Discussion Paper

Applied Stochastic Models in Business and Industry

Received 30 December 2015, Revised 15 July 2016, Accepted 15 August 2016 Published online 21 September 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2193

# Post selection shrinkage estimation for high-dimensional data analysis

Xiaoli Gao[a*†], S. E. Ahmed[b] and Yang Feng[c]

**In high-dimensional data settings where $p \gg n$, many penalized regularization approaches were studied for simultaneous variable selection and estimation. However, with the existence of covariates with weak effect, many existing variable selection methods, including Lasso and its generations, cannot distinguish covariates with weak and no contribution. Thus, prediction based on a subset model of selected covariates only can be inefficient. In this paper, we propose a post selection shrinkage estimation strategy to improve the prediction performance of a selected subset model. Such a post selection shrinkage estimator (PSE) is data adaptive and constructed by shrinking a post selection weighted ridge estimator in the direction of a selected candidate subset. Under an asymptotic distributional quadratic risk criterion, its prediction performance is explored analytically. We show that the proposed post selection PSE performs better than the post selection weighted ridge estimator. More importantly, it improves the prediction performance of any candidate subset model selected from most existing Lasso-type variable selection methods significantly. The relative performance of the post selection PSE is demonstrated by both simulation studies and real-data analysis. Copyright © 2016 John Wiley & Sons, Ltd.**

**Keywords:** asymptotic risk; lasso; ridge regression; (positive) shrinkage estimation; post selection; sparse model

## 1. Introduction

Many high-dimensional data arise in biological, medical, social, and economical studies. Because of the trade-off between model complexity and model prediction, the statistical inference of model selection becomes extremely important and challenging in high-dimensional data analysis. Consider a classical high-dimensional linear regression model with $i$th observed response variable $y_i$ and covariates $x_{ij}$s,

$$y_i = \sum_{j=1}^{p_n} x_{ij}\beta_j + \varepsilon_i, \quad 1 \leqslant i \leqslant n, \tag{1.1}$$

where $\varepsilon_i$s is independent and identically distributed random errors with center 0 and variance $\sigma^2$. Without loss of generality, we do not include the intercept in the model by assuming all data have been centered. Here, the subscript $n$ in $p_n$ indicates that the number of coefficients may increase with the sample size $n$. Such a notation will be used throughout the paper without further explanation.

Over the past two decades, many penalized regularization approaches have been developed to do variable selection and estimation simultaneously. Among them, the Lasso [1] is one of the most popular approaches because of its convexity and computation efficiency. In general, the Lasso penalty tends to select an over-fitted model because it penalizes all coefficients equally [2]. Many endeavors have been undertaken to improve the Lasso to reach both variable selection consistency and the estimation consistency. To list a few, smoothly clipped absolute deviation [3, 4], adaptive Lasso [5] and minimax concave penalty [6], among others. An overview of variable selection in high-dimensional feature space can be found in [7].

In order to have nice estimation and selection properties, most Lasso-type penalties make some important assumptions about both true model and designed covariates. For example, the true model is often assumed to be sparse, insofar

[a] *Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC, USA*
[b] *Department of Mathematics, Brock University, St. Catharines, ON, Canada*
[c] *Department of Statistics, Columbia University, New York, NY, USA*
**Correspondence to: Xiaoli Gao, Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC, USA.*
[†] *E-mail: x_gao2@uncg.edu*

Copyright © 2016 John Wiley & Sons, Ltd.

that (i) most $\beta_j$s are zeros except for a few ones and (ii) all those nonzero $\beta_j$'s are larger than an inflated noise level, $c\sigma\sqrt{(2/n)\log(p_n)}$ with $c \geqslant 1/2$ [8]. Additional assumptions made on the designed covariates include the adaptive irrepresentable condition and the restricted eigenvalue conditions. For detailed information, we refer to [9], [10], and [11].

However, those conditions are somewhat restrictive and are not judiciously justified in real applications. Consequently, Lasso and its generalizations may have lower prediction efficiency once those assumptions are violated. To fix the idea, we take the sparse model assumption (ii) as an example. Suppose we can divide the index set $\{1, \dots, p_n\}$ into three disjoint subsets: $S_1$, $S_2$, and $S_3$. In particular, $S_1$ includes indexes of nonzero $\beta_i$'s which are moderately large and easily detected; $S_3$ includes indexes with only zero coefficients; $S_2$, being the intermediate, includes indexes of those nonzero $\beta_j$ with weak but nonzero effects. Thus, $S_1$ is able to be detected using some existing variable selection techniques, while $S_2$ may not be separated from $S_3$ in general using existing Lasso-type methods. A more detailed description can be found in [8]. Following the spirit of model parsimony, covariates in $S_1$ are kept in the model, and some or all covariates in $S_2$ are left aside with ones in $S_3$. Author in [12] has showed using simulation studies that such a Lasso estimate often performs worse than the post selection least squares estimate. To improve the prediction error of a Lasso-type variable selection approach, some (modified) post least squares estimators are studied in [13] and [14]. However, this work still assume the irrepresentable condition, and those post estimations are only based upon the chosen subset after the Lasso. Consequently, the simultaneous weak effects in $S_2$ are still ignored. An ideal strategy would be able to incorporate the joint contribution from covariates in $S_2$, even though a parsimonious model without including covariates in $S_2$ is adopted.

