathbf{z}_{ij}) = \boldsymbol{\mu}_i$. It follows that $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + \boldsymbol{\Gamma}\boldsymbol{\delta}$, where

$$\boldsymbol{\Gamma} = \begin{pmatrix} C_1 & 0 & \ldots & 0 \\ C_2A_2 & C_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_p\prod_{i=2}^{p} A_i & \ldots & C_pA_p & C_p \end{pmatrix}.$$

Hence we can set $\boldsymbol{\beta}_1 = \boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_1$ and $\boldsymbol{\beta}_2 = \boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_2$. As a consequence,

$$\tilde{\boldsymbol{\beta}}_i = \boldsymbol{\Gamma}^{-1}\bar{\mathbf{z}}_i, \text{ and } \widehat{\boldsymbol{\Omega}}_i = \boldsymbol{\Gamma}^{-1}\widehat{\boldsymbol{\Sigma}}_i(\boldsymbol{\Gamma}^{-1})^T,$$

where $\bar{\mathbf{z}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$ are the estimated mean vector and covariance matrix for the i-th dataset, respectively.

We now compare the LEB procedure with some existing alternatives. Note that Zhou et al. (2003) mainly studied the diagnosability of linear-mixed multistage models and Zou and Tsung's (2008) diagnosis method is only applicable for single-fault cases. Li and Tsung (2009) proposed a fault isolation method which is able to identify multiple fault stages, and thus we use it as a benchmark for comparison. 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. Hereafter we refer to it as OSFE-FDR. In this subsection, we set $p = 20$, $(A_k, C_k) = (1.0, 1.0)$, $(A_k, C_k) = (1.2, 0.8)$ and $(A_k, C_k) = (0.8, 1.2)$, which have been investigated and are consistent with the numerical comparison settings in Zou and Tsung (2008). $\sigma^2_{v_k}, \sigma^2_{w_k}$ and σ^2_ε are all chosen as 0.5. Three types of faults are chosen for comparison: (i) $(\delta_1, \delta_2) = (1, -1)$; (ii) $(\delta_1, \delta_2, \delta_5) = (1, -1, 2)$; and (iii) $(\delta_1, \delta_2, \delta_5, \delta_{15}) = (1, -1, 2, -2)$. Note that the OSFE-FDR approach also depends on the pre-specification of type I error probability, which is fixed at 0.001 here for illustration. The simulation results are summarized in Table 2.

The LEB appears to be much more accurate in estimating the changed stages in most of the cases. The OSFE-FDR procedure performs reasonably well only in some cases of

Table 3: Diagnostic performance comparisons between the LEB and OSFE-FDR procedures for the multistage problem.

(A_k, C_k)	scenarios	C	PSS	C	PSS	C	PSS	C	PSS
		$(n_1, n_2) = (60, 40)$				$(n_1, n_2) = (50, 100)$			
		OSFE-FDR		LEB		OSFE-FDR		LEB	
	(i)	0.09	1.40	0.53	0.62	0.23	1.00	0.69	0.39
(1.0,1.0)	(ii)	0.04	2.23	0.49	0.74	0.07	2.04	0.68	0.42
	(iii)	0.02	2.89	0.53	0.69	0.02	2.95	0.69	0.41
	(i)	0.29	0.95	0.66	0.43	0.54	0.54	0.78	0.27
(0.8,1.2)	(ii)	0.11	1.86	0.65	0.47	0.17	1.66	0.77	0.28
	(iii)	0.06	2.42	0.66	0.44	0.07	2.40	0.77	0.28
	(i)	0.01	1.73	0.30	1.03	0.04	1.46	0.49	0.67
(1.2,0.8)	(ii)	0.01	2.39	0.28	1.21	0.02	2.33	0.45	0.85
	(iii)	0.00	3.20	0.30	1.15	0.01	3.30	0.48	0.79

scenario (i). Unlike our proposed method, OSFE-FDR has no theoretical result to assure its effectiveness. This can be clearly seen from the case of $(A_k, C_k) = (1.2, 0.8)$, in which the accuracy of OSFE-FDR hardly improves as the sample sizes become larger. This is not surprising because the effectiveness of OSFE-FDR relies on the implicit assumption that the OSFE's at the stages $i \in s_T$ significantly deviate away from zero but those at the stages $i \notin s_T$ are close to zero (Li and Tsung 2009). However, as shown in the theoretical and numerical analysis in Zou and Tsung (2008), this assumption is not always satisfied. That's why the accuracy of OSFE-FDR would be extremely poor in certain cases, although the type I error probability can be adjusted to alleviate the problem. On the other hand, the LEB procedure always has a higher diagnostic power than OSFE-FDR. Moreover, its power can be significantly improved when sample sizes become larger in all cases considered here. When $(A_k, C_k) = (1.2, 0.8)$, its diagnostic probability increases from around 0.30 to 0.50 when (n_1, n_2) changes from (60, 40) to (50, 100). Although this is a specific example and

not representative of any multistage systems, it illustrates that the LEB procedure is more accurate and also more robust than OSFE-FDR for diagnosing different types of faults in multistage systems.

4.3 Diagnosis in profile problems

In many applications of manufacturing and service industries, the quality of a process is often characterized by functional relationships between response variables and explanatory variables. The problems in monitoring and diagnosing the stability of such relationships are referred to as SPC for profile data. An extensive discussion of research problems on this topic is presented by Woodall et al. (2004). As Woodall (2007) pointed out, a systematic approach for profile diagnosis is most critical in many practical applications. However, it is challenging to isolate the type of profile parameter changes in a high-dimensional profile problem, especially for complicated general profiles. Among others, Zou et al. (2007) proposed to use parametric tests to identify the shifted parameters for general linear profiles.

