# Multivariate Statistical Process Control Using LASSO

Changliang Zou[1] and Peihua Qiu[2]

[1] *Department of Statistics, Nankai University, China*

[2] *School of Statistics, University of Minnesota, USA*

**Abstract**

This paper develops a new multivariate statistical process control (SPC) methodology based on adapting the LASSO variable selection method to the SPC problem. The LASSO method has the *sparsity* property that it can select exactly the set of nonzero regression coefficients in multivariate regression modeling, which is especially useful in cases when the number of nonzero coefficients is small. In multivariate SPC applications, process mean vectors often shift in a small number of components. Our major goal is to detect such a shift as soon as it occurs and identify the shifted mean components. Using this connection between the two problems, a LASSO-based multivariate test statistic is proposed, which is then integrated into the multivariate EWMA charting scheme for Phase II multivariate process monitoring. It is shown that this approach balances protection against various shift levels and shift directions, and hence provides an effective tool for multivariate SPC applications.

**Keywords:** Hotelling's $T^2$ statistic; Model selection; Multivariate statistical process control; Penalized likelihood; Regression-adjusted variables.

# 1 Introduction

In *multivariate statistical process control* (MSPC), one monitors several quality characteristics of a process. The fundamental tasks of MSPC are to determine whether a multivariate process mean $\boldsymbol{\mu}$ has changed; to identify when a detected shift in $\boldsymbol{\mu}$ occurred; and to isolate the shifted components of $\boldsymbol{\mu}$. Methods for accomplishing these tasks are usually derived under the assumption that the observed measurement vectors $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{pi})'$ are distributed $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ for $i = 1, 2, \ldots, \tau$, and $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ for $i = \tau+1, \ldots, n$, with $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$ known. Throughout we take $\boldsymbol{\mu}_0 = \mathbf{0}$ without loss of generality. Then a portmanteau test for detecting a mean shift occurring at $\tau$ is based on testing $H_0 : \boldsymbol{\mu} = \mathbf{0}$ versus $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$, using the likelihood ratio test (LRT) statistic $n\bar{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}}$ where $\bar{\boldsymbol{X}} = \sum_{i=1}^{n} \boldsymbol{X}_i/n$. This procedure assumes that the covariance matrix does not change. Replacing $\boldsymbol{\Sigma}$ with the sample covariance matrix $\mathbf{S}$ results in the Hotelling's $T^2$ statistic. Based on such test statistics, several MSPC control charts have been proposed, in the framework of cumulative sum (CUSUM) or exponentially weighted moving average (EWMA). Most charting statistics take quadratic forms of the related test statistics; for instance, Healy (1987), Croisier (1988), Pignatiello and Runger (1990), Hawkins (1991, 1993), Lowry et al. (1992), Runger and Prabhu (1996), Zamba and Hawkins (2006), and Hawkins and Maboudou-Tchao (2007). After a control chart signals a mean shift, a separate diagnostic procedure is often used for identifying the shift time and the component(s) of $\boldsymbol{\mu}$ that shifted. Commonly used diagnostic procedures include the ones based on decomposition of $T^2$ (e.g., Mason et al. 1995, 1997; Li et al. 2008) and various step-down procedures (e.g., Hawkins 1993; Mason and Young 2002; Sullivan et al. 2007).

MSPC control charts with quadratic charting statistics are powerful when one is interested in detecting shifts that occur in majority components of $\boldsymbol{\mu}$. In practice,

however, shifts often occur in only a few of the mean components. In such cases, more powerful control charts are possible. To be specific, if we know that the potential shift is in one specific direction $\boldsymbol{d}$, then the corresponding hypothesis testing problem becomes $H_0 : \boldsymbol{\mu} = \mathbf{0}$ versus $H_1' : \boldsymbol{\mu} = \delta\boldsymbol{d}$, where $\delta$ is an unknown constant. The corresponding LRT statistic is $n(\boldsymbol{d'}\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}})^2/\boldsymbol{d'}\boldsymbol{\Sigma}^{-1}\boldsymbol{d} \overset{H_0}{\sim} \chi_1^2$. In the case when only one measurement component shifts but the component index is unknown, the testing problem becomes

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{versus} \quad H_1'' : \boldsymbol{\mu} = \delta\boldsymbol{d}_1 \text{ or } \boldsymbol{\mu} = \delta\boldsymbol{d}_2 \text{ or } \ldots \text{ or } \boldsymbol{\mu} = \delta\boldsymbol{d}_p,$$

where $\boldsymbol{d}_j$ is the $p$-dimensional vector with 1 at the $j$-th element and 0 elsewhere. In this case, likelihood ratio arguments would lead to the following $p$ test statistics:

$$V_j = \sqrt{n}(\boldsymbol{d}_j'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}})/(\boldsymbol{d}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{d}_j)^{\frac{1}{2}}, \quad \text{for } j = 1, \ldots, p. \tag{1}$$

It can be checked that $V_j$s are just the regression-adjusted variables defined in Hawkins (1991). Hawkins (1991; 1993) suggested a MSPC control chart using $\max_{j=1,\ldots,p} |V_j|$ as a charting statistic, and showed that this chart was more effective than the one based on $T^2$ when the potential shift occurs in only one measurement component. Furthermore, after a shift is signaled, $\arg\max_j |V_j|$ can be used as a diagnostic tool to specify the shifted component. On the other hand, as noted by Hawkins (1991), the chart based on $\boldsymbol{V} = (V_1, \ldots, V_p)$ can perform poorly, as when the shift occurs in several highly correlated components and when the shift is proportional to the two leading principal components of the covariance matrix. These findings are consistent with the properties of LRT with misspecified alternatives. Moreover, when the shift occurs in more than one component, diagnosis based on $\boldsymbol{V}$ could be misleading, because it cannot indicate multiple shift components correctly.

Based on the above discussion, it is natural to consider more directions as *a priori* alternatives and to construct corresponding test statistics. For instance, we

can assume that two or more measurement components may shift; the resulting control charts would alleviate certain problems mentioned above. However, such control charts are infeasible when the dimension $p$ is large, since the total number of possible shift directions would increase exponentially with $p$. Furthermore, their shift-detection power would decline substantially when $p$ increases, due to the well-known problem of multiple tests. Regarding post-detection diagnosis, conventional approaches (e.g., the decomposition of $T^2$ procedure and the step-down test) are theoretically sound, but are inefficient when $p$ is large. For instance, the decomposition of $T^2$ procedure considers $p!$ different decompositions of $T^2$, which is computationally expensive in such cases. Certain parameters in these approaches (e.g., the threshold values) affect their diagnostic ability significantly, but are generally difficult to determine. See Sullivan et al. (2007) for related discussion.

In this paper, we propose a MSPC chart that has three features: (1) It is capable of detecting shifts in one or more components efficiently; (2) its computation is reasonably simple; and (3) it provides an effective post-signal diagnostic tool for identifying shifted measurement components. By adapting certain recent results on variable selection in multiple regression to the SPC problem, we propose a novel MSPC approach that has all these three properties. This approach integrates a *least absolute shrinkage and selection operator* (LASSO) test statistic (cf., Tibshirani 1996) into a multivariate EWMA charting scheme for on-line process monitoring, which is easy to implement using the *least angle regression* (LARS) algorithm (cf., Efron et al. 2004). Because of certain good properties of LASSO, it can determine the shift direction "automatically," and can offer balanced protection against various possible shifts. Moreover, due to the sparsity property of LASSO estimators that some of their components would be exactly zero after certain tuning parameters are chosen properly, the proposed control chart can provide an effective diagnostic tool as well.

We describe in detail the proposed MSPC control chart in Section 2, and investigate its numerical performance in Section 3. In Section 4, we demonstrate the proposed method using a real-data example from manufacturing. Several remarks conclude the article in Section 5. Technical details are included in an appendix, available online as supplementary materials.

# 2 Methodology

Our proposed MSPC methodology is described in four parts. Subsection 2.1 provides a brief introduction to the LASSO in the context of variable selection in multiple regression. We demonstrate its connection to the MSPC problem and propose a LASSO-based test statistic in Subsection 2.2. In Subsection 2.3, a Phase II control chart is constructed by integrating the LASSO-based test statistic into the EWMA charting scheme. Finally, its diagnostic property is discussed in Subsection 2.4.

## 2.1 LASSO for regression variable selection

Consider the following multiple linear regression model:

$$y_i = \boldsymbol{Z}_i\boldsymbol{\beta} + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots, n,$$

where $y_i$, $\boldsymbol{Z}_i$, and $\boldsymbol{\beta}$ are the response variable, vector of predictors, and vector of regression coefficients, respectively, and $\varepsilon_i$ are i.i.d. random errors with distribution $N(0, \sigma^2)$. In practice, some predictors do not provide much useful information about the response, given the remaining predictors; thus, they can be deleted from the model. To this end, stepwise or all subset selection procedures along with a model selection criterion (e.g., AIC or BIC) are usually used. Such model selection procedures are practically useful; but they have several limitations, including lack of

stability (cf., Breiman 1996), lack of theoretical properties (cf., Fan and Li 2001), and extensive computation (cf., Tibshirani 1996). To overcome these limitations, some authors suggest using the following penalized least squares method (or equivalently, the penalized likelihood method in the normal error distribution case):

$$\sum_{i=1}^{n}(y_i - \boldsymbol{Z}_i\boldsymbol{\beta})^2 + n\sum_{j=1}^{p} g_{\gamma_j}\left(|\boldsymbol{\beta}^{(j)}|\right),$$

where $\boldsymbol{\beta}^{(j)}$ denotes the $j$-th component of $\boldsymbol{\beta}$, $\gamma_j$ are the penalty parameters (or called the regularization parameters, see Bickel and Li 2006), and $g_{\gamma_j}$ are the penalty functions. When $g_{\gamma_j}(|\boldsymbol{\beta}^{(j)}|) = \gamma|\boldsymbol{\beta}^{(j)}|$ where $\gamma$ is a constant parameter, the corresponding penalized least squares method is called LASSO (cf., Tibshirani 1996). Besides LASSO, another major penalized least squares method is the so-called *smoothly clipped absolute deviation* method (SCAD, Fan and Li 2001). See Fan and Li (2006), Wang and Leng (2007), and Zou and Li (2008) for related discussion. Among other good properties, Fan and Li (2001) have shown that, with the proper choice of the penalty functions and regularization parameters, the penalized likelihood estimators would perform asymptotically as well as if the correct submodel was known, which is referred to as the oracle property in the literature. When the penalty functions are chosen to be continuous, the coefficient estimates that correspond to insignificant predictors would shrink towards 0 as $\gamma$ increases, and these coefficient estimates could be exactly 0 if $\gamma$ is sufficiently large. In addition, the penalized likelihood approach, especially those with LASSO-type penalty functions, enjoys efficient computation using Efron et al.'s (2004) LARS algorithm. See Zou and Li (2008) for detailed discussion about computational issues of LASSO and SCAD.

## 2.2 LASSO-based testing

In the variable selection problem discussed in the previous subsection, an *a priori* assumption is that some components of the coefficient vector $\boldsymbol{\beta}$ are zero. In high-dimensional cases, it is often reasonable to assume that only a few coefficients are non-zero, which is the so-called sparsity characteristic. Analogously, in the process monitoring problem, we assume in MSPC that only a few components in the shift vector $\boldsymbol{\delta} = \boldsymbol{\mu}_1$ are expected to be non-zero when a shift occurs. Therefore, the penalized likelihood method should have a potential in solving the MSPC problem, which is investigated below.

As described in Section 1, we are interested in testing $H_0 : \boldsymbol{\mu} = \mathbf{0}$ versus $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$ in the MSPC testing problem. After a constant term is ignored, the corresponding penalized likelihood function can be written as

$$PL(\boldsymbol{\mu}) = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) + n\sum_{j=1}^{p} g_{\gamma_j}\left(|\boldsymbol{\mu}^{(j)}|\right).$$

If the adaptive LASSO (ALASSO) penalty function (cf., Zou 2006) is used, then it becomes

$$PL(\boldsymbol{\mu}) = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) + n\gamma\sum_{j=1}^{p}\frac{1}{|\bar{\boldsymbol{X}}^{(j)}|^a}|\boldsymbol{\mu}^{(j)}|, \tag{2}$$

where $a > 0$ is a pre-specified constant and $\boldsymbol{\mu}^{(j)}$ is the $j$-th component of $\boldsymbol{\mu}$. As recommended by Zou (2006) and Wang and Leng (2007), we fix $a = 1$ in this paper. The ALASSO penalty used in (2) is slightly different from the traditional one by Tibshirani (1996). The latter uses the same amount of shrinkage for each regression coefficient. Because of that, its coefficient estimator cannot be as efficient as the oracle estimator (Fan and Li 2001), and its model selection results could be inconsistent in certain cases. As a comparison, the ALASSO applies different amount of shrinkage to different regression coefficients. Consequently, its LASSO estimator is asymptotically

unbiased and has certain oracle properties. See Fan and Li (2001), Zou (2006), and Zhao and Yu (2006) for related discussion. The ALASSO estimator of $\boldsymbol{\mu}$ is defined by

$$\widehat{\boldsymbol{\mu}}_\gamma = \arg\min PL(\boldsymbol{\mu}).$$

It can be checked that $\widehat{\boldsymbol{\mu}}_\gamma$ is the same as the one-step sparse estimator in Zou and Li (2008) (see also Bühlmann and Meier 2008), and it is also the same as the least squares approximation estimator in Wang and Leng (2007). Because of its sparsity assumption (i.e., some of its components are exactly zero after $\gamma$ is properly chosen), it is an ideal estimator of the shift direction in the MSPC problem. Then, a LASSO-based test statistic can be defined as

$$\tilde{T}_\gamma = \frac{n(\widehat{\boldsymbol{\mu}}_\gamma' \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{X}})^2}{\widehat{\boldsymbol{\mu}}_\gamma' \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_\gamma}.$$

Theoretically, $\widehat{\boldsymbol{\mu}}_\gamma \neq \mathbf{0}$ almost surely. For completeness, $\tilde{T}_\gamma$ can be defined to be any negative number when $\widehat{\boldsymbol{\mu}}_\gamma = \mathbf{0}$. Obviously, $\tilde{T}_\gamma$ can be regarded as a data-driven version of the test statistics in (1), with the shift direction estimator $\widehat{\boldsymbol{\mu}}_\gamma$ substituting for the pre-specified shift directions $\boldsymbol{d}_j$.