Let us consider an extreme case where $S_1$ is a null set and $p$ is fixed. It has been studied extensively that shrinkage estimators can have uniformly smaller risk compared with the ordinary maximum likelihood estimators (MLEs) since the discussion papers in [15] and [16]. The relative risk properties of shrinkage estimators were also investigated in low-dimensional regression models under a restricted linear submodel space. See, for example, [17–20] and many others.

However, in high-dimensional settings where $p > n$, *a priori* information on $S_1$ is not guaranteed, not mentioning the existence of an MLE. Thanks to the existing variable selection techniques, an estimated candidate subset $\hat{S}_1$ is selected. Once $\hat{S}_1$ is obtained, the next question we want to ask is: can we construct a post selection shrinkage estimate to improve the risk of the post selection least squares estimators?

As we know, ridge regression [21, 22] has been widely used when the design matrix is ill-conditioned such that a regular MLE is not available. In this paper, we follow the model parsimony spirit and extend shrinkage estimation to the high-dimensional data setting using both ridge penalty and Lasso-type penalty separately. In particular, we use a ridge penalty to construct a data-adaptive post selection shrinkage estimator (PSE) to improve the risk of a post selection least squares estimator based upon a Lasso-type variable selection result.

We summarize our main contributions as follows:

(1) We propose a post selection shrinkage strategy to improve the risk of the Lasso-type estimators in high-dimensional settings. This post selection shrinkage strategy is data adaptive and has some practical applications, especially when an 'important' subset is generated and some covariates with joint weak effects are not selected.
(2) We investigate the asymptotic risk of the proposed PSEs. Corresponding asymptotic properties of a predecessor generating those PSEs are also investigated under some regularity conditions.

The rest of the paper is organized as follows. In Section 2, we describe some preliminary model information involved in building a PSE. As preparation, we introduce some sparsity definitions under certain signal strength levels. Some existing variable selection results from Lasso are also summarized in this section. We propose three steps in constructing the shrinkage strategy in Section 3. In Section 4, we investigate some asymptotic properties of those post selection estimators during three steps in Section 3. We first investigate some asymptotic normality properties of the designed weighted ridge (WR) estimators under some conditions. Then, we investigate the asymptotic distributional risks of the linear combination of the proposed PSEs. In Sections 5 and 6, we perform some numerical studies using some simulated examples and a real-data application, respectively. We summarize the paper with some discussions in the final section. All proofs are given in the Appendix.

## 2. Model description and basic notations

Let $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_{p_n}^*)'$ be the true coefficients vector in model (1.1). For any subset $S \subset \{1, \dots, p_n\}$ with a cardinal value $|S|$, denote $\boldsymbol{\beta}_S^*$ a subvector of $\boldsymbol{\beta}^*$ indexed by $S$. Similar subscripts are used for other submatrices and subvectors.

Copyright © 2016 John Wiley & Sons, Ltd.

### 2.1. *Model sparsity and signal strength*

As introduced in the previous section, the effect of all $p_n$ covariates is characterized into three categories based upon their signal strength: important covariates with strong effects in $S_1$, covariates with no effect in $S_3$, and an intermediate group in $S_2$ with joint weak effects. In particular, those *signal strength assumptions* of the true model are made explicitly as follows:

**(A1)** There exists a positive constant $c_1$, such that $|\beta_j^*| > c_1\sqrt{(\log p_n)/n}$ for $\forall j \in S_1$;
**(A2)** The parameter vector $\boldsymbol{\beta}^*$ satisfies that $\|\boldsymbol{\beta}_{S_2}^*\| = O(n^\tau)$ for some $0 < \tau < 1$, where $\|\cdot\|$ is the $\ell_2$ norm;
**(A3)** $\beta_j^* = 0$, for $\forall j \in S_3$.

Assumptions (A1–A3) specify those signal strength levels in the strong signals set $S_1$, weak signals set $S_2$, and sparse signal set $S_3$ explicitly. In particular, (A2) indicates that joint weak effects in $\boldsymbol{\beta}_{S_2}^*$ only grow with $n$ at a certain rate, even though the dimension $p_n$ grows with $n$ fast. For example, if (A1) holds for some $c_1 > 0$ and we let $|\beta_{0j}| < c_1\sqrt{(\log p_n)/n}$ for $j \in S_2$ with $|S_2| < n$, then $\|\boldsymbol{\beta}_{S_2}^*\| < c_1\sqrt{\log(p_n)} < O(n^\tau)$ even though $p_n = O(\exp(n^{2\tau}))$.

Most existing high-dimensional sparse models investigate the variable selection consistency by only considering the existence of the strong signals in (A1) and sparse signals in (A3). There is very limited work assuming the existence of weak signals in $S_2$. For example, besides a strong signal set in (A1), [23] does not separate $S_2$ and $S_3$ and makes an alternative sparse model assumption,

**(A2')** $\sum_{j \notin S_1} |\beta_j^*| \leqslant \eta_1$ for some $\eta_1 \geqslant 0$.

In their work, some sufficient conditions are investigated under which the Lasso can select the strong signal set $S_1$ consistently, following the spirit of the model parsimony.

Our weak and sparse conditions in (A2–A3) are different from the sparse condition in (A2') where $S_2$ and $S_3$ are not separated. If we replace (A2) by (A2') in our signal strength assumptions, then (A2) becomes $\|\boldsymbol{\beta}_{S_2}^*\| \leqslant \sum_{j\in S_2} |\beta_j^*| = \eta_1$, the joint effects in $S_2$ being bounded uniformly. Thus, a true model under (A2') only is less sparse than one under (A3) only but more sparse than one in both (A2) and (A3). On the contrary, a sparse model under both (A2) and (A3) includes the most weak signals; a sparse model under (A3) only does not have any weak signals, while a sparse model under (A2') only is in the middle.