To demonstrate the diverse applicability of the proposed LASSO-based framework, we will extend the linear profile model in Zou et al. (2007) to a more general case - multivariate linear profile model in which multiple response variables may be involved simultaneously. Assume we have n observations on q responses of $\mathbf{y} = (y_1, \ldots, y_q)^T$ and p explanatory variables $\mathbf{x} = (x_1, \ldots, x_p)^T$. The process observations are collected from the following general multivariate linear profile model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{15}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_q)$ is an $m \times q$ matrix, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^T$ is an $m \times p$ matrix, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_q)$ is a $p \times q$ coefficients matrix, $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_m)^T$ is the regression noise with $\mathbf{e}_k = (\varepsilon_{k1}, \ldots, \varepsilon_{kq})^T$ and $\mathbf{e}$'s are independently sampled from $(\mathbf{0}, \mathbf{\Sigma})$ with the diagonal components of $\mathbf{\Sigma}$ being $\mathbf{s} = (\sigma^2_{(11)}, \ldots, \sigma^2_{(qq)})^T$. Now assume there is a change point in the profile relationship (15) and we collect $\mathbf{z}_{ij} = (\mathbf{Y}_{ij}, \mathbf{X}_{ij})$ from each side of the change point, where $(\mathbf{Y}_{ij}, \mathbf{X}_{ij})$ are random samples collected from the model (15) with model parameters $(\mathbf{B}_i, \mathbf{s}_i)$ for $i = 1$ or 2. The difference between the parameters $\boldsymbol{\beta}_i = (\mathrm{Vec}^T(\mathbf{B}_i), \mathbf{s}_i^T)^T$ are of interest here, where $\mathrm{Vec}(\mathbf{S})$ denotes the $(p \cdot q)$-dimensional vector formed by stacking the

columns of a $(p \times q)$-dimensional matrix $\mathbf{S}$. In this example, $d = pq + q$.

A direct way to obtain $\tilde{\boldsymbol{\beta}}_i$ is

$$\tilde{\mathbf{B}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\tilde{\mathbf{B}}_{ij}, \quad \tilde{\mathbf{B}}_{ij} = (\mathbf{X}_{ij}^T\mathbf{X}_{ij})^{-1}\mathbf{X}_{ij}^T\tilde{\mathbf{Y}}_{ij}, \quad \tilde{\mathbf{s}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\tilde{\mathbf{s}}_{ij},$$
$$\tilde{\mathbf{s}}_{ij} = \mathrm{diag}[\tilde{\boldsymbol{\Sigma}}_{ij}], \quad \tilde{\boldsymbol{\Sigma}}_{ij} = (\mathbf{Y}_{ij} - \mathbf{X}_{ij}\tilde{\mathbf{B}}_{ij})^T(\mathbf{Y}_{ij} - \mathbf{X}_{ij}\tilde{\mathbf{B}}_{ij})/(m-p).$$

Thus, $\tilde{\boldsymbol{\beta}}_i = (\mathrm{Vec}^T(\tilde{\mathbf{B}}_i), \tilde{\mathbf{s}}_i^T)^T$. Note that the asymptotic covariance matrix of $\tilde{\boldsymbol{\beta}}_i$ can be approximated by

$$\widehat{\boldsymbol{\Omega}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathrm{diag}\{\tilde{\boldsymbol{\Sigma}}_{ij} \otimes (\mathbf{X}_{ij}^T\mathbf{X}_{ij})^{-1}, \tilde{\boldsymbol{\Gamma}}_{ij}\}.$$

The (k,l) elements in $\tilde{\boldsymbol{\Gamma}}_{ij}$, $\tilde{\rho}_{kl}$, is given by (Anderson 2003; Chapter 7),

$$\tilde{\rho}_{k,l} \approx \frac{\tilde{\sigma}_{(kl)}^4}{\tilde{\sigma}_{(kk)}\tilde{\sigma}_{(ll)}},$$

where $\tilde{\boldsymbol{\Sigma}}_{ij} = (\tilde{\sigma}_{(kl)}^2)$. It can be seen that when $m > pq$ the assumptions in Appendix A hold. This is not restrictive and can easily be satisfied in practical applications.

We now present some simulation results to assess the effectiveness of the proposed method for diagnosing any parameter changes in the profile problem. Comparing the LEB procedure with alternative methods turns out to be difficult, due to the lack of an obvious comparative method because most of the approaches in the literature are designed for univariate linear profiles. Thus, we only consider single profile case, say $q = 1$, and use the parametric testing approach (called PT for abbreviation) provided in Section 7.2 of Zou et al. (2007). We consider the linear profile model with covariates $(x_1, x_1^2, x_2, x_2^2, x_1x_2, \ldots, x_l, x_l^2, x_{1l}, \ldots x_{l-1l})$, where the design points $\mathbf{x}_i$ are generated from a uniform distribution. We also center those design points so that their means are zero. This example, containing both multiple and polynomial regression terms, is quite common in practice. For illustration purpose, $l = 4$ is used and there are $p = 1 + 2l + l(l-1)/2 = 15$ covariates. Without loss of generality, the variance of the error term is set to be one and the following three types of faults are considered for comparison: (i) $(\delta_1, \delta_2) = (1, 1)$; (ii) $(\delta_1, \delta_2, \delta_5) = (1, 1, 1)$; and (iii) $(\delta_1, \delta_2, \delta_5, \delta_{16}) = (1, 1, 1, 1)$. Note that δ_{16} indicates a change in the variance of profile. The simulation results with two choices of sample sizes are considered in Table 3. The type I error probability

in the PT approach is fixed at 0.002 in this table. It can be seen that the two methods perform reasonably satisfactorily in most of the cases as shown by the correct identification probability and PSS index. The LEB has a significant advantage over PT, especially for small values of m and (n_1, n_2). Although the type I error probability can be adjusted to improve the sensitivity of the PT approach, it should be emphasized that the PT approach contains multiple tests, including a t-test, a χ^2-test and a multivariate F-test, and thus requires more statistical and sophisticated knowledge on the choice of false alarm rate for each test. In contrast, the LEB procedure provides a unified diagnosis, which does not depend on the pre-specified type I error probability, and is thus easier to implement in practice.