Before $\tilde{T}_\gamma$ can be used for MSPC, the regularization parameter $\gamma$ should be chosen properly, since it plays an important role in balancing robustness and sensitivity of $\tilde{T}_\gamma$ to various shifts. By Theorem 1 in Wang and Leng (2007) and Theorem 5 in Zou and Li (2008), we can obtain results in the following proposition without much difficulty, which provide us some insights about selection of $\gamma$. Without loss of generality, hereafter we assume that, under $H_1$, the first $p_0$ components of $\boldsymbol{\mu}$ are nonzero and the remaining $p - p_0$ components are all zero, where $0 < p_0 \leq p$.

**Proposition 1** *Assume that there are two regularization parameters $\gamma_{1n}$ and $\gamma_{2n}$ satisfying the conditions that $1/(n\gamma_{jn}) = o(1)$, for $j = 1, 2$, and $\gamma_{1n}/\gamma_{2n} = o(1)$. Then,*

*(i) under* $H_0$, $P(\tilde{T}_{\gamma_{2n}} \geq 0)/P(\tilde{T}_{\gamma_{1n}} \geq 0) = o(1)$, *and*

*(ii) under* $H_1$, *if* $\boldsymbol{\mu}$ *satisfies* $\min\{\boldsymbol{\mu}^{(j)}, j \leq p_0\} = O(n^s)$ *and* $\max\{\boldsymbol{\mu}^{(j)}, j \leq p_0\} = O(n^t)$, *where* $-\frac{1}{2} < s \leq t < 0$, *then* $P(\tilde{T}_{\gamma_{jn}} \geq 0) - 1 = o(1)$ *if and only if* $n^{-2t}\gamma_{jn} = o(1)$, *for* $j = 1$ *or 2. Furthermore, when* $n^{-s+\frac{1}{2}}\gamma_{2n} = o(1)$, $\tilde{T}_{\gamma_{1n}}$ *and* $\tilde{T}_{\gamma_{2n}}$ *converge to infinity with the same rate.*

By this proposition, from an asymptotic viewpoint, selection of $\gamma$ should depend on the shift size. For detecting a large shift (i.e., $t$ is large in part (ii) of the proposition), $\gamma$ should be chosen large, and small for detecting a small shift. One can choose $\gamma$ is to use CV, GCV, AIC, BIC, or other model selection criteria. However, our simulation results show that these criteria tailored for estimation often do not produce a powerful test. This finding is not surprising because similar conclusions have been made in the nonparametric regression testing problem. There, the power and size of a typical test would depend on the bandwidth used in regression function estimation (cf., Bickel and Li 2006), and it is recognized that optimal bandwidth for nonparametric curve estimation may not be optimal for testing (cf., Hart 1997).

To overcome this difficulty, we follow the approach proposed by Horowitz and Spokoiny (2001) in the nonparametric testing setup, by combining several values of $\gamma$ to make the resulting test nearly optimal. Let $\Gamma_q = \{\gamma_j, j = 1, \ldots, q\}$ be a set of penalty parameters, where $q$ is a pre-specified constant. Then, the modified penalized test statistic is defined as

$$\tilde{T} = \max_{j=1,\ldots,q} \frac{\tilde{T}_{\gamma_j} - \mathrm{E}(\tilde{T}_{\gamma_j})}{\sqrt{\mathrm{Var}(\tilde{T}_{\gamma_j})}},$$

where $\mathrm{E}(\tilde{T}_{\gamma_j})$ and $\mathrm{Var}(\tilde{T}_{\gamma_j})$ are respectively the mean and variance of $\tilde{T}_{\gamma_j}$ under $H_0$. Ideally, $q$ should be chosen to be the number of non-zero components of $\boldsymbol{\mu}$, and $\gamma_j$ to be the penalty parameter that is good for detecting $\boldsymbol{\mu}$ with $j$ non-zero components.

In practice, however, $\boldsymbol{\mu}$ is unknown. To overcome this difficulty, we propose a simple and efficient method to determine $\Gamma_q$, as described below.

Let us first rewrite the ALASSO-type penalized likelihood (2) as

$$PL(\boldsymbol{\alpha}) = n(\bar{\boldsymbol{X}} - \boldsymbol{\Lambda}\boldsymbol{\alpha})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\Lambda}\boldsymbol{\alpha}) + n\gamma \sum_{i=1}^{p} |\boldsymbol{\alpha}^{(i)}|, \tag{3}$$

where $\boldsymbol{\alpha}^{(i)} = \boldsymbol{\mu}^{(i)}/|\bar{\boldsymbol{X}}^{(i)}|$ and $\boldsymbol{\Lambda} = \text{diag}(|\bar{\boldsymbol{X}}^{(1)}|, \ldots, |\bar{\boldsymbol{X}}^{(p)}|)$. This is exactly a LASSO-type penalized likelihood function. According to Zou et al. (2007), for given $\bar{\boldsymbol{X}}$ in (3), there is a finite sequence

$$\tilde{\gamma}_0 > \tilde{\gamma}_1 > \ldots > \tilde{\gamma}_K = 0 \tag{4}$$

such that (1) for all $\gamma > \tilde{\gamma}_0$, $\widehat{\boldsymbol{\alpha}}_\gamma = 0$, and (2) in the interval $(\tilde{\gamma}_{m+1}, \tilde{\gamma}_m)$, the active set $\mathfrak{B}(\gamma) = \{j : \text{sgn}[\boldsymbol{\alpha}_\gamma^{(j)}] \neq 0\}$ and the sign vector $\boldsymbol{S}(\gamma) = \{\text{sgn}[\boldsymbol{\alpha}_\gamma^{(1)}], \ldots, \text{sgn}[\boldsymbol{\alpha}_\gamma^{(p)}]\}$ are unchanged with $\gamma$. These $\tilde{\gamma}_m$'s are called transition points because the active set changes at each $\tilde{\gamma}_m$. By Efron et al. (2004), the random integer $K$ can be larger than $p$. Thus, we suggest using $\tilde{\gamma}_{m_j^{\text{last}}}$, for $j = 1, \ldots, q$, to construct $\Gamma_q$, where $m_j^{\text{last}}$ is the index of last $\tilde{\gamma}$ in the sequence $\{\tilde{\gamma}_0, \tilde{\gamma}_1, \ldots, \tilde{\gamma}_K\}$ defined in (4) that the corresponding active set contains exactly $j$ elements. By Lemma 7 in Zou et al. (2007), $\tilde{T}_{\tilde{\gamma}_{m_j^{\text{last}}}}$ are well defined, because the "one at a time" condition (cf., Efron et al. 2004) holds almost everywhere, where "one at a time" means that two active sets of two consecutive $\tilde{\gamma}$s differ on at most a single index. The following asymptotic result provides some insight about the use of the transition points for constructing $\Gamma_q$.

**Proposition 2** *Under $H_1$, assume that $\min\{\boldsymbol{\mu}^{(j)}, j \leq p_0\} = O(n^s)$, where $-\frac{1}{2} < s < 0$. Then, the test using $\tilde{T}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}}}$ as a testing statistic would be asymptotically more powerful than the test using $T^2$.*

In practice, $p_0$ is often unknown. So, we suggest combining all $\tilde{\gamma}_{m_j^{\text{last}}}$, for $j =$

$1, \ldots, q$, in our test. The resulting test statistic is

$$\tilde{T}_L = \max_{j=1,\ldots,q} \frac{\tilde{T}_{\tilde{\gamma}_{m_j^{\text{last}}}} - \mathrm{E}(\tilde{T}_{\tilde{\gamma}_{m_j^{\text{last}}}})}{\sqrt{\mathrm{Var}(\tilde{T}_{\tilde{\gamma}_{m_j^{\text{last}}}})}}. \tag{5}$$

Note that $\tilde{T}_{\tilde{\gamma}_{m_p^{\text{last}}}}$ is exactly the $T^2$ test statistic. Therefore, test (5) should share certain properties of the $T^2$ test and certain properties of the likelihood ratio tests constructed when the potential shift has some specific structures (cf., expression (1)).

Next, we make several remarks about the transition points used in (5). First, $\tilde{\gamma}_m$ can be easily determined by the LARS algorithm (cf., Efron et al. 2004). The whole testing procedure (5) involves $O(np + p^3)$ computations, which are about the same as the computational cost of a least squares regression with $p$ covariates. This property is especially desirable for on-line process monitoring. Second, each $\tilde{\gamma}_m$ corresponds to a different active set $\mathfrak{B}_{\tilde{\gamma}_m}$. Thus, by combining different $\tilde{\gamma}_m$'s, test (5) actually combines some cases with different numbers of shifted components, and these cases are determined by the observed data. Third, based on Lemma 1 in Appendix, for any $\gamma \in (\tilde{\gamma}_{m_j^{\text{last}}}, \tilde{\gamma}_{m_j^{\text{last}}-1})$, we have $l(\widehat{\boldsymbol{\mu}}_\gamma) > l(\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_j^{\text{last}}}})$, where $l(\boldsymbol{\mu}) = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$ and $\widehat{\boldsymbol{\mu}}_\gamma = \boldsymbol{\Lambda}\widehat{\boldsymbol{\alpha}}_\gamma$. That is, in the interval $[\tilde{\gamma}_{m_j^{\text{last}}}, \tilde{\gamma}_{m_j^{\text{last}}-1})$, $\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_j^{\text{last}}}}$ is the "optimal" estimator in terms of the likelihood value in the case when $j$ components are shifted, which provides a strong evidence that the transition points should be used in our test.

## 2.3 Phase II MSPC

In this part, we propose a framework to construct Phase II MSPC control charts using the LASSO-based test statistic (5). We consider EWMA-type control charts only in this paper; but, other types of control charts (e.g., CUSUM) can be developed in a similar way.

Let $\boldsymbol{X}_j$ be the $j$-th observed measurement vector collected over time. Then, a

multivariate EWMA (MEWMA) sequence of statistics (cf., Lowry et al. 1992) is often defined as

$$\boldsymbol{U}_j = \lambda \boldsymbol{X}_j + (1-\lambda)\boldsymbol{U}_{j-1}, \quad \text{for } j = 1, 2, \ldots, \tag{6}$$

where $\boldsymbol{U}_0 = \boldsymbol{0}$ and $\lambda$ is a weighting parameter in $(0, 1]$. For each $\boldsymbol{U}_j$, we suggest computing $q$ LASSO estimators $\widehat{\boldsymbol{\mu}}_{j,\tilde{\gamma}_{m_k^{\text{last}}}}$, for $k = 1, 2, \ldots, q$, from the penalized likelihood function

$$(\boldsymbol{U}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{U}_j - \boldsymbol{\mu}) + \gamma \sum_{k=1}^{p} \frac{1}{|\boldsymbol{U}_j^{(k)}|}|\boldsymbol{\mu}^{(k)}|. \tag{7}$$

Our proposed control chart signals a shift if

$$Q_j = \max_{k=1,\ldots,q} \frac{W_{j,\tilde{\gamma}_{m_k^{\text{last}}}} - \mathrm{E}(W_{j,\tilde{\gamma}_{m_k^{\text{last}}}})}{\sqrt{\mathrm{Var}(W_{j,\tilde{\gamma}_{m_k^{\text{last}}}})}} > L, \tag{8}$$

where

$$W_{j,\gamma} = \frac{2-\lambda}{\lambda[1-(1-\lambda)^{2j}]} \frac{(\boldsymbol{U}_j'\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\mu}}_\gamma)^2}{\widehat{\boldsymbol{\mu}}_\gamma'\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\mu}}_\gamma},$$

and $L > 0$ is a control limit chosen to achieve a given in-control (IC) average run length (ARL). In practice, $(2-\lambda)/\{\lambda[1-(1-\lambda)^{2j}]\}$ can be replaced by its asymptotic form $(2-\lambda)/\lambda$, as conventionally done in the EWMA literature. Hereafter, the above control chart is called LASSO-based multivariate EWMA (LEWMA) chart.

In LEWMA (6)–(8), selection of $\lambda$ mainly depends on a target shift (cf. Proposition 1). It should be chosen large if the target shift is large and small otherwise, as in conventional EWMA charts (e.g., Lucas and Saccucci 1990; Prabhu and Runger 1997). Regarding $q$, recall that an implicit *a priori* assumption underlying the LEWMA chart is that a potential shift occurs only in a small number of measurement components. If prior information indicating potential shifts in at most $r$ components, with $1 \le r \le p$, then our numerical studies show that using $q = r + 1$ or $q = r + 2$ provides satisfactory performance in practice (cf., Section 3). When prior information is unavailable,

numerical results (Section 3) show that the LEWMA chart with $q = p$ performs reasonably well in all cases considered there.

In (8), quantities $\mathrm{E}(W_{j,\tilde{\gamma}_{m_k^{\mathrm{last}}}})$ and $\mathrm{Var}(W_{j,\tilde{\gamma}_{m_k^{\mathrm{last}}}})$ are usually unknown. The following result suggests replacing them by their approximations.