### 2.2. *Parsimonious model selection*

As discussed in Section 1, a penalized least squares (PLS) estimator is often adopted to select a parsimonious model for a high-dimensional regression model in (1.1),

$$\hat{\boldsymbol{\beta}}_n^{\text{PLS}} = \arg\ \min\left\{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p_n} x_{ij}\beta_j\right)^2 + \sum_{j=1}^{p_n} p_\lambda(\beta_j)\right\}, \tag{2.1}$$

where $p_\lambda(\beta_j)$ is the penalty term on $\beta_j$ with a tuning parameter controlling the size of selected candidate subset model. For example, the Lasso takes $p_\lambda(\beta_j) = \lambda|\beta_j|$, and the adaptive Lasso takes $p_\lambda(\beta_j) = \lambda|\beta_j|/|w_j|$, where $w_j$ can be taken as an initial estimator of $\beta_j$. The size of selected subset model depends strongly on the choice of tuning parameters in (2.1). As pointed out by [8], one turns to ignore weak signals in $S_2$ together with $S_3$ and select a candidate subset model with only strong signals in $S_1$, following the model parsimony spirit.

If we let $\hat{S}_1 \subset \{1, \ldots, p_n\}$ index an active subset from (2.1), then a data-adaptive candidate subset model is produced such that

$$\hat{\beta}_j^{\text{PLS}} = 0 \quad \text{if and only if } j \notin \hat{S}_1. \tag{2.2}$$

Denote the response vector $\mathbf{y} = (y_1, \ldots, y_n)'$, all covariates vectors $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})'$ for $j = 1, \ldots, p_n$, and the design matrix $\mathbf{X} = (\mathbf{x}_1 \ldots \mathbf{x}_{p_n})$. Without loss of generality, we rearrange the designed vectors such that $\mathbf{X} = (\mathbf{X}_{S_1}|\mathbf{X}_{S_2}|\mathbf{X}_{S_3})$, where $\mathbf{X}_S$ is the submatrix consists of vectors indexed by $S \subset \{1, \ldots, p_n\}$. Next, we give two scenarios where $S_2$ cannot be separated from $S_3$.

*Case* 1 ([24])
Consider an orthonormal design with $\mathbf{X}'\mathbf{X}/n = \mathbf{I}_n$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_n)$. The PLS with Lasso penalty provides a soft-threshold estimator with $\hat{\beta}_j^{\text{Lasso}} = \tilde{\beta}_j - \lambda/(2n)\text{sgn}(\tilde{\beta}_j)$ and 0, for $|\tilde{\beta}_j| > \lambda/(2n)$ and $|\tilde{\beta}_j| < \lambda/(2n)$, respectively. Here, $\tilde{\beta}_j = \mathbf{x}_j'\mathbf{y}/n \sim N(\beta_{0j}, 1/n)$ is the least squares solution, and $\text{sgn}(\cdot)$ is the sign mapping function. If $\min_{j\in S_1} |\beta_j^*| > \lambda/(2n) > c > \max_{j\in S_2} |\beta_j^*|$

for some $c > 0$, then $P(\hat{S}_1 = S_1) \to 1$; that is, $P(\hat{\beta}_j^{\text{PLS}} = 0) \to 1$ for $j \notin S_1$. Thus, all weak signals in $S_2$ are omitted together with sparse signals in $S_3$ using the Lasso approach.

*Case* 2 ([25])
Consider a non-singular design such that the smallest eigenvalue of $\mathbf{X}'_{S_3^c}\mathbf{X}_{S_3^c}/n$ is larger than some positive constant $c$. If there exists some $j \in S_2$ such that $|\beta_{0j}| < |g_j(\lambda)|$, where $g_j(\lambda) = \lambda \mathbf{e}'_j(\mathbf{X}'_{S_3^c}\mathbf{X}_{S_3^c})^{-1}\text{sgn}(\boldsymbol{\beta}_{0S_3^c})$ with $\mathbf{e}_j$ being the $j$th column of the identity matrix, then $P(\{S_1 \cup S_2 \subseteq \hat{S}_1\} \cap \{S_3 \subseteq \hat{S}_1^c\}) < 1$. Thus, $S_2$ and $S_3$ cannot be separated using the Lasso.

Some post selection estimators were proposed to improve the prediction performance of the PLS estimator. For example, under some regularity conditions, [13] and [14] studied some post selection least square estimators,

$$\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{RE}} = \left(\mathbf{X}'_{\hat{S}_1}\mathbf{X}_{\hat{S}_1}\right)^{-1}\mathbf{X}'_{\hat{S}_1}\mathbf{y}. \tag{2.3}$$

Here, we denote such a post selection least squares estimator as a restricted estimator (RE), written as $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{RE}}$ in this paper. For notation's convenience, we omit the phase of 'post selection' in some future short notations without causing any confusion.