Table 4: Diagnostic comparisons between the LEB and parametric testing procedures for the profile problem.

m	scenarios	C	PSS	C	PSS	C	PSS	C	PSS
		$(n_1, n_2) = (20, 20)$				$(n_1, n_2) = (40, 60)$			
		PT		LEB		PT		LEB	
30	(i)	0.52	0.87	0.68	0.57	0.67	0.73	0.82	0.39
	(ii)	0.38	1.81	0.56	1.16	0.55	1.34	0.72	0.71
	(iii)	0.20	2.53	0.37	1.81	0.36	2.06	0.55	1.30
50	(i)	0.77	0.45	0.89	0.20	0.92	0.13	0.97	0.04
	(ii)	0.69	0.72	0.80	0.42	0.91	0.18	0.95	0.08
	(iii)	0.39	1.61	0.54	1.12	0.70	0.68	0.82	0.37
100	(i)	0.92	0.16	0.96	0.06	0.99	0.01	0.99	0.01
	(ii)	0.89	0.21	0.94	0.10	0.99	0.01	0.99	0.01
	(iii)	0.57	0.97	0.72	0.54	0.90	0.18	0.96	0.07

5 A Real-Data Application

In this section, we demonstrate the proposed methodology by applying it to a real dateset from a white wine production process. The data contains totally 4898 observations, and is publicly available in the dataset "Wine Quality" of the UCI repository (can be downloaded from the web site http://archive.ics.uci.edu/ml/datasets/Wine+Quality). The data were collected from May/2004 to February/2007 using only protected designation of origin samples that were tested at the official certification entity, which is an inter-professional organization with the goal of improving the quality and marketing of Portuguese *Vinho Verde* wine. The data were recorded by a computerized system, which automatically manages the process of wine sample testing from producer requests to laboratory and sensory analysis. For each observation, there are eleven continuous measurements (based on physicochemical tests) including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, PH, sulphates and alcohol. A categorical variable, quality, indicating the wine quality between 0 (very bad) and 10 (very excellent), is also provided based on sensory analysis. The goal of this data analysis is mainly to model and monitor wine quality based on physicochemical tests. More detailed discussion about this example and dataset can be found in Cortez et al. (2009) and the references therein.

As mentioned by Cortez et al. (2009), we are not sure if all input variables are relevant to distinguish two categories (quality levels) of wine. So, it could be interesting to consider some feature selection methods to determine which physicochemical factor(s) are responsible for the change of quality. This is closely related to our diagnosis problem discussed above, and hence we may consider the use of the LEB method.

Under the SPC context of sequential monitoring the wine production process, we suppose that the standard quality level is the index "seven" (LV7; as also suggested by Cortez et al. 2009) and use all the observations belonging to this level (totally 880) as a training sample except for the last ten observations. Some summary statistics of the LV7 white-wine data are presented in Table 5. Then, we artificially assume that we firstly monitor those ten observations from LV7 and then obtain the observations categorized as the level "six" (LV6) sequentially. Similar to Cortez et al. (2009), the location parameter is of greatest interest and

Table 5: Summary statistics of the white-wine quality data with quality index "seven"

Sample mean vector										
6.735	0.263	0.326	5.186	0.038	34.10	125.1	0.992	3.214	0.503	11.37
Sample standard deviations										
0.756	0.091	0.079	4.298	0.011	13.24	32.74	0.003	0.158	0.130	1.247
Sample correlation matrix										
1.000	-0.090	0.265	0.234	0.139	-0.000	0.174	0.409	-0.492	-0.091	-0.274
-0.090	1.000	-0.260	-0.024	-0.282	-0.169	-0.081	-0.276	0.048	-0.018	0.502
0.265	-0.260	1.000	0.044	0.166	0.157	0.117	0.141	-0.125	-0.027	-0.155
0.234	-0.024	0.044	1.000	0.275	0.118	0.455	0.822	-0.335	-0.109	-0.480
0.139	-0.282	0.166	0.275	1.000	0.191	0.398	0.494	-0.088	0.049	-0.554
-0.000	-0.169	0.157	0.118	0.191	1.000	0.532	0.181	0.029	0.155	-0.200
0.174	-0.081	0.117	0.455	0.398	0.532	1.000	0.577	-0.030	0.009	-0.465
0.409	-0.276	0.141	0.822	0.494	0.181	0.577	1.000	-0.167	0.030	-0.837
-0.492	0.048	-0.125	-0.335	-0.088	0.029	-0.030	-0.167	1.000	0.178	0.106
-0.091	-0.018	-0.027	-0.109	0.049	0.155	0.009	0.030	0.178	1.000	-0.046
-0.274	0.502	-0.155	-0.480	-0.554	-0.200	-0.465	-0.837	0.106	-0.046	1.000

thus we construct the standard multivariate exponentially weighted moving average control chart (Lowry et al. 1992) to monitor the wine quality. We choose $\lambda = 0.1$ and in-control average run length as 1,000. Accordingly, the control limit is 1.56. As shown in Figure 1-(a), the control chart triggers a signal after the eleventh LV6 observation is obtained. Then, by adapting the multivariate change-point estimator (Zamba and Hawkins 2006), we find that the estimator of change-point is 10, which accurately indicates the change-point location.

Now, we have $n_1 = 880$ and $n_2 = 11$. Then, we use the proposed LEB diagnostic procedure to identify the influential variables. Table 6 tabulates the resulting eleven LASSO estimates $\widehat{\boldsymbol{\delta}}_{\tilde{\theta}_j}$, for $j = 1, \ldots, 11$, and the corresponding values of $\text{EBIC}_{\tilde{\theta}_j}$. These values indicates that the shift may have occurred in the three factors, chlorides, density and alcohol. Figure 1 (b)-(d) show the time series plots of the raw data for these three factors, from which we can also clearly observe significant changes before and after the change-point. Such a feature selection result would be used to support the oenologist's wine evaluations, potentially