**Proposition 3** *When the process is IC,* $\mathrm{E}(W_{j,\tilde{\gamma}_{m_k^{\mathrm{last}}}})$ *and* $\mathrm{Var}(W_{j,\tilde{\gamma}_{m_k^{\mathrm{last}}}})$ *do not depend on* $\lambda$ *and* $j$*.*

By Proposition 3, $\mathrm{E}(W_{j,\tilde{\gamma}_{m_k^{\mathrm{last}}}})$ and $\mathrm{Var}(W_{j,\tilde{\gamma}_{m_k^{\mathrm{last}}}})$ can be approximated by the empirical expectation and variance of $W_{j,\tilde{\gamma}_{m_k^{\mathrm{last}}}}$ computed from simulated IC measurement vectors, since $\boldsymbol{U}_j$ would be the same as $\boldsymbol{X}_j$ when $\lambda = 1$. The control limit $L$ in (8) can also be determined by simulation as follows. First, choose an initial value for $L$. Then, compute the IC ARL of the LEWMA chart based on a large number of replicated simulation runs in which the IC observations are generated from the IC distribution of the process. If the computed IC ARL value is smaller than the nominal IC ARL value, then increase the value of $L$. Otherwise, choose a smaller $L$ value. The above process is repeated until the nominal IC ARL is reached and is within a desired precision. In this process, some numerical searching algorithms, such as the bisection search, can be applied (cf., e.g., Qiu 2008). For given $\lambda$, $\boldsymbol{\Sigma}$ and a desired IC ARL, computation involved in finding $L$ is not difficult, partly due to the fact that the LARS algorithm used in LASSO computation is efficient. For instance, when IC ARL=200 and $p = 15$, it requires about 15 minutes to complete the bisection searching procedure based on 25,000 simulations, using a Pentium-M 1.6MHz CPU. Since this is a one-time computation before the Phase II online process monitoring, it is convenient to accomplish. Fortran code for implementing the proposed procedure is available from the authors upon request.

## 2.4 Post-signal diagnostic

Assume that the LEWMA chart (6)–(8) triggers a mean shift signal at the $k$-th observation. Estimates of the shift locations and identification of the specific measurement components that have shifted will help engineers to eliminate the root causes of the shift in a timely fashion. With respect to the shift location, the generalized maximum likelihood estimation approach for change-point detection can be used (cf., e.g., Zamba and Hawkins 2006). After an estimate of the shift location is obtained, which is denoted as $\widehat{\tau}$, we have $(k-\widehat{\tau})$ out-of-control (OC) observations in which a few measurement components have shifts in their means. Among these $(k-\widehat{\tau})$ observations, a few might actually be IC observations, because $\widehat{\tau}$ is only an estimator of the true shift location $\tau$. After a mean shift is detected and the shift location is estimated, the LASSO methodology can be used for specifying the shifted components, by choosing one of the LASSO estimators (cf., (4)) with a model selection criterion (e.g., $C_p$, GCV, AIC, or BIC). To be specific, our proposed diagnostic algorithm tries to find $\widehat{\boldsymbol{\mu}}_{\gamma^*}$

$$\gamma^* = \arg\min_{\gamma}(k-\widehat{\tau})(\bar{\boldsymbol{X}}_{\widehat{\tau},k} - \widehat{\boldsymbol{\mu}}_{\gamma})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}_{\widehat{\tau},k} - \widehat{\boldsymbol{\mu}}_{\gamma}) + \eta \cdot \widehat{\mathrm{df}}(\widehat{\boldsymbol{\mu}}_{\gamma}), \tag{9}$$

$\bar{\boldsymbol{X}}_{\widehat{\tau},k} = \sum_{i=\widehat{\tau}+1}^{k} \boldsymbol{X}_i/(k-\widehat{\tau})$, $\widehat{\boldsymbol{\mu}}_{\gamma}$ is the ALASSO estimator from $\bar{\boldsymbol{X}}_{\widehat{\tau},k}$, $\eta$ is a parameter, and $\widehat{\mathrm{df}}(\widehat{\boldsymbol{\mu}}_{\gamma})$ is the number of nonzero coefficients in $\widehat{\boldsymbol{\mu}}_{\gamma}$. If AIC is used, then $\eta = 2$. It equals $\ln(k-\widehat{\tau})$ if BIC is used. By Theorem 3 in Zou et al. (2007), $\widehat{\boldsymbol{\mu}}_{\gamma^*}$ is one of $\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_1}, \ldots, \widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_K}$.

In the literature, it is well demonstrated that BIC tends to identify the true sparse model well if the true model is included in the candidate set (cf., e.g., Yang 2005; Zou et al. 2007; Wang et al. 2007). However, unlike the fixed sample case, in the current SPC problem, we usually do not have a large sample (relative to the shift magnitude) to implement the diagnostic procedure, because the whole point of SPC

is to detect the shift (and hence stop the process) as quickly as possible. Based on extensive simulations, we find that the conventional BIC criterion does not perform well in certain cases (not reported here). Instead, the diagnostic procedure (9) would perform well if it uses the risk inflation criterion (RIC) proposed by George and Foster (1994) based on the minimax principle, in which $\eta = 2\ln(p)$. Since some components of $\widehat{\boldsymbol{\mu}}_{\gamma^*}$ would be exactly zero, we can simply take its nonzero components as the shifted measurement components, without making any extra tests that are necessary in most existing diagnostic methods, such as the decomposition of $T^2$ method and the step-down method. As a side note, in cases when estimation of the shift location $\tau$ is not our major concern, after the LEWMA chart gives an OC signal at the $k$-th observation, we can use $\boldsymbol{U}_k$ directly to implement the diagnostic procedure. Namely, we can find $\widehat{\boldsymbol{\mu}}_{\gamma^*}$ so that

$$\gamma^* = \arg\min_{\gamma} \frac{2-\lambda}{\lambda[1-(1-\lambda)^{2k}]}(\boldsymbol{U}_k - \widehat{\boldsymbol{\mu}}_\gamma)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{U}_k - \widehat{\boldsymbol{\mu}}_\gamma) + \eta \cdot \widehat{\mathrm{df}}(\widehat{\boldsymbol{\mu}}_\gamma), \qquad (10)$$

where as before, $\eta = 2\ln(p)$. Compared to (9), procedure (10) is more convenient to use and is faster as well because the LASSO estimators from $\boldsymbol{U}_k$ have been computed in the monitoring process. Our numerical studies show that (9) and (10) provide similar diagnostic results.

# 3 A Simulation Study

We present some simulation results in this section regarding the numerical performance of the proposed LEWMA control chart and the corresponding diagnostic procedure. In the LEWMA chart, the parameter $q$ needs to be chosen, although $q = p$ is always an option. We first investigate the potential effect of $q$ on the performance of the LEWMA chart in the following setting. Let $p = 15$ and the IC measurement distribution be $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\sigma_{ij})$ is chosen to be $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.75^{|i-j|}$,

for $i, j = 1, 2, \ldots, p$. In the LEWMA chart, $\lambda$ is chosen to be 0.2, and its IC ARL is fixed at 500. Two representative shifts are considered. The first one has three shifted components $\delta_2 = 0.25$, $\delta_3 = 0.25$ and $\delta_8 = 0.50$, where $\delta_j$ denotes the shift size of the $j$-th component, and the second one has all eight odd components shifted of the same size 0.25. The two plots in Figure 1 show the OC ARLs of the LEWMA chart when it is used for detecting the two shifts and when $q$ changes. The plots show that the LEWMA chart performs the best when $q = 5$ and $q = 10$, respectively, and its performance is close to optimal when $q = p$. Therefore, $q = p$ is indeed a reasonable choice when no prior information is available about the number of shifted components.

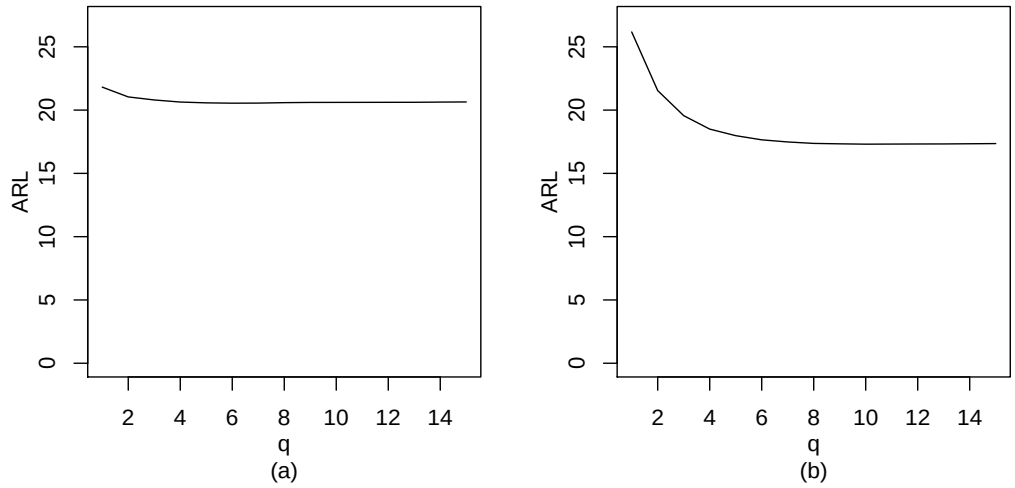


Figure 1: OC ARLs of the LEWMA chart when $q$ changes, $p = 15$ and two representative shifts are detected. Plot (a): the shift has three shifted components $\delta_2 = 0.25$, $\delta_3 = 0.25$ and $\delta_8 = 0.50$. Plot (b): all eight odd components are shifted of 0.25.

Next, we compare the LEWMA chart with the conventional MEWMA chart (i.e., the multivariate EWMA chart using a quadratic charting statistic) and REWMA chart (i.e., the multivariate EWMA chart using $\boldsymbol{V}$). We first consider a low-dimensional case when $p = 5$ that was discussed in Hawkins (1991). In all three charts, $\lambda = 0.2$ and the IC ARL is fixed at 500. The true shift location $\tau$ is fixed at 25, and any

simulations in which shift signals occur before $\tau$ are discarded. In the LEWMA chart, to investigate its sensitivity to selection of $q$, we consider two different $q$ values: $q = 3$ and $q = p = 5$. Twenty different shifts are considered, which include the following cases: shifts in a single component, equal shifts in a pair of components, and shifts in the principle component directions of $\boldsymbol{X}$. As was done in Hawkins (1991), all shifts are scaled so that the largest displacement in $\boldsymbol{V}$ equals one. The simulation results over 20,000 replications are presented in Table 1. Control limits $L$ of the three control charts are included in the last row of this table. In order to assess the overall performance of the charts, besides OC ARLs, we also compute their relative mean index (RMI) values. The RMI index of a control chart is suggested by Han and Tsung (2006), and is defined as

$$\text{RMI} = \frac{1}{N}\sum_{l=1}^{N}\frac{\text{ARL}_{\boldsymbol{\delta}_l} - \text{MARL}_{\boldsymbol{\delta}_l}}{\text{MARL}_{\boldsymbol{\delta}_l}}, \tag{11}$$

where $N$ is the total number of shifts considered, $\text{ARL}_{\boldsymbol{\delta}_l}$ is the OC ARL of the given control chart when detecting shift $\boldsymbol{\delta}_l$, and $\text{MARL}_{\boldsymbol{\delta}_l}$ is the smallest OC ARL among all OC ARL values of the charts considered when detecting shift $\boldsymbol{\delta}_l$. So, $\frac{\text{ARL}_{\boldsymbol{\delta}_l}-\text{MARL}_{\boldsymbol{\delta}_l}}{\text{MARL}_{\boldsymbol{\delta}_l}}$ can be considered as a relative efficiency measure of the given control chart, compared to the best chart, when detecting shift $\boldsymbol{\delta}_l$, and RMI is the average of all such relative efficiency values. By this index, a control chart with a smaller RMI value is considered better in its overall performance. From Table 1, we observe the following results. First, LEWMA offers a good balance between sensitivity and robustness to potential shifts. This can be seen from cases when REWMA outperforms MEWMA in quite large margins or vice versa. In such cases, LEWMA may not be best, but it is always close to the best. Second, in several cases in which LEWMA performs the best, the margins are generally quite small. Third, in terms of the RMI index, LEWMA performs the best overall. Fourth, the OC ARL values of the LEWMA chart when $q = 3$ are close to those when $q = 5$, although the LEWMA chart with $q = 3$

performs slightly better when the number of shifted components is small and slightly worse when the number of shifted components is large. Fifth, the last shift considered in the table is large in magnitude and MEWMA performs the best as expected. The LEWMA chart with $q = p$ also performs well in this case. In Table 1, $\lambda$ is fixed at 0.2 for all three charts. To diminish the possible effect of $\lambda$ on the performance of the three charts, we also compute the optimal OC ARLs with respect to $\lambda$ (i.e., the smallest OC ARLs when $\lambda$ changes) for all three charts in the cases considered in Table 1. The results are presented in Table S1 given in the online supplementary materials. They are slightly better than those in Table 1, as expected; but, the conclusions made from Table 1 are also true in Table S1.