When $S_1$ and $S_2$ are not separable, we tend to select the important subset $\hat{S}_1$, such that $\hat{S}_1 \subseteq S_1$ for a large enough $\lambda$, or $S_1 \subset \hat{S}_1 \subset S_1 \cup S_2$ for a smaller $\lambda$, following the spirit of model parsimony. Although $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{RE}}$ is more estimation efficient than $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{PLS}}$, the prediction risk of $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{RE}}$ can still be high because many weak signals in $S_2$ are ignored in $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{RE}}$. Our interest is to improve the risk performance of $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{RE}}$ given in (2.3) by picking up some information from $\hat{S}_1^c$, a complement subset of the selected candidate submodel.

### 2.3. *Some additional notations*

Based upon a subset partition $S_1, S_2, S_3$, we can partition the true parameters $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}_1^{*'}, \boldsymbol{\beta}_2^{*'}, \boldsymbol{\beta}_3^{*'}\right)'$, without loss of generality. Some notations are shortened for notation's simplicity such that $\boldsymbol{\beta}_{S_k}^* = \boldsymbol{\beta}_k^*$ for $k = 1, 2$ and 3. Similar notations are also adopted for other subvectors and matrices. For example, after the same partition, the design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_{p_n})$ can be written as $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$. We also write $\mathbf{X} = (\mathbf{Z}, \mathbf{X}_3)$ with $\mathbf{Z} = (\mathbf{X}_1, \mathbf{X}_2)$. The row vector of $\mathbf{Z}$ is denoted as $\mathbf{z}_i = (z_{i1}, \cdots, z_{i,p_1+p_2})$ for $1 \leqslant i \leqslant n$.

We denote $p_k = |S_k|$ for $1 \leqslant k \leqslant 3$ and $p_n = p_1 + p_2 + p_3$. In this paper, we allow $p_n = \sum_{k=1}^3 p_k$ to be very large but restrict $q = p_1 + p_2 \leqslant n$ such that $\boldsymbol{\Sigma}_n = n^{-1}\mathbf{Z}'\mathbf{Z}$ is non-singular. If $\boldsymbol{\Sigma}_n$ is singular, then a generalized inverse matrix is adopted when needed in computations. Some other submatrices of $\boldsymbol{\Sigma}_n$ are defined as follows:

$$\begin{aligned}
\boldsymbol{\Sigma}_{n11} &= \mathbf{X}_1'\mathbf{X}_1/n, \quad \boldsymbol{\Sigma}_{n22} = \mathbf{X}_2'\mathbf{X}_2/n, \\
\boldsymbol{\Sigma}_{n12} &= \mathbf{X}_1'\mathbf{X}_2/n, \quad \boldsymbol{\Sigma}_{n21} = \mathbf{X}_2'\mathbf{X}_1/n, \\
\boldsymbol{\Sigma}_{n22.1} &= n^{-1}\mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 \\
\boldsymbol{\Sigma}_{n11.2} &= n^{-1}\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1
\end{aligned} \tag{2.4}$$

Let $\mathbf{U} = (\mathbf{X}_2, \mathbf{X}_3)$ be a $n \times (p_n - p_1)$ submatrix of $\mathbf{X}$. Then, another partition is written as $\mathbf{X} = (\mathbf{X}_1, \mathbf{U})$. Let $\boldsymbol{M}_1 = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. Then, $\mathbf{U}'\boldsymbol{M}_1\mathbf{U}$ is a $(p_n - p_1) \times (p_n - p_1)$ dimensional singular matrix with rank $k_n \geqslant 0$. We denote $\varrho_{1n} \leqslant \ldots \leqslant \varrho_{k_n n}$ as all $k_n$ positive eigenvalues of $\mathbf{U}'\boldsymbol{M}_1\mathbf{U}$.

## 3. Post selection shrinkage estimation strategy

We propose a high-dimensional post selection shrinkage estimation strategy based upon the following three steps:

*Step 1:* Obtain a data-adaptive candidate subset $\hat{S}_1$ following a model parsimony spirit and construct a post selection least square estimator $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{RE}}$ using (2.3);

*Step 2:* Obtain a post selection WR estimator, $\hat{\boldsymbol{\beta}}_n^{\text{WR}} = \left(\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{WR}}, \hat{\boldsymbol{\beta}}_{\hat{S}_1^c}^{\text{WR}}\right)$, using a threshold ridge penalty to be introduced and a submodel $\hat{S}_1$ selected from Step 1;

*Step 3:* Obtain a PSE by shrinking $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{WR}}$ from Step 2 in the direction of $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{RE}}$ from Step 1.

The post selection WR estimator in Step 2 can handle three scenarios simultaneously: (a) the sparsity in high-dimensional data analysis; (b) the strong correlation among covariates; and (c) the jointly weak contribution from some covariates.

Copyright © 2016 John Wiley & Sons, Ltd.

*Remark* 1
This post selection shrinkage estimation is expected to improve the risk performance on the selected submodel once a variable selection approach in Step 1 tends to select those and only those variables with strong signal strength, that is, $S_1 \supset \hat{S}_1$ or $S_1 \subset \hat{S}_1 \subset S_1 \cup S_2$. However, if the model parsimony spirit is not followed and $\lambda$ in (2.1) is too small such that $\hat{S}_1 \supset S_1 \cup S_2$, this post selection shrinkage estimation is not suggested. Therefore, the effect of the PSE is data adaptive and depends on $\hat{S}_1$.

As a preparation, we first construct a post selection WR estimation based upon $\hat{S}_1$. This post selection weight ridge estimation itself is constructed from two steps introduced in Section 3.1 and 3.2.