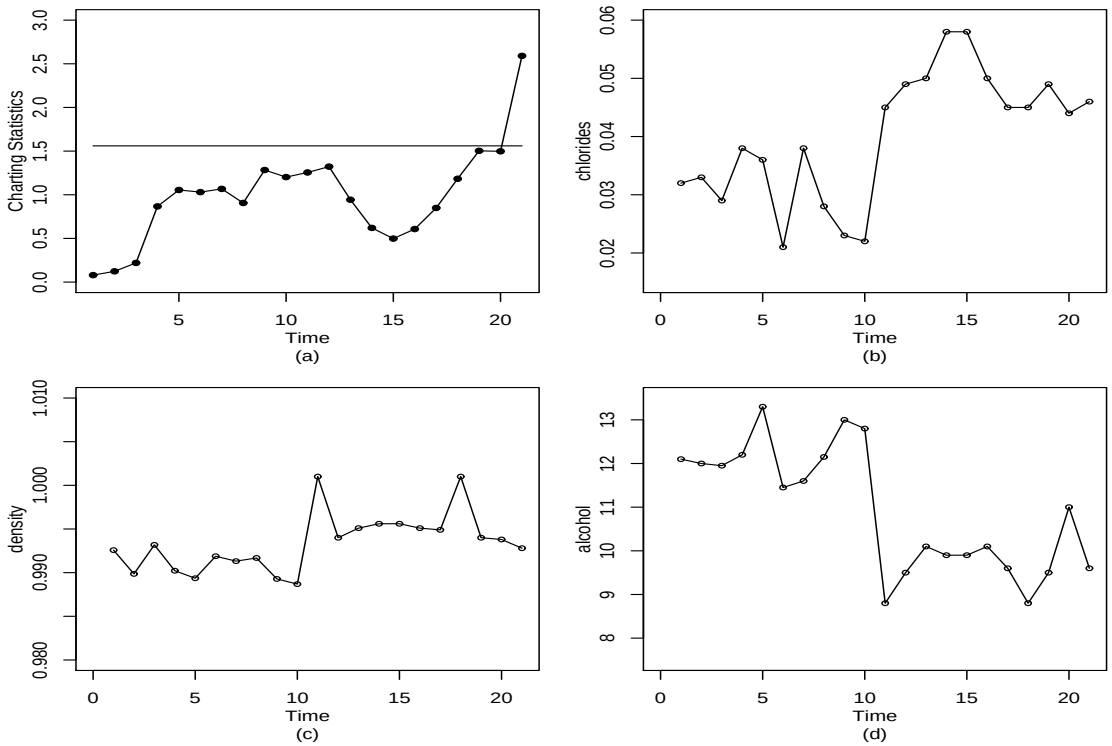

Figure 1: (a) The MEWMA control chart for monitoring the white-wine production process; (b)-(d) represent the time series plots of the observations for the variables chlorides, density and alcohol, respectively

improving the quality and speed of their decisions. Moreover, identifying the most important impacts among all the physicochemical tests in the wine quality is useful for improving the production process as suggested by Cortez et al. (2009) as well.

6 Concluding Remarks

Fault diagnosis in statistical process control remains a challenging problem for complex processes due to the complex nature of various systems. In this research, we start from a generic parametric model and frame the fault diagnosis as a variable (model) selection problem. A two-sample BIC is derived based on the least-squares approximation of the loss function in general for evaluating tentative models. A practical LASSO-based diagnostic procedure which combines an extension of BIC with the adaptive LASSO algorithm is proposed and its diagnostic consistency is shown under some mild conditions. The consistency theory

Table 6: Diagnostic results of the LEB procedure about white-wine quality data.

j	$\widehat{\boldsymbol{\delta}}_{\tilde{\theta}_j}$											$\text{EBIC}_{\tilde{\theta}_j}$
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-2×10^{-4}	0.000	0.000	0.000	49.61
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.001	0.000	0.000	0.584	39.76
3	0.000	0.000	0.000	0.000	**-0.007**	0.000	0.000	**-0.001**	0.000	0.000	**1.163**	**32.34**
4	0.000	0.000	0.000	0.000	-0.007	0.000	0.000	-0.002	0.000	0.010	1.197	38.45
5	-0.027	0.000	0.000	0.000	-0.007	0.000	0.000	-0.002	0.000	0.015	1.220	44.90
6	-0.066	0.000	0.000	-0.233	-0.008	0.000	0.000	-0.002	0.000	0.021	1.275	50.70
7	-0.127	0.000	0.000	-0.763	-0.008	0.000	-3.910	-0.002	0.000	0.030	1.412	55.12
8	-0.281	-0.018	0.000	-2.127	-0.010	0.000	-14.01	-0.003	0.000	0.053	1.639	58.04
9	-0.284	-0.020	0.002	-2.205	-0.010	0.000	-14.54	-0.003	0.000	0.055	1.647	65.08
10	-0.293	-0.021	0.003	-2.266	-0.010	0.000	-14.80	-0.003	0.002	0.056	1.653	72.16
11	-0.338	-0.026	0.010	-2.540	-0.010	1.398	-14.33	-0.003	0.019	0.063	1.658	78.99

guarantees correct fault diagnosis without any prior knowledge of thresholds and sensitivity. This property turns out to be a major advantage of the proposed method over other existing fault diagnosis methods in different applications. Furthermore, due to the piecewise linear property of LASSO solution path, the total computational effort is minimal, which makes our method very attractive in practice. Through industrial examples shown in Section 4, we demonstrate that the proposed procedure provides an efficient alternative for multivariate process diagnosis.

In our framework of the two-sample diagnosis problem, we assume that the change point time can be correctly identified from other SPC procedures. In practice, it may be possible that the change point identification is not accurate, or it may be coupled with the fault diagnosis task. In theory, this can be handled similarly by the Bayesian framework with some *a priori* knowledge of the change point. We will discuss it in another paper. In addition, considerable efforts have been devoted to construct post-signal change-point estimators in various monitoring problem (e.g., see Pignatiello and Samuel 2001, Zamba and Hawkins 2006, Zou et al. 2007 and Zou and Tsung 2008), but there lacks a systematic study

of the whole post-signal diagnostic procedure (including both the change-point estimate and fault isolation) for Phase II monitoring.

It is also worth noting again that we assume the observations in each partition come from the same model. Thus, the proposed model does not include the drift or ramp changes. Although our simulation results (available from the authors) reveal that the LEB procedure does work reasonably well in certain cases, it is definitely not be optimal because our model (1) is incorrect in such situations. This issue requires further investigation. Moreover, in this method, we require that there exists full-rank estimators of the (asymptotic) covariance matrix of $\tilde{\boldsymbol{\beta}}_i$, say, $\widehat{\boldsymbol{\Omega}}_i$. However, in an ultra-high dimensional situation, i.e., when d is very large and even larger than the sample sizes, this condition may not be easily satisfied. It is of great interest to design an appropriate procedure in such cases and to verify whether the diagnostic consistent property is still valid, which warrants further research.

Acknowledgement

The authors thank the editor, associate editor, and two anonymous referees for their many helpful comments that have resulted in significant improvements in the article.