Table 1: OC ARL values of control charts MEWMA, REWMA, and LEWMA in the case when $p = 5$, $\lambda = 0.2$, and IC ARL=500. Numbers in parentheses are standard errors. They are shown as ".00" when they are smaller than 0.01.

| shifts | | | | | MEWMA | REWMA | LEWMA | |
|---|---|---|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | | | $q = 3$ | $q = 5$ |
| 0.91 | 0.00 | 0.00 | 0.00 | 0.00 | 17.3 (.09) | 14.4 (.07) | 14.6 (.07) | 14.9 (.07) |
| 0.00 | 0.36 | 0.00 | 0.00 | 0.00 | 17.0 (.08) | 13.3 (.06) | 13.9 (.06) | 14.3 (.07) |
| 0.00 | 0.00 | 0.48 | 0.00 | 0.00 | 17.3 (.09) | 13.8 (.07) | 14.6 (.07) | 15.0 (.07) |
| 0.00 | 0.00 | 0.00 | 0.34 | 0.00 | 17.2 (.09) | 13.4 (.06) | 14.2 (.07) | 14.6 (.07) |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 17.9 (.09) | 14.1 (.07) | 14.9 (.07) | 15.2 (.07) |
| 0.36 | 0.36 | 0.00 | 0.00 | 0.00 | 15.0 (.07) | 13.5 (.06) | 13.4 (.06) | 13.7 (.06) |
| 0.54 | 0.00 | 0.54 | 0.00 | 0.00 | 12.8 (.06) | 12.9 (.06) | 12.4 (.05) | 12.4 (.05) |
| 0.32 | 0.00 | 0.00 | 0.32 | 0.00 | 15.2 (.07) | 13.1 (.06) | 13.4 (.06) | 13.6 (.06) |
| 0.49 | 0.00 | 0.00 | 0.00 | 0.49 | 13.2 (.06) | 12.6 (.06) | 12.4 (.06) | 12.5 (.06) |
| 0.00 | 0.54 | 0.54 | 0.00 | 0.00 | 8.79 (.03) | 12.3 (.06) | 8.75 (.03) | 8.88 (.03) |
| 0.00 | 1.60 | 0.00 | 1.60 | 0.00 | 3.48 (.00) | 9.03 (.03) | 3.59 (.00) | 3.57 (.00) |
| 0.00 | 0.28 | 0.00 | 0.00 | 0.28 | 13.0 (.06) | 10.5 (.05) | 11.1 (.05) | 11.3 (.05) |
| 0.00 | 0.00 | 0.28 | 0.28 | 0.00 | 13.0 (.06) | 10.5 (.05) | 11.3 (.05) | 11.4 (.05) |
| 0.00 | 0.00 | 1.26 | 0.00 | 1.26 | 4.28 (.01) | 8.70 (.03) | 4.41 (.01) | 4.34 (.01) |
| 0.00 | 0.00 | 0.00 | 0.56 | 0.56 | 8.60 (.03) | 12.2 (.06) | 8.60 (.03) | 8.70 (.03) |
| 0.01 | -0.15 | 0.07 | 0.17 | -0.09 | 15.0 (.07) | 11.5 (.05) | 12.3 (.06) | 12.5 (.06) |
| 0.07 | -0.13 | -0.40 | 0.19 | 0.35 | 10.1 (.04) | 12.4 (.06) | 10.0 (.04) | 10.1 (.04) |
| 0.40 | 0.63 | -0.57 | 0.47 | -0.68 | 4.74 (.01) | 9.28 (.04) | 4.97 (.01) | 4.94 (.01) |
| -1.11 | 0.26 | -0.17 | 0.34 | -0.04 | 11.8 (.05) | 14.6 (.07) | 12.5 (.05) | 12.2 (.05) |
| 2.51 | 7.11 | 7.05 | 7.11 | 7.08 | 1.19 (.00) | 7.52 (.02) | 2.88 (.00) | 1.30 (.00) |
| RMI | | | | | 0.123 | 0.514 | 0.103 | 0.046 |
| Control Limit | | | | | 18.13 | 3.425 | 5.181 | 5.262 |

Next, we consider the case when $p = 15$ and all IC parameters are chosen to be the same as those in the example of Figure 1. In this example, besides the OC ARL, the "worst case" OC ARL is also evaluated. The "worst case" scenario is that, immediately before a shift occurs, the control charting statistic is at the value that would maximize the OC ARL (cf., e.g., Yashchin 1993; Woodall and Mahmoud 2005). The "worst case" OC ARL reflects the worst case performance of a given control chart. The simulation results based on 10,000 replications are presented in Table 2, where 27 shifts with various numbers of shifted components are considered. In this example, we simply choose $q = p$ in the LEWMA chart. The results show that the LEWMA chart is either the best or near the best in all cases, in terms of both OC ARL and "worst case" OC ARL. Even in the last two cases with large shifts in all fifteen components, the LEWMA chart still performs reasonably well. RMI values in the table confirm the claimed advantage of the LEWMA chart that it does offer a balanced protection against various shifts.

In all previous examples, shifts are pre-specified. For a more objective comparison of different control charts, next we consider an example in which $r$ indices are randomly generated from $\{1, \ldots, p\}$ without replacement as the indices of the shifted components, and $r$ shift magnitudes are randomly generated from $N(\delta, \sigma^2)$. A sequence of multivariate normal vectors are generated, with the shift vector specified above. We generate control charts to this sequence and obtain their corresponding run-lengths. This simulation is repeated $N$ times, and we compute the RMI value for each control chart using (11), except that ARLs in that formula are replaced by those run-lengths. The resulting performance measure is referred to as relative run length index (RRLI). By properly choosing $\delta$, $\sigma^2$ and a large $N$, this index would be a good measure of the relative run length performance of the considered control charts. Figure 2 shows the RRLI values (in log-scale) of the LEWMA, REWMA and

Table 2: OC ARL and "worst case" OC ARL values of control charts MEWMA, REWMA, and LEWMA in the case when $p = 15$, $\lambda = 0.2$, and IC ARL=500. All numbers in parentheses are standard errors. For a given shift, components not specified are all 0.

| | OC ARL | | | "worst case" OC ARL | | |
|---|---|---|---|---|---|---|
| shifts | MEWMA | REWMA | LEWMA | MEWMA | REWMA | LEWMA |
| $\delta_1 = 0.50$ | 62.5 (.58) | 39.8 (.35) | 40.8 (.35) | 66.3 (.58) | 43.1 (.35) | 44.7 (.36) |
| $\delta_1 = 1.00$ | 11.2 (.06) | 7.84 (.04) | 8.11 (.04) | 14.9 (.06) | 10.6 (.04) | 10.9 (.04) |
| $\delta_3 = 0.50$ | 34.1 (.29) | 21.5 (.16) | 22.5 (.17) | 38.6 (.29) | 24.9 (.16) | 26.4 (.20) |
| $\delta_3 = 1.00$ | 7.26 (.03) | 5.41 (.02) | 5.62 (.02) | 10.6 (.03) | 7.72 (.02) | 8.02 (.02) |
| $\delta_1 = 0.50; \delta_2 = 0.25$ | 106 (1.01) | 138 (1.30) | 109 (1.00) | 110 (1.02) | 143 (1.34) | 114 (1.03) |
| $\delta_1 = 0.50; \delta_2 = 0.50$ | 57.3 (.52) | 91.0 (.88) | 57.7 (.51) | 61.5 (.53) | 93.4 (.89) | 62.3 (.50) |
| $\delta_1 = 0.50; \delta_2 = 0.75$ | 21.2 (.16) | 19.7 (.14) | 17.8 (.12) | 25.5 (.16) | 23.2 (.14) | 21.3 (.12) |
| $\delta_1 = 0.50; \delta_3 = 0.25$ | 39.3 (.34) | 29.9 (.24) | 30.2 (.23) | 43.8 (.34) | 33.1 (.25) | 33.9 (.25) |
| $\delta_1 = 0.50; \delta_3 = 0.50$ | 18.0 (.13) | 14.9 (.10) | 14.8 (.09) | 22.2 (.13) | 18.3 (.10) | 18.3 (.10) |
| $\delta_1 = 0.50; \delta_3 = 0.75$ | 9.78 (.05) | 7.92 (.04) | 8.02 (.04) | 13.4 (.05) | 10.8 (.04) | 10.9 (.04) |
| $\delta_3 = 0.50; \delta_8 = 0.25$ | 25.5 (.19) | 20.3 (.15) | 20.0 (.14) | 29.9 (.20) | 23.7 (.15) | 23.4 (.14) |
| $\delta_3 = 0.50; \delta_8 = 0.50$ | 14.4 (.09) | 13.8 (.09) | 12.7 (.07) | 18.5 (.09) | 17.1 (.09) | 16.1 (.08) |
| $\delta_3 = 0.50; \delta_8 = 0.75$ | 8.78 (.04) | 8.04 (.04) | 7.70 (.03) | 12.3 (.04) | 10.9 (.04) | 10.5 (.04) |
| $\delta_1 = 0.50; \delta_2 = 0.25; \delta_3 = 0.25$ | 103 (.99) | 127 (1.23) | 104 (1.01) | 106 (.99) | 130 (.24) | 107 (1.02) |
| $\delta_1 = 0.25; \delta_2 = 0.25; \delta_3 = 0.50$ | 55.9 (.52) | 44.0 (.39) | 44.1 (.37) | 60.5 (.53) | 47.3 (.39) | 47.5 (.39) |
| $\delta_2 = 0.50; \delta_3 = 0.25; \delta_8 = 0.25$ | 33.5 (.28) | 34.5 (.29) | 30.2 (.24) | 37.7 (.28) | 38.2 (.30) | 34.1 (.24) |
| $\delta_2 = 0.25; \delta_3 = 0.25; \delta_8 = 0.50$ | 25.4 (.20) | 21.1 (.16) | 20.6 (.14) | 29.8 (.20) | 24.6 (.16) | 24.0 (.15) |
| $\delta_7 = 0.50; \delta_8 = 0.25; \delta_9 = 0.50$ | 24.3 (.18) | 25.4 (.20) | 22.4 (.16) | 28.5 (.19) | 28.5 (.20) | 26.3 (.17) |
| $\delta_7 = 0.25; \delta_8 = 0.75; \delta_9 = 0.50$ | 23.1 (.18) | 60.4 (.54) | 26.5 (.20) | 27.2 (.18) | 61.1 (.55) | 28.9 (.23) |
| $\delta_6 = 0.50; \delta_8 = 0.25; \delta_{10} = 0.50$ | 20.6 (.15) | 18.1 (.13) | 17.1 (.11) | 21.6 (.16) | 20.2 (.16) | 19.0 (.15) |
| $\delta_6 = 0.25; \delta_8 = 0.75; \delta_{10} = 0.50$ | 6.96 (.03) | 6.95 (.03) | 6.49 (.03) | 10.3 (.04) | 9.86 (.03) | 9.55 (.03) |
| Even components shift 0.25 | 15.9 (.11) | 23.8 (.18) | 17.2 (.11) | 20.9 (.11) | 27.6 (.18) | 21.9 (.12) |
| Even components shift 0.50 | 4.60 (.02) | 6.50 (.03) | 4.90 (.02) | 7.50 (.02) | 9.22 (.04) | 7.78 (.02) |
| Odd components shift 0.25 | 17.1 (.11) | 24.4 (.20) | 17.9 (.12) | 20.2 (.10) | 26.8 (.17) | 21.3 (.11) |
| Odd components shift 0.50 | 4.75 (.02) | 6.64 (.03) | 5.03 (.02) | 7.33 (.02) | 9.63 (.04) | 7.67 (.04) |
| Even (odd) components shift 0.50 (0.25) | 13.7 (.09) | 24.3 (.19) | 16.8 (.11) | 17.7 (.09) | 27.4 (.18) | 21.0 (.11) |
| Even (odd) components shift 0.25 (0.50) | 12.2 (.07) | 22.5 (.17) | 15.1 (.10) | 16.4 (.09) | 25.8 (.16) | 18.8 (.10) |
| RMI | 0.164 | 0.251 | 0.040 | 0.165 | 0.180 | 0.030 |
| Control limit | 34.75 | 3.749 | 4.950 | 34.75 | 3.749 | 4.950 |

MEWMA charts when $r$ changes, $p = 15$ (plot (a)), and $p = 30$ (plot (b)). In this example, we choose $\delta = 0.25$, $\sigma = 0.5$, $N$=50,000 in plot (a), $N = 100,000$ in plot (b), and $q = p$ in the LEWMA chart. From the plots, we can see that (i) for small $r$ (e.g., $r \leq 2$), the LEWMA and REWMA charts have similar performance and both significantly outperform the MEWMA chart, (ii) for moderate $r$ (e.g., $2 < r \leq p/2$), the LEWMA chart performs the best, and (iii) for large $r$ (e.g., $r > p/2$), the MEWMA chart performs slightly better than the LEWMA and both of them are much better than the REWMA chart. These findings confirm our statement made earlier that the

LEWMA chart is effective when the sparsity condition holds and it offers a protection against various OC conditions.

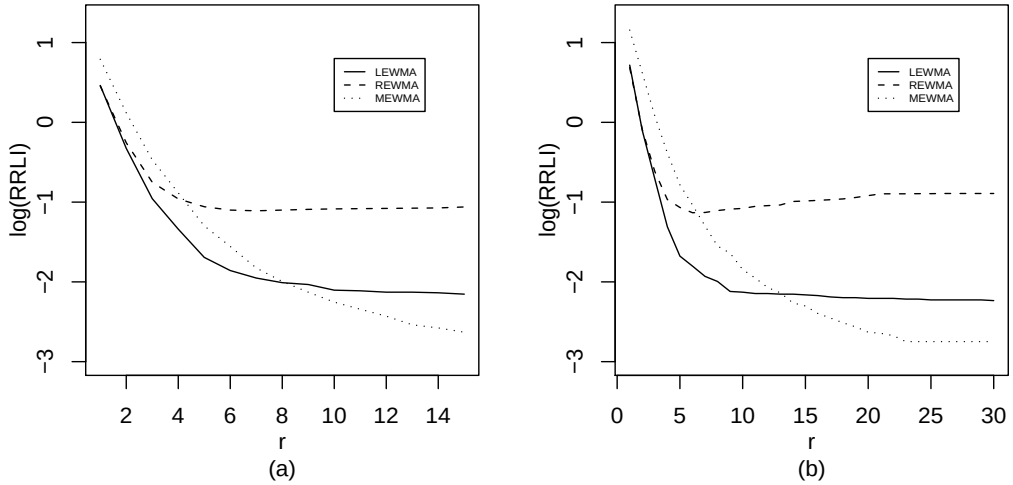


Figure 2: RRLI comparison (in log-scale) of the LEWMA, REWMA and MEWMA charts when $p = 15$ (plot (a)) and $p = 30$ (plot (b)). In the plots, $r$ denotes the number of shifted components.