### 3.1. *Weighted ridge estimation*

Once $\hat{S}_1$ is obtained from Step 1, we seek to minimize a penalized objective function with a ridge penalty on coefficients in $\hat{S}_1^c$,

$$\widetilde{\boldsymbol{\beta}}(r_n) = \arg\ \min\{L(\boldsymbol{\beta}; \hat{S}_1)\} = \arg\ \min\left\{\|\mathbf{Y} - \mathbf{X}_n\boldsymbol{\beta}_n\|^2 + r_n\|\boldsymbol{\beta}_{\hat{S}_1^c}\|^2\right\} \tag{3.1}$$

where $r_n > 0$ is a tuning parameter controlling the penalty effect on $\boldsymbol{\beta}_{\hat{S}_1}$. Then, a post selection WR estimator $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}(r_n, a_n; \hat{S}_1) = (\beta^{\mathrm{WR}}_{\hat{S}_1}(r_n), \beta^{\mathrm{WR}}_{\hat{S}_1^c}(r_n, a_n)$ is obtained from,

$$\hat{\beta}_j^{\mathrm{WR}}(r_n, a_n) = \begin{cases} \tilde{\beta}_j(r_n), & j \in \hat{S}_1; \\ \tilde{\beta}_j(r_n) I\left(\tilde{\beta}_j(r_n) > a_n\right), & j \in \hat{S}_1^c, \end{cases} \tag{3.2}$$

where $I(\cdot)$ is the indicator function and $a_n$ is a threshold parameter. Thus, we obtain estimators of the weak signal subset

$$\hat{S}_2 := \hat{S}_2(\hat{S}_1) = \left\{j \in \hat{S}_1^c : \hat{\beta}_j^{\mathrm{WR}}(r_n, a_n) \neq 0\right\} \tag{3.3}$$

and of the sparse subset

$$\hat{S}_3 := \hat{S}_3(\hat{S}_1) = \left(\hat{S}_1 \cup \hat{S}_2\right)^c. \tag{3.4}$$

Our post selection strategy is only applied when the threshold parameter $a_n$ satisfies $|\hat{S}_2| > 2$ and $|\hat{S}_3^c| < n$. In particular, we set

$$a_n = c_1 n^{-\alpha}, \quad 0 < \alpha \leqslant 1/2, \text{ for some } c_1 > 0. \tag{3.5}$$

*Remark* 2
We call $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}(r_n, a_n)$ a post selection WR estimator from two facts: (i) we only penalize parameters in $\boldsymbol{\beta}_{\hat{S}_1^c}$ instead of the entire coefficients vector $\boldsymbol{\beta}_n$, and (ii) the threshold step in (3.2) can be interpreted as a WR penalty $r_n \sum_{j \in \hat{S}_1^c}\left(\beta_j^2/w_j^2\right)$ in (3.1), where $w_j = 0$ and 1 for $j \in \hat{S}_3$ and $j \in \hat{S}_2$.

*Remark* 3
Similar to the discussion in Remark 2, we can also understand the post selection step into the WR estimator, $r_n \sum_{j \in \hat{S}_1^c}\left(\beta_j^2/w_j^2\right)$ with $w_j = \infty$ for $j \in \hat{S}_1$. We do not enforce an additional ridge penalty on $\hat{S}_1$ to reduce some unnecessary biases during the WR step. This is different from the post selection threshold regression studied in [26], where the $\ell_2$ penalty is applied on the entire $\boldsymbol{\beta}_n$ equally.

*Remark* 4
The idea of the WR regression is connected to the regularization after retention framework proposed in [27]. In that framework, a retention step is conducted to find the important set $\hat{S}_1$ with large marginal-correlation coefficients with the response. Then, a regularization step is conducted by a penalized least square with $L_1$ regularization only on the covariates that are not in $\hat{S}_1$. Compared with that framework, the current framework focused more on prediction by using the ridge penalty, and the estimate $\hat{S}_1$ is also different.

Notice that for every selected candidate subset $\hat{S}_1$, $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1}(r_n)$ depends on $r_n$ and $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1^c}(r_n, a_n)$ depends on both $r_n$ and $a_n$. For convenience, we omit those tuning parameters and denote above post selection WR estimators as $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1^c}$, respectively.

Copyright © 2016 John Wiley & Sons, Ltd.

### 3.2. *Post selection shrinkage estimation*

Now, we are ready to propose a shrinkage estimation based upon two post selection estimators: $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{S_1}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{S_1}$.

An initial PSE $\hat{\boldsymbol{\beta}}^{\mathrm{SE}}_{\hat{S}_1}$ is defined as

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\mathrm{SE}}_{\hat{S}_1} &= \hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1} + \left(\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1} - \hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1}\right)\left(1 - (\hat{s}_2 - 2)/\hat{T}_n\right)\\ &= \hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1} - \left((\hat{s}_2 - 2)/\hat{T}_n\right)\left(\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1} - \hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1}\right),\end{aligned} \tag{3.6}$$

where $\hat{s}_2 = |\hat{S}_2|$ and $\hat{T}_n$ are given by

$$\hat{T}_n = \left(\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_2}\right)' \left(\mathbf{X}'_{\hat{S}_2} \boldsymbol{M}_{\hat{S}_1} \mathbf{X}_{\hat{S}_2}\right) \hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_2} / \sigma^2, \tag{3.7}$$

where $\boldsymbol{M}_{\hat{S}_1} = \mathbf{I}_n - \mathbf{X}_{\hat{S}_1}\left(\mathbf{X}'_{\hat{S}_1}\mathbf{X}_{\hat{S}_1}\right)^{-1}\mathbf{X}'_{\hat{S}_1}$. If $\sigma^2$ is unknown, it is replaced by a consistent estimator $\hat{\sigma}^2$. In the numerical studies, $\sigma^2$ is replaced by $\hat{\sigma}^2 = \sum_{i=1}^{n}\left(y_i - \mathbf{x}'_i\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_2}\right)^2/(n - \hat{s}_2)$, and a generalized inverse is used if $\left(\mathbf{X}'_{\hat{S}_1}\mathbf{X}_{\hat{S}_1}\right)^{-1}$ is not singular.