Appendices

Recall the penalized least-squares problem (13) and the associated notation. Throughout this section, we will use the following notation for ease of exposition. Denote $n = n_1 n_2/(n_1 + n_2)$, $a_n = \max\{\theta_j, j \in s_T\}$ and $b_n = \min\{\theta_j, j \notin s_T\}$. Denote by $\boldsymbol{\delta}_s$ the d-dimensional parameter vector $\boldsymbol{\delta}$ with those elements outside s being set to 0, that is, $\forall i \in s^c, \delta_i = 0$, where s^c is the complement of s in $\mathcal{S}$. Given a d-dimensional vector $\boldsymbol{\delta}$, $\boldsymbol{\delta}^{(s)}$ indicates the sub-vector which consists of all elements of $\boldsymbol{\delta}$ whose indices are in s. Given a $p \times p$ matrix $\mathbf{K}$, $\mathbf{K}^{(s,s')}$ indicates the sub-matrix which consists of all rows and columns of $\mathbf{K}$ whose indices are in s and s', respectively, where s' is another subset of $\{1, \ldots, d\}$.

Appendix A: Assumptions

Assumption 1: $\tilde{\boldsymbol{\beta}}_i$ are $\sqrt{n_i}$-consistent estimators of $\boldsymbol{\beta}_i$ for $i = 1, 2$.

Assumption 2: $\sqrt{n/\log n}\liminf_{n\to\infty}(\min_{i\in s_T}|\delta_{nj}|) \to \infty$.

Assumption 3: $\widehat{\boldsymbol{\Omega}}_i$ are consistent estimators of $\boldsymbol{\Omega}_i$ and positive-definite for $i = 1, 2$.

These three assumptions imposed here are all used for obtaining the property of the penalized estimator $\widehat{\boldsymbol{\delta}}$. They are quite mild and typically hold in many problems (see the examples in Section 4).

Appendix B: A quadratic approximation to $L(\boldsymbol{\beta}_1, \boldsymbol{\delta})$

Recall the notation $\tilde{\boldsymbol{\delta}} = \tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1$. A standard Taylor series expansion of $L_i(\boldsymbol{\beta}_i)$ at $\tilde{\boldsymbol{\beta}}_i$ is as follows,

$$L_i(\boldsymbol{\beta}_i) \approx L_i(\tilde{\boldsymbol{\beta}}_i) + \dot{L}_i(\tilde{\boldsymbol{\beta}}_i)(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i) + \frac{1}{2}(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i)^T \ddot{L}_i(\tilde{\boldsymbol{\beta}}_i)(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i),\ i = 1, 2,$$

where $\dot{L}_i$ and $\ddot{L}_i$ are the first- and second-order derivatives of the loss function $L_i(\cdot)$. Because $\tilde{\boldsymbol{\beta}}_i$ is the minimizer of $L_i(\cdot)$, we know that $\dot{L}_i(\tilde{\boldsymbol{\beta}}_i) = 0$. Thus, by ignoring the constant $L_i(\tilde{\boldsymbol{\beta}}_i)$ and the coefficient $1/2$, the joint minimization function (1) can be simplified and approximated as

$$L'(\boldsymbol{\beta}_1, \boldsymbol{\delta}) \approx (\boldsymbol{\beta}_1 - \tilde{\boldsymbol{\beta}}_1)^T \ddot{L}_1(\tilde{\boldsymbol{\beta}}_1)(\boldsymbol{\beta}_1 - \tilde{\boldsymbol{\beta}}_i) + (\boldsymbol{\beta}_1 + \boldsymbol{\delta} - \tilde{\boldsymbol{\beta}}_2)^T \ddot{L}_2(\tilde{\boldsymbol{\beta}}_2)(\boldsymbol{\beta}_1 + \boldsymbol{\delta} - \tilde{\boldsymbol{\beta}}_2). \quad \text{(A.1)}$$

Furthermore, assume that $\ddot{L}_i(\cdot)$ for $i = 1, 2$ are full-rank. Given $\boldsymbol{\delta}$, the minimizer of $L'(\boldsymbol{\beta}_1, \boldsymbol{\delta})$ for $\boldsymbol{\beta}_1$ is

$$\widehat{\boldsymbol{\beta}}_1(\boldsymbol{\delta}) = [\ddot{L}_1(\tilde{\boldsymbol{\beta}}_1) + \ddot{L}_2(\tilde{\boldsymbol{\beta}}_2)]^{-1}[\ddot{L}_1(\tilde{\boldsymbol{\beta}}_1)\tilde{\boldsymbol{\beta}}_1 + \ddot{L}_2(\tilde{\boldsymbol{\beta}}_2)\tilde{\boldsymbol{\beta}}_2 - \ddot{L}_2(\tilde{\boldsymbol{\beta}}_2)\boldsymbol{\delta}]$$

Consequently, substituting $\widehat{\boldsymbol{\beta}}_1(\boldsymbol{\delta})$ into (A.1), we will have an objective function dependent only on $\boldsymbol{\delta}$

$$L''(\boldsymbol{\delta}) = (\widehat{\boldsymbol{\beta}}_1(\boldsymbol{\delta}) - \tilde{\boldsymbol{\beta}}_1)^T \ddot{L}_1(\tilde{\boldsymbol{\beta}}_1)(\widehat{\boldsymbol{\beta}}_1(\boldsymbol{\delta}) - \tilde{\boldsymbol{\beta}}_1) + (\widehat{\boldsymbol{\beta}}_1(\boldsymbol{\delta}) + \boldsymbol{\delta} - \tilde{\boldsymbol{\beta}}_2)^T \ddot{L}_2(\tilde{\boldsymbol{\beta}}_2)(\widehat{\boldsymbol{\beta}}_1(\boldsymbol{\delta}) + \boldsymbol{\delta} - \tilde{\boldsymbol{\beta}}_2).$$

It is easily verified that the minimizer of $L''(\boldsymbol{\delta})$ is given by $\arg\min_{\boldsymbol{\delta}} L''(\boldsymbol{\delta}) = \tilde{\boldsymbol{\delta}}$. A second-order Taylor series expansion of $L''(\boldsymbol{\delta})$ at $\tilde{\boldsymbol{\delta}}$ will give

$$\begin{aligned} L''(\boldsymbol{\delta}) &\approx L''(\tilde{\boldsymbol{\delta}}) + \dot{L}''(\tilde{\boldsymbol{\delta}})(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}) + \frac{1}{2}(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}})^T \ddot{L}''(\tilde{\boldsymbol{\delta}})(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}) \\ &= L''(\tilde{\boldsymbol{\delta}}) + \frac{1}{2}(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}})^T \ddot{L}''(\tilde{\boldsymbol{\delta}})(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}). \end{aligned}$$