We now compare the proposed diagnostic procedure (9) with the existing step-down procedure. The step-down procedure depends heavily on the pre-specified type I error probability, which is set to be 0.10 here for illustration. The following three schemes are considered: the MEWMA chart with the step-down procedure (denoted as "step-down"), the MEWMA chart with procedure (9) (denoted as "MEWMA-(9)"), and the LEWMA chart with procedure (9) (denoted as "LEWMA-(9)"). Simulation results in the settings of Tables 1 and 2 are presented in Table S2 (given in the online supplemental materials) and Table 3, respectively. In the two tables, the columns labelled "C" present relative frequencies that the diagnostic procedures identify shifted measurement components correctly, and the columns labelled "I" denote relative frequencies that at least one shifted component is missed and at least one identified shifted component is false. So, for a given diagnostic procedure, it performs better in a given case if its value in column "C" is larger and its value in column "I" is smaller. The results show that the proposed LASSO-based approach has comparable

diagnostic ability to that of the step-down procedure. In many situations, especially when the dimension $p$ is large (cf., Table 3), the LASSO-based approach outperforms the step-down procedure. For instance, in the first row of Table 3, the relative frequency of the MEWMA-(9) scheme is $\widehat{\pi} = 0.54$ (SE=$\sqrt{\widehat{\pi}(1-\widehat{\pi})/10000} = 0.005$) for identifying shifted measurement components correctly. As a comparison, the corresponding relative frequency and its standard error of the step-down scheme are 0.07 and 0.003, respectively. So, the MEWMA-(9) scheme is significantly better in this case. After taking into account its computational advantage, we think that the LASSO-based approach provides a reasonable diagnosis tool for MSPC.

Table 3: Diagnostic results of the proposed procedure (9) and the step-down procedure in the case when $p = 15$, $\lambda = 0.2$, and IC ARL=500. In the table, columns labeled "C" present relative frequencies of identifying shifted measurement components correctly, and the columns labelled "I" denote relative frequencies that at least one shifted components is missed and at least one identified shifted component is false. For a given entry with relative frequency $\widehat{\pi}$, its standard error can be computed by the formula $\sqrt{\widehat{\pi}(1-\widehat{\pi})/10000}$.

| shifts | step-down | | MEWMA-(9) | | LEWMA-(9) | |
|---|---|---|---|---|---|---|
| | C | I | C | I | C | I |
| $\delta_1 = 0.50$ | 0.07 | 0.08 | 0.54 | 0.11 | 0.55 | 0.13 |
| $\delta_1 = 1.00$ | 0.07 | 0.04 | 0.57 | 0.08 | 0.61 | 0.06 |
| $\delta_3 = 0.50$ | 0.05 | 0.09 | 0.53 | 0.12 | 0.55 | 0.12 |
| $\delta_3 = 1.00$ | 0.07 | 0.06 | 0.58 | 0.08 | 0.61 | 0.07 |
| $\delta_1 = 0.50;\ \delta_2 = 0.25$ | 0.07 | 0.22 | 0.34 | 0.24 | 0.32 | 0.28 |
| $\delta_1 = 0.50;\ \delta_2 = 0.50$ | 0.07 | 0.13 | 0.40 | 0.18 | 0.37 | 0.22 |
| $\delta_1 = 0.50;\ \delta_2 = 0.75$ | 0.07 | 0.23 | 0.19 | 0.26 | 0.22 | 0.27 |
| $\delta_1 = 0.50;\ \delta_3 = 0.25$ | 0.06 | 0.28 | 0.22 | 0.37 | 0.27 | 0.34 |
| $\delta_1 = 0.50;\ \delta_3 = 0.50$ | 0.06 | 0.36 | 0.24 | 0.34 | 0.29 | 0.32 |
| $\delta_1 = 0.50;\ \delta_3 = 0.75$ | 0.06 | 0.36 | 0.16 | 0.35 | 0.20 | 0.32 |
| $\delta_3 = 0.50;\ \delta_8 = 0.25$ | 0.05 | 0.44 | 0.15 | 0.35 | 0.20 | 0.34 |
| $\delta_3 = 0.50;\ \delta_8 = 0.50$ | 0.06 | 0.34 | 0.27 | 0.29 | 0.30 | 0.30 |
| $\delta_3 = 0.50;\ \delta_8 = 0.75$ | 0.07 | 0.35 | 0.23 | 0.29 | 0.27 | 0.29 |
| $\delta_1 = 0.50;\ \delta_2 = 0.25;\ \delta_3 = 0.25$ | 0.08 | 0.32 | 0.19 | 0.30 | 0.18 | 0.34 |
| $\delta_1 = 0.25;\ \delta_2 = 0.25;\ \delta_3 = 0.50$ | 0.06 | 0.48 | 0.04 | 0.36 | 0.06 | 0.36 |
| $\delta_2 = 0.50;\ \delta_3 = 0.25;\ \delta_8 = 0.25$ | 0.06 | 0.55 | 0.11 | 0.42 | 0.12 | 0.43 |
| $\delta_2 = 0.25;\ \delta_3 = 0.25;\ \delta_8 = 0.50$ | 0.05 | 0.59 | 0.05 | 0.43 | 0.07 | 0.42 |
| $\delta_7 = 0.50;\ \delta_8 = 0.25;\ \delta_9 = 0.50$ | 0.05 | 0.51 | 0.05 | 0.42 | 0.06 | 0.44 |
| $\delta_7 = 0.25;\ \delta_8 = 0.75;\ \delta_9 = 0.50$ | 0.09 | 0.30 | 0.28 | 0.25 | 0.24 | 0.30 |
| $\delta_6 = 0.50;\ \delta_8 = 0.25;\ \delta_{10} = 0.50$ | 0.04 | 0.62 | 0.08 | 0.47 | 0.13 | 0.43 |
| $\delta_6 = 0.25;\ \delta_8 = 0.75;\ \delta_{10} = 0.50$ | 0.07 | 0.55 | 0.12 | 0.43 | 0.09 | 0.47 |

In the previous examples, the IC and OC distributions are assumed multivariate normal and the IC parameters are all known. In the online supplementary materials, a numerical example is presented to study the performance of the three control charts LEWMA, MEWMA and REWMA when these assumptions are violated, which shows that the three control charts are all affected by violations of the assumptions and their sensitivity to the violations is comparable.

# 4 An Application

We illustrate the proposed method using a dataset from an aluminum smelter that produces metallic aluminum from dissolved alumina through a chemical reaction process. The dataset contains five variables–the contents of $SiO_2$, $Fe_2O_3$, MgO, CaO, and $Al_2O_3$ (labeled as $X_1$, $X_2$, $X_3$, $X_4$, and $X_5$) in the cryolite/alumina mixture that goes through the chemical reaction process. These measures are important to the quality of metallic aluminum produced. More detailed information about this data can be found in Qiu and Hawkins (2001) and Zamba and Hawkins (2006).

The dataset has 189 vectors. Figure 3 shows the time series of the raw data. Similar to Qiu and Hawkins (2001), we first pre-whiten the dataset by removing substantial autocorrelation in the original measurements. Then, since our method relies on the joint normality of the 5 variables, we transform $X_1$ and $X_5$ to the log scale, as was done in Zamba and Hawkins (2006). For the transformed data, Mardia's multivariate normality test (Mardia 1970) gives the p-value of about 0.27, which suggests that the multivariate normality assumption is valid in this case. Some summary statistics of the transformed data are presented in Table S3 included in the online supplementary materials, which shows that the correlation matrix $\widehat{\mathbf{\Sigma}}$ contains several large entries. Therefore, multivariate control charts should be more appropriate to

use in this example than univariate control charts.

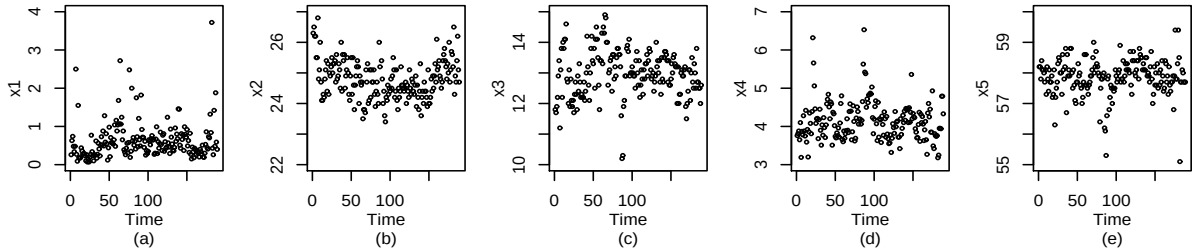


Figure 3: The raw aluminum smelter data.

We then apply the three control charts LEMMA, REWMA and MEWMA to this dataset. In the LEMMA chart, we set $\lambda = 0.2$, $q = 5$, and IC ARL=500. Its control limit is computed to be $L = 5.143$, using estimated IC parameters from the first 75 vectors of the data. Note that a calibration sample of this size is smaller than one would like to fully determine the IC measurement distribution. Based on the simulation results shown in Figure S1 in the online supplementary materials, the actual IC ARLs of the three procedures could be quite different from their nominal IC ARLs in such cases. So, this example is just for illustrating the use of the proposed method in a real-world setting. In practice, a larger calibration sample (say 500 vectors; cf., Figure S1) is desired. Figure 4 shows the resulting LEWMA chart (solid curve connecting the dots), along with its control limit (solid horizontal line). The corresponding REWMA chart (dashed curve connecting little circles), MEWMA chart (dotted curve connecting little diamonds), and their control limits (horizontal dashed and dotted lines, respectively) are also presented in the figure. All three charts signal at the 83rd observation and remain above their control limits in the remainder of the sequence, but the LEMMA chart gives a more convincing evidence of the shift because its charting statistic values after the 83rd observation are much higher than its control limit, compared to the other two charts. We also applied the anti-rank CUSUM chart proposed by Qiu and Hawkins (2001) to this dataset; with a reference value of 0.5, that chart does not give a signal until the 91st observation.

We think that the anti-rank chart is relatively inefficient here because it does not use the normality assumption that approximately is satisfied for the transformed data. Then, we use the proposed diagnostic procedure (9) to identify the shifted variables, after the change-point estimate of $\widehat{\tau} = 81$ is obtained. Table 4 tabulates the resulting five LASSO estimates $\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_j}$, for $j = 1, \ldots, 5$, and the corresponding values of $D_j = 2(\bar{\boldsymbol{X}}_{81,83} - \widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_j})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}_{81,83} - \widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_j}) + 2\ln(5) \cdot j$ (denoted as D-value in Table 4). These values indicate that the shift may have occurred in the last two variables CaO and $Al_2O_3$.

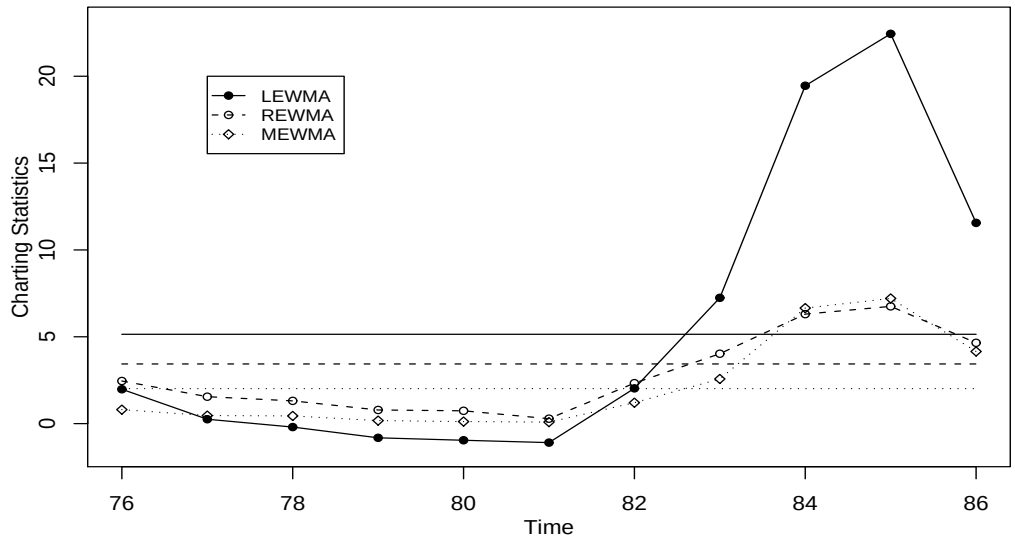


Figure 4: LEWMA, REWMA and MEWMA control charts for monitoring the aluminum smelting process. The solid, dashed and dotted horizontal lines indicate their control limits, respectively.