Observing from (3.6) and (3.7), signs of two estimators of $\boldsymbol{\beta}_{\hat{S}_1}$ can be reversed if $\hat{T}_n$ is too small such that $\hat{s}_2 - 2 > \hat{T}_n$. It is possible because $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1^c}$ consists of nuisance parameters, and over-shrinkage can occur for a large $r_n$ in the WR step. Thus, we also suggest to modify (3.6) as the following post selection PSE,

$$\hat{\boldsymbol{\beta}}^{\mathrm{PSE}}_{\hat{S}_1} = \hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1} - \left([(\hat{s}_2 - 2)/\hat{T}_n] \wedge 1\right)\left(\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1} - \hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1}\right). \tag{3.8}$$

*Remark* 5
Our proposed post selection shrinkage estimation and the classical shrinkage estimation bear some resemblance but are different because of two facts: (i) Post selection shrinkage estimation is associated with a selected candidate subset and has some flexibility of adjusting the shrinkage strength data adaptively because $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1^c}$ depends on tuning parameters $a_n$ and $r_n$; (ii) Post selection shrinkage estimation uses an initial ridge shrinkage step and is tailored for the high-dimensional settings where multiple covariates tend to be correlated and function jointly.

## 4. Asymptotic properties

In order to investigate some asymptotic properties of the proposed post selection estimators, we first make following assumptions on the random error, $\mathbf{U}'\boldsymbol{M}_1\mathbf{U}$, and the model sparsity. One can review some notations at the end of Section 2.

**(B1)** The random error $\varepsilon_i \sim N(0, \sigma^2)$.
**(B2)** $\varrho_{1n}^{-1} = O(n^{-\eta})$, where $\tau < \eta \leqslant 1$ for $\tau$ in (A2).
**(B3)** $\log(p_n) = O(n^{\nu})$ for $0 < \nu < 1$.
**(B4)** There exists a positive definite matrix $\boldsymbol{\Sigma}$ such that $\lim_{n\to\infty}\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}$, where eigenvalues of $\boldsymbol{\Sigma}$ satisfy $0 < \rho_1 < \rho_{\boldsymbol{\Sigma}} < \rho_2 < \infty$.

Here, condition (B1) can be relaxed to a symmetric distribution with some finite moments. To simplify our theoretical investigations and handle the ultra high dimensionality, we only restrict our studies to normal random error in this paper. Condition (B2) guarantees that the positive eigenvalues of the redundant $\boldsymbol{U}'\boldsymbol{M}_1\boldsymbol{U}$ cannot be too small with a rate associated with the weak signals strength in $S_2$. Condition (B3) permits the ultra-high dimensionality such that the number of variables can grow with sample size at an almost exponential rate. Condition (B4) is the regularity condition for $\mathbf{X}_{S_3^c}$. This condition is made in order to obtain the asymptotic normality the WR estimator.

### 4.1. *Asymptotic properties of the weighted ridge estimator*

We have the following asymptotic properties of the WR estimator $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{n}$.

*Theorem* 1
Suppose the sparse model in (1.1) satisfies signal strength assumptions in (A1–A3) and model assumptions in (B1–B3). If we choose $r_n = c_2 a_n^{-2}(\log\log n)^3\log(n \vee p_n)$ for some constant $c_2 > 0$ and $a_n$ defined in (3.5) with $\alpha < (\eta - \nu - \tau)/3$, then $\hat{S}_2$ in (3.3) satisfies

$$P\left(\hat{S}_2 = S_2 | \hat{S}_1 = S_1\right) \geqslant 1 - (n \vee p_n)^{-t} \text{ for some constant } t > 0, \tag{4.1}$$

where $\tau$, $\eta$, and $\nu$ are defined in (A2), (B2), and (B3), respectively.

Theorem 1 is similar to the variable selection result in [28]. We postpone the detailed proof to the Appendix. It tells us that the WR estimator $\hat{\boldsymbol{\beta}}^{\text{WR}}_{S_1^c}$ is able to single out the sparse set $S_3$ with a large probability, if $S_1$ is pre-selected in advance such that $P(\hat{S}_1 = S_1) = 1$. For example, [23] argued that $S_1$ can be recovered with a large probability under the sparse Riesz condition (SRC) with rank $p_1$. Here, a design matrix $\mathbf{X}$ satisfies the SRC with rank $q$ and spectrum bounds $0 < c_* < c^* < \infty$ if

$$c_* \leqslant \frac{\|\mathbf{X}_S \mathbf{v}\|^2}{\|\mathbf{v}\|^2} \leqslant c^* \quad \forall S \text{ with } |S| = q \text{ and } \ \mathbf{v} \in \mathcal{R}^q. \tag{4.2}$$