By ignoring the constant $L''(\tilde{\boldsymbol{\delta}})$ and the coefficient $1/2$, $L''(\boldsymbol{\delta})$ will be simplified as

$$\begin{aligned} &(\tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\delta})^T \ddot{L}''(\tilde{\boldsymbol{\delta}})(\tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\delta}) \\ &\quad = (\tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\delta})^T (\ddot{L}_1^{-1} + \ddot{L}_2^{-1})^{-1} (\tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\delta}), \end{aligned} \tag{A.2}$$

where we use the fact that

$$\ddot{L}_2(\ddot{L}_1 + \ddot{L}_2)^{-1}\ddot{L}_1(\ddot{L}_1 + \ddot{L}_2)^{-1}\ddot{L}_2 + \ddot{L}_1(\ddot{L}_1 + \ddot{L}_2)^{-1}\ddot{L}_2(\ddot{L}_1 + \ddot{L}_2)^{-1}\ddot{L}_1 = (\ddot{L}_1^{-1} + \ddot{L}_2^{-1})^{-1},$$

and $\ddot{L}_i$ instead of $\ddot{L}_i(\tilde{\boldsymbol{\beta}}_i)$ for abbreviation which should not cause any confusion.

Appendix C: Proof of Theorem 1

In order to prove Theorem 1, we need establish the $\sqrt{n}$-consistency and selection consistency of the penalized estimator $\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}}$ which are summarized in the following lemma.

Lemma 1 *If Assumptions 1-3 hold and the penalty parameter $\boldsymbol{\theta}$ satisfies $\sqrt{n}a_n \xrightarrow{p} 0$ and $\sqrt{n}b_n \xrightarrow{p} \infty$ then as $\min\{n_1, n_2\} \to \infty$, the minimizer of (11), $\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}}$, must satisfy:*

(i) $\Pr(\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}}^{(s_T^c)} = \mathbf{0}) \to 1$;

(ii) $\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}}^{(s_T)} = \boldsymbol{\delta}_n^{(s_T)} + O_p(n^{-1/2})$.

These results can be obtained in a similar way to Theorems 1-2 in Wang and Leng (2007), or to Theorems 1-2 in Fan and Li (2001). The only difference is that $\boldsymbol{\delta}_n$ may be of order $o(1)$ here. The technical arguments in the proof of Theorems 1-2 in Wang and Leng (2007) continue to hold in the current setting, and we only need to use (7) to obtain the result that $||\tilde{\boldsymbol{\Lambda}}\sqrt{n}(\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}_n)|| = O_p(1)$. For simplicity, details of these arguments are omitted here. □

For penalized-type estimators, Wang et al. (2007) and Wang and Leng (2007) respectively established the consistency of smoothly clipped absolute deviation method (SCAD) and adaptive LASSO estimators with the tuning parameter chosen by a BIC-type criterion. By using Lemma 1, we generalize the proof of Theorem 4 in Wang and Leng (2007) to the two-sample case and also allow $\boldsymbol{\delta}_n \to 0$.

Following a similar idea of Wang et al. (2007), according to whether the resulting model $s_{\boldsymbol{\theta}}$ is underfitted, correctly fitted, or overfitted, we can partition $\mathbb{R}^{d+}$ into the following three mutually exclusive regions:

$$\begin{aligned}
\mathbb{R}_U^{d+} &= \{\boldsymbol{\theta} \in \mathbb{R}^{d+} : s_{\boldsymbol{\theta}} \not\supseteq s_T\}, \\
\mathbb{R}_T^{d+} &= \{\boldsymbol{\theta} \in \mathbb{R}^{d+} : s_{\boldsymbol{\theta}} = s_T\}, \text{ and} \\
\mathbb{R}_O^{d+} &= \{\boldsymbol{\theta} \in \mathbb{R}^{d+} : s_{\boldsymbol{\theta}} \supset s_T, s_{\boldsymbol{\theta}} \neq s_T\}.
\end{aligned}$$

For the purpose of proof, we could readily define a reference tuning parameter sequence $\boldsymbol{\theta}_* \in \mathbb{R}^{d+}$ which satisfies the conditions in Lemma 1. For instance, we could set $\theta_{*k} = \theta_* |\tilde{\delta}_k|^{-1}$ with $\theta_* = n^{-1}(\log n)^{1/2}$. By Lemma 1-(i), $s_{\boldsymbol{\theta}_*} \in \mathbb{R}_T^{d+}$ with probability tending to 1. Thus, to prove the theorem, it suffices to show that $\Pr(\inf_{\boldsymbol{\theta} \in \mathbb{R}_U^{d+} \cup \mathbb{R}_O^{d+}} \mathrm{BIC}_{\boldsymbol{\theta}} > \mathrm{BIC}_{\boldsymbol{\theta}^*}) \to 1$. The following proof consists of two steps.

Step 1. Let us firstly consider $\boldsymbol{\theta} \in \mathbb{R}_O^{d+}$. We then have $d_{\boldsymbol{\theta}} - d_{\boldsymbol{\theta}_*} \geq 1$. We shall show that with probability approaching one, the BIC favors $s_{\boldsymbol{\theta}_*}$. Before proceeding, to facilitate our proof we need another definition, the unpenalized estimate of $\boldsymbol{\delta}_n$ under the model identified by $\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}}$, say

$$\tilde{\boldsymbol{\delta}}_{s_{\boldsymbol{\theta}}} = \underset{\boldsymbol{\delta} \in \mathbb{R}^d : \boldsymbol{\delta}^{(s_{\boldsymbol{\theta}})} = \mathbf{0}}{\arg\min} \; g(\boldsymbol{\delta}).$$