Table 4: Diagnostic results of procedure (9) about the aluminum smelter data.

| $j$ | $\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_j}$ | | | | | D-value |
|---|---|---|---|---|---|---|
| 1 | 0.000 | 0.000 | 0.000 | 0.000 | -2.383 | 17.332 |
| 2 | 0.000 | 0.000 | 0.000 | 0.977 | -2.970 | 13.914 |
| 3 | 0.248 | 0.000 | 0.000 | 1.134 | -3.112 | 15.388 |
| 4 | 0.799 | 0.000 | -0.423 | 1.747 | -3.671 | 14.552 |
| 5 | 1.263 | 0.778 | -1.184 | 1.984 | -3.689 | 16.094 |

# 5 Concluding Remarks

In this paper, we have proposed a new framework for MSPC using LASSO estimators. A LASSO-based multivariate test statistic is proposed, and it is combined with a multivariate EWMA procedure for Phase II process monitoring. Because of certain good properties of the LASSO estimators, the proposed control chart has a balanced protection against various shifts. We also proposed a LASSO-based diagnostic procedure that is convenient to use. Numerical studies show that the proposed monitoring and diagnosis approaches would be effective in industrial applications.

The current version of the proposed LEWMA chart is designed for detecting mean shifts only, and it is based on the assumptions that the multivariate observations are independent of each other, they follow normal distributions, and their IC distribution is known. We believe that, after certain modifications to the LASSO penalized likelihood function (2), the proposed method should be able to handle cases in which monitoring both the mean vector and the covariance matrix is of interest (cf., e.g., Huwang et al. 2007). In such cases, we would face a much higher dimensional problem (i.e., $[p(p+3)/2]$-dimensional case); we expect that the potential improvement by LEWMA, compared to its peers, would be even larger. In addition, the multivariate normality assumption could be readily relaxed to the elliptical distribution assumption (cf., Liu and Singh 1993). It requires much future research to (i) further extend the proposed approach to nonparametric multivariate SPC cases (cf., Liu 1995; Qiu and and Hawkins 2001, 2003) and (ii) accommodate possible correlation among observations observed at different time points. Future research includes a self-starting version of the LEWMA chart (cf., Hawkins and Maboudou-Tchao 2007) and a study of its properties in cases when the IC parameters in the measurement distribution are unknown (cf., Jones et al. 2001).

**Acknowledgments** The authors thank the editor, the associate editor, and two referees for many constructive comments and suggestions which greatly improved the quality of the paper. This research is supported in part by the grant DMS-0721204 from NSF of USA and the grant 10771107 from NNSF of China.

**References**:

Bickel, P. J., and Li, B. (2006), "Regularization in Statistics," *Test*, 15, 271–344.

Breiman, L. (1996), "Heuristics of Instability and Stabilization in Model Selection," *The Annals of Statistics*, 24, 2350–2383.

Bühlmann, P., and Meier, L. (2008), "Discussion of 'One-step Sparse Estimates in Nonconcave Penalized Likelihood Models'," *The Annals of Statistics*, 36, 1534–1541.

Croisier, R. B. (1988), "Multivariate Generalizations of Cumulative Sum Quality-Control Schemes," *Technometrics*, 30, 243–251.

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.

Fan, J., and Li, R. (2006), "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery," in *Proceedings of the Madrid International Congress of Mathematicians 2006*.

George, E. I., and Foster, D. P. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

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.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*, Springer, New York.

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

Hawkins, D. M., (1993), "Regression Adjustment for Variables in Multivariate Quality Control," *Journal of Quality Technology*, 25, 170–182.

Hawkins, D. M., and Maboudou-Tchao, E. M. (2007), "Self-Starting Multivariate Exponentially Weighted Moving Average Control Charting," *Technometrics*, 49, 199–209.

Healy, J. D. (1987), "A Note on Multivariate CUSUM Procedure," *Technometrics*, 29, 409–412.

Horowitz, J. L., and Spokoiny, V. G. (2001), "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative," *Econometrica*, 69, 599–631.

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

Jones, L. A., Champ, C. W., and Rigdon, S. E. (2001), "The Performance of Exponentially Weighted Moving Average Charts with Estimated Parameters," *Technometrics*, 43, 156–167.

Knight, K., and Fu, W. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378.

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.

Liu, R. (1995), "Control Charts for Multivariate Processes," *Journal of the American Statistical Association*, 90, 1380–1388.

Liu, R., and Singh, K. (1993), "A Quality Index Based on Data Depth Tests," *Journal of the American Statistical Association*, 88, 252–260.

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.

Lucas, J. M., and Saccucci, M. S. (1990), "Exponentially Weighted Moving Average Control Scheme Properties and Enhancements," *Technometrics*, 32, 1–29.

Mardia, K.V. (1970), "Measures of Multivariate Skewness and Kurtosis With Applications," *Biometrika*, 57, 519–530.

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., 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.

Osborne, M., Presnell, B., and Turlach, B. (2000), "A New Approach to Variable Selection in Least Squares Problems," *IMA Journal of Numerical Analysis*, 20, 389–403.

Pignatiello, J. J., and Runger, G. C. (1990), "Comparison of Multivariate Cusum Charts," *Journal of Quality Technology*, 22, 173–186.

Prabhu, S. S., and Runger, G. C. (1997), "Designing a Multivariate EWMA Control Chart," *Journal of Quality Technology*, 29, 8–15.

Qiu, P. (2008), "Distribution-Free Multivariate Process Control Based on Log-Linear Modeling," *IIE Transactions*, 40, 664–677.

Qiu, P., and Hawkins, D. M. (2001), "A Rank-Based Multivariate CUSUM Procedure," *Technometrics*, 43, 120-132.

Qiu, P., and Hawkins, D. M. (2003), "A nonparametric multivariate CUSUM procedure for detecting shifts in all directions," *JRSS (Series D) - The Statistician*, 52, 151–164.

Runger, G. C., and Prabhu, S. S. (1996), "A Markov Chain Model for the Multivariate Exponentially Weighted Moving Averages Control Chart," *Journal of the American Statistical Association*, 91, 1701–1706.

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.

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

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., and Mahmoud, M. A. (2005), "The Inertial Properties of Quality Control Charts," *Technometrics*, 47, 425–436.

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

Yashchin, E. (1993), "Statistical Control Schemes: Methods, Applications and Generalizations," *International Statistical Review*, 61, 41-66.

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

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

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.

Zou, H., and Li, R. (2008), "One-step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics* (with discussion), 36, 1509–1533.

**Supplemental file for the paper titled "Multivariate Statistical Process Control Using LASSO"**

# 1 Some Supplemental Simulation Results

In Table 1 of the printed paper, $\lambda$ is fixed at 0.2 for all three charts MEWMA, REWMA, and LEWMA. To diminish the possible effect of $\lambda$ on the performance of the three charts, we also compute the optimal OC ARLs with respect to $\lambda$ (i.e., the smallest OC ARLs when $\lambda$ changes) for all three charts in the cases considered in Table 1. The results are presented in Table S1.

Table S1: Optimal OC ARL values of control charts MEWMA, REWMA, and LEWMA, in the case when $p = 5$ and IC ARL=500. Numbers in parentheses are standard errors. They are shown as ".00" when they are smaller than 0.01.

| shifts | | | | | MEWMA | REWMA | LEWMA | |
|---|---|---|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | | | $q = 3$ | $q = 5$ |
| 0.91 | 0.00 | 0.00 | 0.00 | 0.00 | 14.7 (.06) | 12.8 (.05) | 13.0 (.05) | 13.2 (.05) |
| 0.00 | 0.36 | 0.00 | 0.00 | 0.00 | 14.5 (.06) | 12.2 (.05) | 12.8 (.05) | 13.1 (.05) |
| 0.00 | 0.00 | 0.48 | 0.00 | 0.00 | 14.8 (.06) | 12.5 (.05) | 13.2 (.05) | 13.5 (.05) |
| 0.00 | 0.00 | 0.00 | 0.34 | 0.00 | 14.7 (.06) | 12.3 (.05) | 12.9 (.05) | 13.3 (.05) |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 15.0 (.06) | 12.7 (.05) | 13.4 (.05) | 14.0 (.05) |
| 0.36 | 0.36 | 0.00 | 0.00 | 0.00 | 13.3 (.05) | 12.3 (.05) | 12.2 (.05) | 12.5 (.05) |
| 0.54 | 0.00 | 0.54 | 0.00 | 0.00 | 11.9 (.04) | 12.0 (.04) | 11.6 (.04) | 11.7 (.04) |
| 0.32 | 0.00 | 0.00 | 0.32 | 0.00 | 13.4 (.05) | 12.0 (.04) | 12.3 (.05) | 12.4 (.05) |
| 0.49 | 0.00 | 0.00 | 0.00 | 0.49 | 12.0 (.04) | 11.6 (.04) | 11.5 (.04) | 11.6 (.04) |
| 0.00 | 0.54 | 0.54 | 0.00 | 0.00 | 8.35 (.03) | 11.4 (.04) | 8.40 (.03) | 8.44 (.03) |
| 0.00 | 1.60 | 0.00 | 1.60 | 0.00 | 3.07 (.01) | 7.93 (.03) | 3.21 (.01) | 3.18 (.01) |
| 0.00 | 0.28 | 0.00 | 0.00 | 0.28 | 11.9 (.04) | 10.2 (.04) | 10.7 (.04) | 10.9 (.04) |
| 0.00 | 0.00 | 0.28 | 0.28 | 0.00 | 11.9 (.04) | 10.2 (.04) | 10.8 (.04) | 10.9 (.04) |
| 0.00 | 0.00 | 1.26 | 0.00 | 1.26 | 3.89 (.01) | 7.68 (.02) | 4.03 (.01) | 4.01 (.01) |
| 0.00 | 0.00 | 0.00 | 0.56 | 0.56 | 8.05 (.03) | 11.3 (.04) | 8.07 (.03) | 8.14 (.03) |
| 0.01 | -0.15 | 0.07 | 0.17 | -0.09 | 13.2 (.05) | 10.9 (.04) | 11.6 (.04) | 11.7 (.04) |
| 0.07 | -0.13 | -0.40 | 0.19 | 0.35 | 9.79 (.03) | 11.5 (.04) | 9.71 (.03) | 9.79 (.03) |
| 0.40 | 0.63 | -0.57 | 0.47 | -0.68 | 4.62 (.01) | 9.01 (.04) | 4.85 (.01) | 4.82 (.01) |
| -1.11 | 0.26 | -0.17 | 0.34 | -0.04 | 11.1 (.05) | 12.9 (.07) | 11.7 (.05) | 11.4 (.05) |
| 2.51 | 7.11 | 7.05 | 7.11 | 7.08 | 1.00 (.00) | 7.45 (.02) | 2.40 (.00) | 1.09 (.00) |
| | | RMI | | | 0.086 | 0.556 | 0.100 | 0.045 |

In the printed paper, we compare the proposed diagnostic procedure (9) with the existing step-down procedure in the setting of Table 2, and the results are presented in Table 3. In Table S2, we provide the comparison results in the setting of Table 1. All labels in Table S2 are the same as those in Table 3 of the printed paper.

Table S2: Diagnostic results of the proposed procedure (12) and the step-down procedure in the case when $p = 5$, $\lambda = 0.2$, and IC ARL=500. In the table, columns labeled "C" present relative frequencies of identifying shifted measurement components correctly, and the columns labelled "I" denote relative frequencies that at least one shifted component is missed while at least one identified shifted component is false.

| shifts | | | | | step-down | | LEWMA-(9) | | MEWMA-(9) | |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | C | I | C | I | C | I |
| 0.91 | 0.0 | 0.0 | 0.0 | 0.0 | 0.36 | 0.02 | 0.51 | 0.01 | 0.48 | 0.02 |
| 0.0 | 0.36 | 0.0 | 0.0 | 0.0 | 0.30 | 0.24 | 0.43 | 0.23 | 0.43 | 0.21 |
| 0.0 | 0.0 | 0.48 | 0.0 | 0.0 | 0.34 | 0.16 | 0.47 | 0.13 | 0.45 | 0.12 |
| 0.0 | 0.0 | 0.0 | 0.34 | 0.0 | 0.30 | 0.25 | 0.44 | 0.22 | 0.43 | 0.22 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.46 | 0.33 | 0.17 | 0.46 | 0.13 | 0.44 | 0.13 |
| 0.36 | 0.36 | 0.0 | 0.0 | 0.0 | 0.22 | 0.44 | 0.18 | 0.38 | 0.18 | 0.38 |
| 0.54 | 0.0 | 0.54 | 0.0 | 0.0 | 0.34 | 0.29 | 0.26 | 0.27 | 0.26 | 0.27 |
| 0.32 | 0.0 | 0.0 | 0.32 | 0.0 | 0.20 | 0.46 | 0.15 | 0.40 | 0.16 | 0.39 |
| 0.49 | 0.0 | 0.0 | 0.0 | 0.49 | 0.30 | 0.33 | 0.24 | 0.29 | 0.24 | 0.30 |
| 0.0 | 0.54 | 0.54 | 0.00 | 0.0 | 0.37 | 0.31 | 0.39 | 0.28 | 0.39 | 0.28 |
| 0.0 | 1.60 | 0.0 | 1.60 | 0.0 | 0.57 | 0.07 | 0.58 | 0.05 | 0.57 | 0.05 |
| 0.0 | 0.28 | 0.0 | 0.0 | 0.28 | 0.23 | 0.44 | 0.24 | 0.41 | 0.25 | 0.41 |
| 0.0 | 0.0 | 0.28 | 0.28 | 0.0 | 0.23 | 0.44 | 0.23 | 0.40 | 0.26 | 0.41 |
| 0.0 | 0.0 | 1.26 | 0.0 | 1.26 | 0.56 | 0.11 | 0.54 | 0.06 | 0.51 | 0.08 |
| 0.0 | 0.0 | 0.0 | 0.56 | 0.56 | 0.37 | 0.30 | 0.40 | 0.27 | 0.39 | 0.27 |
| 0.01 | -0.15 | 0.07 | 0.17 | -0.09 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 |
| 0.07 | -0.13 | -0.40 | 0.19 | 0.35 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 |
| 0.40 | 0.63 | -0.57 | 0.47 | -0.68 | 0.00 | 0.00 | 0.04 | 0.00 | 0.04 | 0.00 |
| -1.11 | 0.26 | -0.17 | 0.34 | -0.04 | 0.03 | 0.00 | 0.03 | 0.00 | 0.03 | 0.00 |
| 2.51 | 7.11 | 7.05 | 7.11 | 7.08 | 0.85 | 0.00 | 0.89 | 0.00 | 0.88 | 0.00 |