*Lemma* 1
Consider the Lasso solution for linear model (1.1) with $\varepsilon_i \sim N(0, \sigma^2)$. Suppose (A1) and (B1) are satisfied, and the sparse condition (A2') holds for some $0 < \eta_1 < O(p_1\sqrt{\log(p_n)/n})$, and the design matrix $\mathbf{X}$ satisfies the SRC with rank $p_1$ in (4.2). Then, $\hat{S}_1$ generated from a PLS with the Lasso penalty in (2.1) satisfies

$$\lim_{n\to\infty} P\left(\{S_1 \subset \hat{S}_1\} \cap \left\{\sum_{j\in S_1} |\beta_j^*| I\left(\hat{\beta}_j^{\text{PLS}} = 0\right) = 0\right\}\right) = \lim_{n\to\infty} P(\hat{S}_1 = S_1) = 1.$$

Lemma 1 is a direct result from Theorem 2 in [23]. Here, the tuning parameter in (2.1) is chosen such that $\lambda \geqslant 2\sigma\sqrt{2(1+c_0)c^* n \log(p_n)}$. Lemma 1 indicates that those and only those strong signals in $S_1$ are included in $\hat{S}_1$ while using the Lasso under sufficient conditions.

In Lemma 1, we have $\sum_{j\notin S_1} |\beta_j^*| < \eta$. The signal of each individual coefficient is trivial if such a joint effect is uniformly distributed on $p_n - p_1$ coefficients when $p_n \gg n$. However, if this joint effect is only distributed on a much smaller number of coefficients, each individual effect may not be negligible. In particular, if we let both (A2') and (A3) hold, then $\sum_{j\in S_2} |\beta_j^*| < \eta$. Thus, (A2) also holds. Combing Lemma 1 and Theorem 1, we have the following result directly.

*Corollary* 1
Suppose all conditions in both Lemma 1 and Theorem 1 hold. Then, we have

$$\lim_{n\to\infty} P\left(\{\hat{S}_2 = S_2\} \cap \{\hat{S}_1 = S_1\}\right) = 1. \tag{4.3}$$

Corollary 1 indicates that $\hat{S}_3 = S_3$ is able to be recovered if an additional WRs step is used post the Lasso under some sufficient conditions. We skip the proof because this is a direct result from Lemma 1 and Theorem 1.

However, Corollary 1 still requires a SRC condition. Although $P(\hat{S}_1 = S_1) = 1$ may not be guaranteed when a SRC condition is not satisfied, we may have

$$P\left(\{S_1 \subset \hat{S}_1 \subset S_1 \cup S_2\}\right) \to 1. \tag{4.4}$$

Thus, we have similar but weaker result.

*Corollary* 2
Suppose all conditions in Theorem 1 hold, and $\hat{S}_1$ satisfies (4.4). Then, we have

$$\lim_{n\to\infty} P\left(\{\hat{S}_2 = \hat{S}_1^c \cap S_2\}\right) = 1. \tag{4.5}$$

Corollary 2 can be interpreted by treating $\hat{S}_1$ as a new $S_1$ and $\hat{S}_1^c \cap S_2$ as a new $S_2$.

The asymptotic properties in Theorem 1 and its derivatives in Corollary 1 and 2 are important for establishing the efficiency of $\hat{\boldsymbol{\beta}}^{\text{WR}}_{\hat{S}_1}$ and $\hat{\boldsymbol{\beta}}^{\text{WR}}_{\hat{S}_2}$.

*Theorem* 2
Let $s_n^2 = \sigma^2 \mathbf{d}_n' \boldsymbol{\Sigma}_n^{-1} \mathbf{d}_n$ for any $(p_{1n} + p_{2n}) \times 1$ vector $\mathbf{d}_n$ satisfying $\|\mathbf{d}_n\| \leqslant 1$. Suppose assumptions (B1–B4) hold. Consider a sparse model with signal strength under (A1), (A3), and (A2) with $0 < \tau < 1/2$. Suppose a pre-selected model such as

$S_1 \subset \hat{S}_1 \subset S_1 \cup S_2$ is obtained with probability 1. If we choose $r_n$ as in Theorem 1 with $\alpha < \{(\eta - \nu - \tau)/3, 1/4 - \tau/2\}$, then we have the asymptotic normality,

$$n^{1/2} s_n^{-1} \mathbf{d}'_n \left( \hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{S_3^c} - \boldsymbol{\beta}^*_{S_3^c} \right) \underset{\to}{\mathrm{d}} N(0,1). \tag{4.6}$$

Theorem 2 studies the asymptotic normality of the WR estimator, $\hat{\boldsymbol{\beta}}_{S_3^c}$. In addition, $\hat{\boldsymbol{\beta}}_{S_3^c}$ has the same estimation efficiency as one from a restricted least square estimator as if $\boldsymbol{\beta}_{S_3} = 0$ is given as *a priori*. However, the result holds if $\|\boldsymbol{\beta}^*_{S_2}\| = o(n^{1/2})$ and $r_n$ is chosen appropriately. More importantly, the strong signal set $S_1$ is detected with a large probability in advance. This can be guaranteed under Lemma 1.