By this definition, we must have $g(\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}}) \geq g(\tilde{\boldsymbol{\delta}}_{s_{\boldsymbol{\theta}}})$. Also, note that if $\boldsymbol{\delta}^{(s_{\boldsymbol{\theta}})} = \mathbf{0}$,

$$\begin{aligned}
g(\boldsymbol{\delta}) =& (\tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}})} - \boldsymbol{\delta}^{(s_{\boldsymbol{\theta}})})^T \tilde{\boldsymbol{\Lambda}}^{(s_{\boldsymbol{\theta}}, s_{\boldsymbol{\theta}})} (\tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}})} - \boldsymbol{\delta}^{(s_{\boldsymbol{\theta}})}) + 2(\tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}})} - \boldsymbol{\delta}^{(s_{\boldsymbol{\theta}})})^T \tilde{\boldsymbol{\Lambda}}^{(s_{\boldsymbol{\theta}}, s_{\boldsymbol{\theta}}^c)} \tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}}^c)} \\
&+ (\tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}}^c)})^T \tilde{\boldsymbol{\Lambda}}^{(s_{\boldsymbol{\theta}}^c, s_{\boldsymbol{\theta}}^c)} \tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}}^c)}.
\end{aligned}$$

Thus, the minimizer $\tilde{\boldsymbol{\delta}}_{s_{\boldsymbol{\theta}}}$ must satisfy the estimating equation

$$-\tilde{\boldsymbol{\Lambda}}^{(s_{\boldsymbol{\theta}}, s_{\boldsymbol{\theta}})} (\tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}})} - \tilde{\boldsymbol{\delta}}_{s_{\boldsymbol{\theta}}}^{(s_{\boldsymbol{\theta}})}) - \tilde{\boldsymbol{\Lambda}}^{(s_{\boldsymbol{\theta}}, s_{\boldsymbol{\theta}}^c)} \tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}}^c)} = 0$$

By Assumption 1, Lemma 1-(ii) and the arguments above, we know $\tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}}^c)} = O_p(n^{-1/2})$. As a result, $\tilde{\boldsymbol{\delta}}_{s_{\boldsymbol{\theta}}}^{(s_{\boldsymbol{\theta}})} = \tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}})} + O_p(n^{-1/2})$. Hence, as $\min\{n_1, n_2\} \to \infty$,

$$\begin{aligned}\mathrm{BIC}_{\boldsymbol{\theta}} - \mathrm{BIC}_{\boldsymbol{\theta}_*} &\geq g(\tilde{\boldsymbol{\delta}}_{s_{\boldsymbol{\theta}}}) - g(\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}_*}) + \log n + 2\log d \\ &= O_p(1) + \log n \to \infty,\end{aligned}$$

which implies $\Pr(\mathrm{BIC}_{\boldsymbol{\theta}} - \mathrm{BIC}_{\boldsymbol{\theta}_*} > 0) \to 1$ for any $\boldsymbol{\theta} \in \mathbb{R}_O^{d+}$.

Step 2. Now consider $\boldsymbol{\theta} \in \mathbb{R}_U^{d+}$. In this case, similar to the Step 1, we have

$$\begin{aligned}\mathrm{BIC}_{\boldsymbol{\theta}} - \mathrm{BIC}_{\boldsymbol{\theta}_*} &\geq g(\tilde{\boldsymbol{\delta}}_{s_{\boldsymbol{\theta}}}) - g(\widehat{\boldsymbol{\delta}}_{\boldsymbol{\theta}_*}) + (d_{\boldsymbol{\theta}} - d_{\boldsymbol{\theta}_*}) \cdot (\log n + 2\log d) \\ &= g(\tilde{\boldsymbol{\delta}}_{s_{\boldsymbol{\theta}}}) + (d_{\boldsymbol{\theta}} - d_{\boldsymbol{\theta}_*}) \cdot \log n + O_p(1) \\ &\geq (\tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}}^c)})^T \left[\tilde{\boldsymbol{\Lambda}}^{(s_{\boldsymbol{\theta}}^c, s_{\boldsymbol{\theta}}^c)} - (\tilde{\boldsymbol{\Lambda}}^{(s_{\boldsymbol{\theta}}, s_{\boldsymbol{\theta}}^c)})^T (\tilde{\boldsymbol{\Lambda}}^{(s_{\boldsymbol{\theta}}, s_{\boldsymbol{\theta}})})^{-1} \tilde{\boldsymbol{\Lambda}}^{(s_{\boldsymbol{\theta}}, s_{\boldsymbol{\theta}}^c)}\right] \tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}}^c)} - d_{\boldsymbol{\theta}_*} \cdot \log n \quad \text{(A.3)} \\ &\geq \lambda_{d-d_\theta}(\mathbf{B}) ||\tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}}^c)}||^2 - d_{\boldsymbol{\theta}_*} \cdot \log n + O_p(1),\end{aligned}$$

where we denote the matrix in the bracket in (A.3) by $\mathbf{B}$ and $\lambda_{d-d_\theta}(\mathbf{B})$ is its smallest eigenvalue. Note that $\mathbf{B}^{-1} = (\tilde{\boldsymbol{\Lambda}}^{-1})^{(s_{\boldsymbol{\theta}}, s_{\boldsymbol{\theta}})}$ and thus $n^{-1}\lambda_{d-d_\theta}(\mathbf{B}) > 0$ where we use (7) again. By Assumption 2, $(n/\log n)||\tilde{\boldsymbol{\delta}}^{(s_{\boldsymbol{\theta}}^c)}||^2 \to \infty$ and it follows immediately that $\Pr(\mathrm{BIC}_{\boldsymbol{\theta}} > \mathrm{BIC}_{\boldsymbol{\theta}_*}) \to 1$ for any $\boldsymbol{\theta} \in \mathbb{R}_U^{d+}$.

Combining the two cases together implies that any $\boldsymbol{\theta}$ failing to identify the true model cannot be selected as the optimal parameter. That is to say, the model associated with the optimal $\boldsymbol{\theta}$ must be the true one. This completes the proof. □

References

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis, 3rd ed*, Wiely, New York.

Broman, K. W., and Speed, T. P. (2002), "A Model Selection Approach for the Identification of Quantitative Trait Loci in Experimental Crosses," *Journal of the Royal Statistical Society: Series B*, 64, 641–56.