In all the simulation examples in the printed paper, it is assumed that the IC and OC distributions are multivariate normal and the IC parameters are known. Next, we study the performance of LEWMA when these assumptions are violated. To this end, we use the example of Table 1 in which $p = 5$ and the nominal IC ARL=500. First, we consider the case when the IC and OC distributions are multivariate normal, but the IC mean and covariance matrix are unknown and need to be estimated from an IC dataset. Figure S1(a) shows the IC ARLs, computed from 20,000 replicated simulations, of the three control charts LEWMA, MEWMA and REWMA when the IC parameters are computed from an IC dataset with varying sample size. From the plot, it can be seen that (i) when the sample size of the IC dataset is relatively small, the actual IC ARLs of the three charts are all quite far away from the nominal level 500, (ii) when the sample size of the IC dataset increases, such biases decrease, and (iii) the biases in IC ARL of the three charts are similar, although MEWMA has a little larger bias when the sample size of the IC dataset is larger than 200. Next, we consider a case when the IC distribution is not normal, which is described below. Multivariate observations are generated in the same way as those in the example of Table 1 except that the first two components of the observations are first transformed by the function $\psi_d^{-1}(\Phi(\cdot))$ and then normalized to have mean 0 and standard deviation 1, where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF) and $\psi_d^{-1}(\cdot)$ is the inverse of the chi-square CDF with $d$ degrees of freedom. Therefore, the first two components of the multivariate observations would follow a normalized chi-square distribution with $d$ degrees of freedom. When $d$ is larger, this distribution is closer to the standard normal distribution. Figure S1(b) shows the IC ARLs, computed from 20,000 replicated simulations, of the three control charts in this case. From the plot, we can see that the three charts are affected similarly by the violation of the normality assumption.

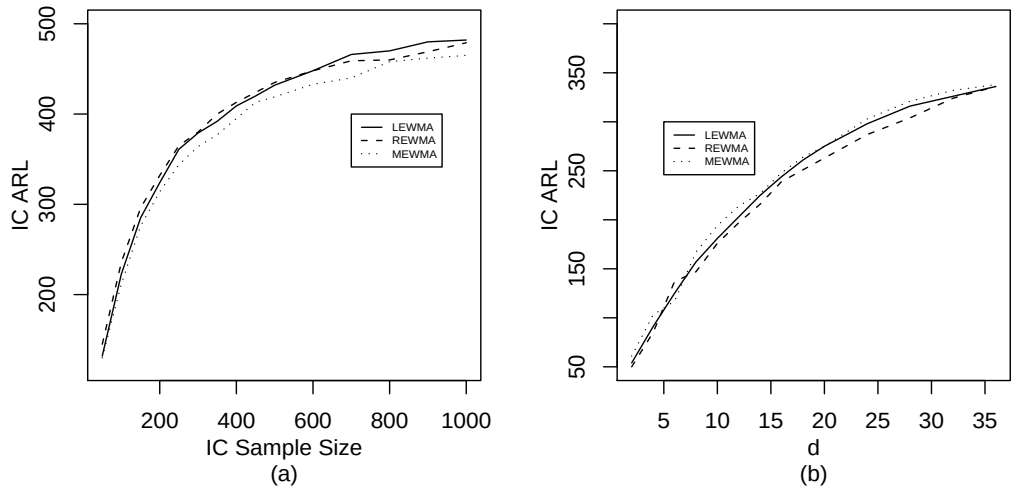


Figure S1: (a): IC ARLs of the LEWMA, REWMA and MEWMA charts when $p = 5$ and the IC mean and covariance matrix are estimated from an IC sample. (b): IC ARLs of the three charts when $p = 5$ and the first two components of the observations follow a normalized chi-square distribution with $d$ degrees of freedom. In both cases, the nominal IC ARL is fixed at 500.

Table S3 below presents some summary statistics of the transformed and pre-whitened aluminum smelter data.

Table S3: Table S3. Summary statistics of the transformed, pre-whitened aluminum smelter data

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|
| Sample means | | | | |
| -0.1083 | 0.0162 | 0.0573 | -0.0085 | 1.0893 |
| Sample standard deviations | | | | |
| 0.1727 | 0.3307 | 0.5313 | 0.1897 | 0.0336 |
| Sample correlation matrix $\widehat{\mathbf{\Sigma}}$ | | | | |
| 1.0000 | 0.2629 | -0.0912 | -0.0704 | -0.0432 |
| 0.2629 | 1.0000 | -0.7201 | -0.0667 | 0.3150 |
| -0.0912 | -0.7201 | 1.0000 | -0.1895 | -0.0051 |
| -0.0704 | -0.0667 | -0.1895 | 1.0000 | -0.3000 |
| -0.0432 | 0.3150 | -0.0051 | -0.3000 | 1.0000 |

# 2 Some Technical Details

In this part, we provide proofs of propositions in Section 2 of the printed paper. For ease of presentation, we will use the following notation. Assume that $\mathbf{M}$ is a matrix with $p$ columns, and $\mathfrak{D}$ is a subset of $\{1, 2, \ldots, p\}$. Then, matrix $\mathbf{M}_{\mathfrak{D}}$ consists of all columns of $\mathbf{M}$ whose indices are in $\mathfrak{D}$. Similarly, $\boldsymbol{\beta}_{\mathfrak{D}}$ denotes a vector consisting of elements of a $p$-dimensional vector whose indices are in $\mathfrak{D}$. Moreover, $\mathbf{M}^{(i,j)}$ denotes the $(i, j)$-th element of matrix $\mathbf{M}$.

**Proof of Proposition 1**

(i) Since $P(\tilde{T}_{\gamma_{in}} \geq 0) = 1 - P(\widehat{\boldsymbol{\mu}}_{\gamma_{in}} = \mathbf{0})$, to prove part (i) of Proposition 1, we only need to show that

$$[1 - P(\widehat{\boldsymbol{\mu}}_{\gamma_{2n}} = \mathbf{0})]/[1 - P(\widehat{\boldsymbol{\mu}}_{\gamma_{1n}} = \mathbf{0})] \to 0.$$

Rewrite (3) as

$$PL(\boldsymbol{\alpha}) = (\boldsymbol{\Sigma}^{-\frac{1}{2}}\sqrt{n}\bar{\boldsymbol{X}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\sqrt{n}\boldsymbol{\Lambda}\boldsymbol{\alpha})'(\boldsymbol{\Sigma}^{-\frac{1}{2}}\sqrt{n}\bar{\boldsymbol{X}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\sqrt{n}\boldsymbol{\Lambda}\boldsymbol{\alpha}) + n\gamma \sum_{i=1}^{p} |\boldsymbol{\alpha}^{(i)}|. \tag{A.1}$$

It is straightforward to verify that (A.1) satisfies the setup of the LASSO optimization in Knight and Fu (2000). Hence, by using (10) in Knight and Fu (2000) or the Karush-Kuhn-Tucker (KKT) optimality condition (cf., Osborne et al. 2000), we have the conclusion that $\widehat{\boldsymbol{\mu}}_{\gamma_{in}} = \mathbf{0}$ if and only if

$$| - 2(\sqrt{n}\boldsymbol{\Lambda})\boldsymbol{\Sigma}^{-1}(\sqrt{n}\bar{\boldsymbol{X}})| \leq n\gamma_{in}, \tag{A.2}$$

where "$\leq$" holds for each component. Let $\boldsymbol{R} = 2(\sqrt{n}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1/2})(\sqrt{n}\boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{X}})$. Then, under $H_0$, the $i$-th component of $|\boldsymbol{R}|$ can be expressed as $|\boldsymbol{R}^{(i)}| = |c_i a_i b_i|$, where $c_i = 2[\boldsymbol{\Sigma}^{(i,i)}]^{\frac{1}{2}}[(\boldsymbol{\Sigma}^{-1})^{(i,i)}]^{\frac{1}{2}}$, and $a_i$ and $b_i$ are standard Normal random variables. Thus,

we have

$$\begin{aligned}\frac{P(\widehat{\boldsymbol{\mu}}_{\gamma_{1n}} \neq \mathbf{0})}{P(\widehat{\boldsymbol{\mu}}_{\gamma_{2n}} \neq \mathbf{0})} &\geq \frac{P(|\boldsymbol{R}^{(j)}| > n\gamma_{1n}, \text{ for some } j)}{\sum_{i=1}^{p} P(|\boldsymbol{R}^{(i)}| > n\gamma_{2n})} \\ &> \frac{P(|a_j| > \sqrt{n\gamma_{1n}/\min_i |c_i|}) \cdot P(|b_j| > \sqrt{n\gamma_{1n}/\min_i |c_i|})}{p \cdot [P(|b_j| > \sqrt{n\gamma_{2n}/\max_i |c_i|}) + P(|a_j| > \sqrt{n\gamma_{2n}/\max_i |c_i|})]} \\ &= C\frac{\gamma_{1n}}{\gamma_{2n}} \exp^{n\gamma_{2n}/(2\max_i |c_i|) - n\gamma_{1n}/\min_i |c_i|}(1 + o(1)) \to \infty,\end{aligned}$$

where $C$ is a positive constant. In the above expression, we have used the facts that $1 - \Phi(t) \approx \frac{\phi(t)}{t}$ for large $t$ and the conditions that $n\gamma_{in} \to \infty$ for $i = 1, 2$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and the density function of the standard Normal distribution.

(ii) Apparently, under $H_1$, $P(\tilde{T}_{\gamma_{in}} \geq 0) \to 1$ if and only if $P(\widehat{\boldsymbol{\mu}}_{\gamma_i} = \mathbf{0}) \to 0$. By (A.2) and the facts that $\boldsymbol{\Lambda} = O_p(n^t)$ and $\boldsymbol{\Sigma}^{-1}\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{-1})$, we obtain the necessary and sufficient condition that $n^{-2t}\gamma_{in} \to 0$ in part (ii) of Proposition 1. Next, we show that

$$\begin{aligned}\left(\widehat{\boldsymbol{\mu}}_{\gamma_{in}}\right)_a &= \bar{\boldsymbol{X}}_a + O_p(n^{-1/2}) \\ P\left[\left(\widehat{\boldsymbol{\mu}}_{\gamma_{in}}\right)_b = 0\right] &\to 1, \quad \text{for } i = 1, 2,\end{aligned}$$

where $\boldsymbol{\beta}_a$ and $\boldsymbol{\beta}_b$ are sub-vectors of a $p$-dimensional vector $\boldsymbol{\beta}$ consisting of the first $p_0$ and the remaining $p - p_0$ elements, respectively. These results can be obtained in a similar way to Theorems 1-3 in Wang and Leng (2007), or to Theorems 1-3 in Fan and Li (2001). The only difference is that $\boldsymbol{\mu}$ is of order o(1) here. The technical arguments in the proof of Theorems 1-3 in Wang and Leng (2007) continue to hold in the current setting, and we only need to change the condition to $n^{-s+\frac{1}{2}}\gamma_{in} \to 0$ in the current setting (cf., the last term of (A.2) in Wang and Leng 2007). For simplicity, details of these arguments are omitted here. Based on these results, it can be easily checked that

$$\tilde{T}_{\gamma_{in}} = n(\bar{\boldsymbol{X}}'_a, \mathbf{0}'_b)\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}}(1 + o_p(1)), \quad i = 1, 2,$$

which implies that, under $H_1$, $\tilde{T}_{\gamma_{1n}}$ and $\tilde{T}_{\gamma_{2n}}$ converge to infinity with the same rate $O(n^{1+2t})$. □

To prove Proposition 2, we first give the following two lemmas.