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criterion for Model Selection with Large Model Spaces," *Biometrika*, 95, 759–771.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009), "Modeling Wine Preferences by Data Mining from Physicochemical Properties," *Decision Support Systems*, 47, 547–553.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–489.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Han D., and Tsung, F. (2006), "A Reference-Free Cuscore Chart for Dynamic Mean Change Detection and a Unified Framework for Charting Performance Comparison," *Journal of the American Statistical Association*, 101, 368–386.

Hawkins, D. M., (1991), "Multivariate Quality Control Based on Regression-Adjusted Variables," *Technometrics*, 33, 61–75.

Hettmansperger, T. P., and McKean, J. W. (1998), *Robust Nonparametric Statistical Methods*, Arnold, London.

Huwang, L., Yeh, A. B., Wu, C. (2007), "Monitoring Multivariate Process Variability for Individual Observations," *Journal of Quality Technology*, 39, 258–278.

Jiang, W., and Tsui, K.-L. (2008), "A Theoretical Framework and Efficiency Study of Multivariate Control Charts," *IIE Transactions*, 40, 650–663.

Jin, J., and Shi, J. (1999), "State Space Modeling of Sheet Metal Assembly for Dimensional Control," *ASME Transactions, Journal of Manufacturing Science and Engineering*, 121, 756–762.

Li, J., Jin, J., and Shi, J. (2008), "Causation-Based T^2 Decomposition for Multivariate Process Monitoring and Diagnosis," *Journal of Quality Technology*, 40, 46–58.

Li, Y., and Tsung, F. (2009), "False Discovery Rate-Adjusted Charting Schemes for Multistage Process Monitoring and Fault Identification," *Technometrics*, 51, 186–205.

Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E. (1992), "Multivariate Exponentially Weighted Moving Average Control Chart," *Technometrics*, 34, 46–53.

Marcus, M., and Minc, H. (1992), *Survey of Matrix Theory and Matrix Inequalities*, Dover, New York.

Marvelakis, P. E., Bersimis, S., Panaretos, J., and Psarakis, S. (2002), "Identifying the out of Control Variable in a Multivariate Control Chart," *Communications in Statistics: Theory and Methods*, 31, 2391–2408.

Mason, R. L., Tracy, N. D., and Young, J. C. (1995), "Decomposition of T^2 for Multivariate Control Chart Interpretation," *Journal of Quality Technology*, 27, 99–108.

Mason, R. L., Chou, Y. M., and Young, J. C. (2001), "Applying Hotelling's T^2 Statistic to Batch Processes," *Journal of Quality Technology*, 33, 466–479.

Mason, R. L., Tracy, N. D., and Young, J. C. (1997), "A Practical Approach for Interpreting Multivariate T^2 Control Chart Signals," *Journal of Quality Technology*, 29, 396–406.

Mason, R. L., and Young, J. C. (2002), *Multivariate Statistical Process Control With Industrial Application*, Philadelphia: SIAM.

Neudecker, H., and Wesselman, A. M. (1990), "The Asymptotic Covariance of the Sample Correlation Matrix," *Linear Algebra and Its Applications*, 127, 589–599.

Nishii, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *The Annals of Statistics*, 12, 758–765.

Pignatiello, J. J., Jr., and Samuel, T. R. (2001), "Estimation of the Change Point of a Normal Process Mean in SPC Applications," *Journal of Quality Technology*, 33, 82–95.

Raftery, A. E. (1996), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," *Biometrika*, **83**, 251–266.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shi, J. and Zhou, S. (2009), "Quality Control and Improvement for Multistage Systems: A Survey," *IIE Transactions*, 41, 744–753.

Sullivan, J. H., Stoumbos, Z. G., Mason, R. L., and Young, J. C. (2007), "Step-Down Analysis for Changes in the Covariance Matrix and Other Parameters," *Journal of Quality Technology*, 39, 66–84.

Sullivan, J. H., and Woodall, W. H. (2000), "Change-Point Detection of Mean Vector or Covariance Matrix Shifts Using Multivariate Individual Observations," *IIE Transactions*, 32, 537–549.

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Wang, K., and Jiang, W. (2009) "High-Dimensional Process Monitoring and Fault Isolation via Variable Selection," *Journal of Quality Technology*, 41, 247–258.

Wang, H., and Leng, C. (2007), "Unified Lasso Estimation by Least Square Approximation", *Journal of the American Statistical Association*, 102, 1039–1048.

Wang, H., Li, R., and Tsai, C. L. (2007), "On the Consistency of SCAD Tuning Parameter Selector," *Biometrika*, 94, 553–568.

Woodall, W. H. (2007), "Current Research on Profile Monitoring," *Revista Producão*, 17, 420–425.

Woodall, W. H., Spitzner, D. J., Montgomery, D. C., and Gupta, S. (2004), "Using Control Charts to Monitor Process and Product Quality Profiles," *Journal of Quality Technology*, 36, 309–320.

Yang, Y. (2005), "Can the Strengths of AIC and BIC Be Shared?–A Conflict between Model Identification and Regression Estimation," *Biometrika*, 92, 937–950.

Zamba, K. D., and Hawkins, D. M. (2006), "A Multivariate Change-Point for Statistical Process Control," *Technometrics*, 48, 539–549.

Zhou, S., Ding, Y., Chen, Y. and Shi, J. (2003), "Diagnosability Study of Multistage Manufacturing Processes Based on Linear Mixed-Effects Models," *Technometrics* 45, 312–325.

Zhu, Y. and Jiang, W. (2009) "An Adaptive T^2 Chart for Multivariate Process Monitoring and Diagnosis," *IIE Transactions*, 41, 1007–1018.

Zou, C., and Qiu, P. (2009), "Multivariate Statistical Process Control Using LASSO," *Journal of the American Statistical Association*, 104, 1586–1596.

Zou, C., and Tsung, F. (2008), "Directional MEWMA Schemes for Multistage Process Monitoring and Diagnosis," *Journal of Quality Technology*, 40, 407–427.

Zou, C., Tsung, F., Liu, Y. (2008), "A Change Point Approach for Phase I Analysis in Multistage Processes," *Technometrics*, 50, 344–356.

Zou, C., Tsung, F., and Wang, Z. (2007), "Monitoring General Linear Profiles Using Multivariate EWMA Schemes," *Technometrics*, 49, 395–408.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007), "On the 'Degrees of Freedom' of Lasso," *The Annals of Statistics*, 35, 2173–2192.