**Lemma 1** *When $\gamma \in [\tilde{\gamma}_K, \tilde{\gamma}_0]$ decreases, $l(\widehat{\boldsymbol{\mu}}_\gamma)$ strictly decreases, and $\tilde{T}_\gamma$ strictly increases.*

**Proof.** Let $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\sqrt{n}\bar{\boldsymbol{X}}$, $\mathbf{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\sqrt{n}\boldsymbol{\Lambda}$ and $\zeta = n\gamma$. For the LASSO objective function (A.1), the LARS-LASSO transition points are assumed to be $\tilde{\zeta}_0 > \tilde{\zeta}_1 > \ldots > \tilde{\zeta}_K = 0$. For any $\zeta \in (\tilde{\zeta}_{m+1}, \tilde{\zeta}_m)$, where $0 \leq m \leq K-1$, by the definition of transition points, $\mathfrak{B}(\zeta)$ and $\boldsymbol{S}(\zeta)$ do not change with $\zeta$ and take the values of, say $\mathfrak{B}(\zeta) = \mathfrak{B}_m$ and $\boldsymbol{S}(\zeta) = \boldsymbol{S}_m$. Then, by Lemma 1 in Zou et al. (2007), we have

$$\widehat{\boldsymbol{\alpha}}_{\zeta\mathfrak{B}_m} = (\mathbf{Z}'_{\mathfrak{B}_m}\mathbf{Z}_{\mathfrak{B}_m})^{-1}(\mathbf{Z}'_{\mathfrak{B}_m}\boldsymbol{Y} - \frac{\zeta}{2}\boldsymbol{S}_{m\mathfrak{B}_m}). \tag{A.3}$$

Similar to the proof of Theorem 3 in Zou et al. (2007), we have

$$\begin{aligned} l(\widehat{\boldsymbol{\mu}}_\gamma) &= (\boldsymbol{\Sigma}^{-\frac{1}{2}}\sqrt{n}\bar{\boldsymbol{X}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\sqrt{n}\boldsymbol{\Lambda}\widehat{\boldsymbol{\alpha}}_\gamma)'(\boldsymbol{\Sigma}^{-\frac{1}{2}}\sqrt{n}\bar{\boldsymbol{X}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\sqrt{n}\boldsymbol{\Lambda}\widehat{\boldsymbol{\alpha}}_\gamma) \\ &= \boldsymbol{Y}'(\mathbf{I} - \mathbf{H}_{\mathfrak{B}_m})\boldsymbol{Y} + \frac{\zeta^2}{4}\boldsymbol{S}'_{m\mathfrak{B}_m}(\mathbf{Z}'_{\mathfrak{B}_m}\mathbf{Z}_{\mathfrak{B}_m})^{-1}\boldsymbol{S}_{m\mathfrak{B}_m}, \end{aligned}$$

where $\mathbf{H}_{\mathfrak{B}_m} = \mathbf{Z}_{\mathfrak{B}_m}(\mathbf{Z}'_{\mathfrak{B}_m}\mathbf{Z}_{\mathfrak{B}_m})^{-1}\mathbf{Z}'_{\mathfrak{B}_m}$. Thus, $l(\widehat{\boldsymbol{\mu}}_\gamma)$ is strictly increasing in interval $(\tilde{\zeta}_{m+1}, \tilde{\zeta}_m)$. By the continuity property of LASSO solutions (cf., Lemma 4 of Zou et al. 2007), we have the conclusion that $l(\widehat{\boldsymbol{\mu}}_\gamma)$ monotonically decreases as $\gamma$ decreases.

For $\tilde{T}_\gamma$, when $\zeta \in (\tilde{\zeta}_{m+1}, \tilde{\zeta}_m)$, it is easy to check that

$$\tilde{T}_{n\zeta} = \frac{(\widehat{\boldsymbol{\alpha}}'_{\zeta\mathfrak{B}_m}\mathbf{Z}'_{\mathfrak{B}_m}\boldsymbol{Y})^2}{\widehat{\boldsymbol{\alpha}}'_{\zeta\mathfrak{B}_m}(\mathbf{Z}'_{\mathfrak{B}_m}\mathbf{Z}_{\mathfrak{B}_m})\widehat{\boldsymbol{\alpha}}_{\zeta\mathfrak{B}_m}}.$$

After substituting (A.3) into the above equation and certain additional mathematical manipulations, we have

$$\frac{\partial\tilde{T}_{n\zeta}}{\partial\zeta} = \frac{1}{2}\zeta\widehat{\boldsymbol{\alpha}}'_{\zeta\mathfrak{B}_m}\mathbf{Z}'_{\mathfrak{B}_m}\boldsymbol{Y}\frac{[\boldsymbol{S}'_{m\mathfrak{B}_m}(\mathbf{Z}'_{\mathfrak{B}_m}\mathbf{Z}_{\mathfrak{B}_m})^{-1}\mathbf{Z}'_{\mathfrak{B}_m}\boldsymbol{Y}]^2 - \boldsymbol{Y}'\mathbf{H}_{\mathfrak{B}_m}\boldsymbol{Y} \times [\boldsymbol{S}'_{m\mathfrak{B}_m}(\mathbf{Z}'_{\mathfrak{B}_m}\mathbf{Z}_{\mathfrak{B}_m})^{-1}\boldsymbol{S}_{m\mathfrak{B}_m}]}{[\widehat{\boldsymbol{\alpha}}'_{\zeta\mathfrak{B}_m}(\mathbf{Z}'_{\mathfrak{B}_m}\mathbf{Z}_{\mathfrak{B}_m})\widehat{\boldsymbol{\alpha}}_{\zeta\mathfrak{B}_m}]^2}.$$

Using the Cauchy-Schwarz inequality, it is straightforward to see that

$$[\boldsymbol{S}'_{m\mathfrak{B}_m}(\mathbf{Z}'_{\mathfrak{B}_m}\mathbf{Z}_{\mathfrak{B}_m})^{-1}\mathbf{Z}'_{\mathfrak{B}_m}\boldsymbol{Y}]^2 \leq \boldsymbol{Y}'\mathbf{H}_{\mathfrak{B}_m}\boldsymbol{Y} \times [\boldsymbol{S}'_{m\mathfrak{B}_m}(\mathbf{Z}'_{\mathfrak{B}_m}\mathbf{Z}_{\mathfrak{B}_m})^{-1}\boldsymbol{S}_{m\mathfrak{B}_m}],$$

and the probability of attaining the equality is zero. Note that

$$2\widehat{\boldsymbol{\alpha}}'_{\zeta\mathfrak{B}_m}\mathbf{Z}'_{\mathfrak{B}_m}\boldsymbol{Y} - \widehat{\boldsymbol{\alpha}}'_{\zeta\mathfrak{B}_m}(\mathbf{Z}'_{\mathfrak{B}_m}\mathbf{Z}_{\mathfrak{B}_m})^{-1}\widehat{\boldsymbol{\alpha}}_{\zeta\mathfrak{B}_m} = l(\mathbf{0}) - l(\widehat{\boldsymbol{\mu}}_\gamma) > 0,$$

we have $\widehat{\boldsymbol{\alpha}}'_{\zeta\mathfrak{B}_m}\mathbf{Z}'_{\mathfrak{B}_m}\boldsymbol{Y} > 0$. Therefore, $\tilde{T}_{n\zeta}$ strictly increases when $\gamma$ decreases. □

**Lemma 2** *Under the conditions in Proposition 2, with probability tending to 1,* $\left(\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}}}\right)_a = \bar{\boldsymbol{X}}_a + O_p(n^{-1/2})$ *and* $\left(\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}}}\right)_b = 0$*, where* $\left(\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}}}\right)_a$ *and* $\left(\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}}}\right)_b$ *denote parts of* $\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}}}$ *with the first* $p_0$ *and the remaining* $p - p_0$ *elements, respectively.*

**Proof.** Similar to Theorems 2-3 in Wang and Leng (2007), we can show $\widehat{\boldsymbol{\mu}}_{\gamma_n}$ has the oracle property that, if $n\gamma_n \to \infty$ and $n^{-s+\frac{1}{2}}\gamma_n \to 0$, then with probability tending to 1, $\left(\widehat{\boldsymbol{\mu}}_{\gamma_n}\right)_a - \boldsymbol{\mu}_a = \bar{\boldsymbol{X}}_a + O_p(n^{-1/2})$ and $\left(\widehat{\boldsymbol{\mu}}_{\gamma_n}\right)_b = 0$. Hence, there exists at least one $\gamma_n^* \in (\tilde{\gamma}_K, \tilde{\gamma}_0)$ such that $\widehat{\boldsymbol{\mu}}_{\gamma_n^*}$ satisfies the oracle property. If $\tilde{\gamma}_{m_{p_0}^{\text{last}}} = \gamma_n^*$, then the results are obvious. Next, we show that $\tilde{\gamma}_{m_{p_0}^{\text{last}}} \leq \gamma_n^*$ with probability tending to 1. Since the "one at a time" condition holds almost everywhere, if $\tilde{\gamma}_{m_{p_0}^{\text{last}}} > \gamma_n^*$, then we have $\gamma_n^* \in (\tilde{\gamma}_{m_{p_0}^{\text{last}}+1}, \tilde{\gamma}_{m_{p_0}^{\text{last}}})$. Therefore, at the transition point $\tilde{\gamma}_{m_{p_0}^{\text{last}}+1}$, an index is removed from $\mathfrak{B}_{m_{p_0}^{\text{last}}}$. On the other hand, it is easy to see that $l(\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}+1}}) \geq |O_p(n^{1+2s})|$ while $l(\widehat{\mu}_{\gamma_n^*}) \sim |O_p(1)|$. By the fact that $l(\widehat{\boldsymbol{\mu}}_\gamma)$ is strictly decreasing as $\gamma$ decreases, as shown in Lemma 1 above, we have the conclusion that, with probability tending to 1, $\tilde{\gamma}_{m_{p_0}^{\text{last}}} \leq \gamma_n^*$. Similarly, the monotonic property of $l(\widehat{\boldsymbol{\mu}}_\gamma)$ would lead to the conclusion that $\mathfrak{B}(\tilde{\gamma}_{m_{p_0}^{\text{last}}}) = \mathfrak{B}(\gamma_n^*)$, i.e, $\left(\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}}}\right)_b = 0$. By (A.3) and the facts that $\tilde{\gamma}_{m_{p_0}^{\text{last}}} \leq \gamma_n^*$ and $n^{-s+\frac{1}{2}}\tilde{\gamma}_{m_{p_0}^{\text{last}}} \to 0$, we have $\left(\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}}}\right)_a = \bar{\boldsymbol{X}}_a + O_p(n^{-1/2})$. □

**Proof of Proposition 2**

Under the null hypothesis, suppose there are two constants $c_1$ and $c_2$ such that

$$P\left(n\bar{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}} > c_1\right) = P\left(\tilde{T}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}}} > c_2\right) = \alpha_0,$$

where $\alpha_0$ is a pre-specified type I error probability. Then, by Lemmas 1 and 2, and the fact that $\tilde{\gamma}_{m_{p_0}^{\text{last}}} < \tilde{\gamma}_K$, it follows that $c_1 > c_2$ and thus there exists $\varepsilon > 0$ so that $c_1 - (c_2 + \varepsilon) > 0$. Under the alternative hypothesis, by Lemma 2 and the Taylor expansions, after certain high-order terms are ignored, we have

$$\tilde{T}_{\tilde{\gamma}_{m_{p_0}^{\text{last}}}} = n\bar{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}}(1 + o_p(1)).$$

Then, under $H_1$, with probability tending to 1, we have

$$\begin{aligned}\frac{P\left(n\bar{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}} < c_1\right)}{P\left(\tilde{T}_{\tilde{\gamma}_{m_{p0}^{\text{last}}}} < c_2\right)} &= \frac{P\left(n\bar{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}} < c_1\right)}{P\left(n\bar{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}} < c_2(1+o_p(1))\right)} \\ &\geq \frac{P\left(V < c_2+\varepsilon\right)}{P\left(V < c_2(1+o_p(1))\right)} P\left(U < c_1 - (c_2+\varepsilon)\right) \\ &\approx \exp\left\{\left(\sqrt{c_2+\varepsilon} - \sqrt{c_2}\right)(n\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{\frac{1}{2}} - \frac{\varepsilon}{2}\right\} P\left(U < c_1 - (c_2+\varepsilon)\right),\end{aligned}$$

where $U$ and $V$ are respectively distributed as $\chi^2_{p-1}$ and the non-central chi-square distribution $\chi^2_1(n\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$. In the above expression, we have used the fact that $n\bar{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}}$ can be partitioned into two independent parts $U$ and $V$. By noting that the probability $Pr\left\{U < c_1 - (c_2+\varepsilon)\right\}$ is a value greater than zero for the fixed $\varepsilon$, it follows that $P\left(n\bar{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{X}} < c_1\right)/P\left(\tilde{T}_{\tilde{\gamma}_{m_{p0}^{\text{last}}}} < c_2\right) \to \infty$, which completes the proof. □

**Proof of Proposition 3**

Note that $\sqrt{\frac{2-\lambda}{\lambda[1-(1-\lambda)^{2j}]}}\boldsymbol{U}_j \sim N(0,\boldsymbol{\Sigma})$ under the IC condition. Thus, to prove Proposition 3, it is equivalent to show that the distributions of $\tilde{T}_{\tilde{\gamma}_{m_i^{\text{last}}}}$, for $i \geq 1$, do not depend on $n$. According to the LARS-LASSO algorithm of Efron et al. (2004), the distribution of LASSO solution path (cf., (4)) does not depend on $n$, using (A.1) and the fact that the distributions of $\sqrt{n}\bar{\boldsymbol{X}}$ and $\sqrt{n}\boldsymbol{\Lambda}$ are free of $n$ under the IC condition. Consequently, the distributions of $\sqrt{n}\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_i^{\text{last}}}} = \sqrt{n}\boldsymbol{\Lambda}\widehat{\boldsymbol{\alpha}}_{\tilde{\gamma}_{m_i^{\text{last}}}}$, for $i = 1,\ldots,p$, are free of $n$. By the fact that $\tilde{T}_{\tilde{\gamma}_{m_i^{\text{last}}}} = (\sqrt{n}\widehat{\boldsymbol{\mu}}'_{\tilde{\gamma}_{m_i^{\text{last}}}}\boldsymbol{\Sigma}^{-1}\sqrt{n}\bar{\boldsymbol{X}})^2/(\sqrt{n}\widehat{\boldsymbol{\mu}}'_{\tilde{\gamma}_{m_i^{\text{last}}}}\boldsymbol{\Sigma}^{-1}\sqrt{n}\widehat{\boldsymbol{\mu}}_{\tilde{\gamma}_{m_i^{\text{last}}}})$, we complete the proof. □