### 4.2. Asymptotic distributional risk analysis

In this section, we provide the relative performance of the post selection shrinkage estimation regarding the asymptotic distribution risk (ADR) introduced in [29]. For simplicity and notation's convenience, we focus on the ADR analysis by assuming $\hat{S}_1 = S_1$, following the spirit of model parsimony. If $S_1 \subset \hat{S}_1 \subset S_1 \cup S_2$, a similar analysis can be carried out by redefining $(S_1, S_2) = (\hat{S}_1, \hat{S}_1 \cap S_2)$, as discussed in Section 4.1. Together with the results in Theorem 1, such that $P(\hat{S}_3 = S_3) \to 1$, $S_3$ is also removed from the PSE with a large probability. Thus, the risk analysis in this section will be conducted by assuming both $S_1$ and $S_3$ are known in advance.

*Definition* 1
For any estimator $\boldsymbol{\beta}^{\diamond}_{1n}$ and $p_{1n}-$dimensional vector, $\mathbf{d}_{1n}$, satisfying $\|\mathbf{d}_{1n}\| \leqslant 1$, the ADR of $\mathbf{d}'_{1n}\boldsymbol{\beta}^{\diamond}_{1n}$ is

$$\mathrm{ADR}\left(\mathbf{d}'_{1n}\boldsymbol{\beta}^{\diamond}_{1n}\right) = \lim_{n\to\infty} E\left\{ \left[ n^{1/2} s_{1n}^{-1} \mathbf{d}'_{1n} (\boldsymbol{\beta}^{\diamond}_{1n} - \boldsymbol{\beta}^*_1) \right]^2 \right\}, \tag{4.7}$$

where $s_{1n}^2 = \sigma^2 \mathbf{d}'_{1n} \boldsymbol{\Sigma}^{-1}_{n11.2} \mathbf{d}_{1n}$ with $\boldsymbol{\Sigma}_{n11.2}$ defined in (2.4).

We will provide some analytic expressions of ADRs under specific weak coefficients in (A2”). In particular,

**(A2”)** $\beta_j^* = \delta_j / \sqrt{n}$ for $j \in S_2$, where $|\delta_j| < \delta_{\max}$ for some $\delta_{\max} > 0$.

Denote $\boldsymbol{\delta} = \left(\delta_1, \dots, \delta_{p_{2n}}\right)' \in \mathcal{R}^{p_{2n}}$. Then, $\Delta_n = \boldsymbol{\delta}' \boldsymbol{\Sigma}_{n22.1} \boldsymbol{\delta} \leqslant \rho_2 p_{2n} \delta_{\max}$, where $\rho_2$ is defined in (B4).
Define

$$\Delta_{\mathbf{d}_{1n}} = \frac{\mathbf{d}'_{1n}\left(\boldsymbol{\Sigma}^{-1}_{n11}\boldsymbol{\Sigma}_{n12}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\Sigma}_{n21}\boldsymbol{\Sigma}^{-1}_{n11}\right)\mathbf{d}_{1n}}{\mathbf{d}'_{1n}\left(\boldsymbol{\Sigma}^{-1}_{n11}\boldsymbol{\Sigma}_{n12}\boldsymbol{\Sigma}^{-1}_{n22.1}\boldsymbol{\Sigma}_{n21}\boldsymbol{\Sigma}^{-1}_{n11}\right)\mathbf{d}_{1n}}. \tag{4.8}$$

We obtain the following results on the expression of ADRs of PSEs.

*Theorem* 3
Let $\mathbf{d}_{1n}$ be any $p_{1n}-$ dimensional vector satisfying $0 < \|\mathbf{d}_{1n}\| \leqslant 1$ and $s_{1n}^2 = \sigma^2 \mathbf{d}'_{1n}\boldsymbol{\Sigma}^{-1}_{n11.2}\mathbf{d}_{1n}$. Suppose all assumptions in Theorem 2 hold except that (A2) is replaced by (A2”). Then, we have

$$\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{1n}\right) = 1, \tag{4.9a}$$

$$\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{1n}\right) = 1 - (1-c)(1-\Delta_{\mathbf{d}_{1n}}), \tag{4.9b}$$

$$\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{SE}}_{1n}\right) = 1 - E[g_1(\mathbf{z}_2 + \boldsymbol{\delta})], \tag{4.9c}$$

$$\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{PSE}}_{1n}\right) = 1 - E[g_2(\mathbf{z}_2 + \boldsymbol{\delta})]. \tag{4.9d}$$

Here, $c = \lim_{n\to\infty} \mathbf{d}'_{1n}\boldsymbol{\Sigma}^{-1}_{n11}\mathbf{d}_{1n} / \left(\mathbf{d}'_{1n}\boldsymbol{\Sigma}^{-1}_{n11.2}\mathbf{d}_{1n}\right) \leqslant 1$, $\mathbf{z}_2$ satisfies that $s_{2n}^{-1}\mathbf{d}'_{2n}\mathbf{z}_2 \to N(0,1)$ with $\mathbf{d}_{2n} = \sigma^2 \boldsymbol{\Sigma}_{n21}\boldsymbol{\Sigma}^{-1}_{n11}\mathbf{d}_{1n}$ and $s_{2n}^2 = \mathbf{d}'_{2n}\boldsymbol{\Sigma}^{-1}_{n22.1}\mathbf{d}_{2n}$. In addition,

$$g_1(\mathbf{x}) = \lim_{n\to\infty}(1-c)\frac{p_{2n}-2}{\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}}\left[2 - \frac{\mathbf{x}'((p_{2n}+2)\mathbf{d}_{2n}\mathbf{d}'_{2n})\mathbf{x}}{s_{2n}^2\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}}\right], \tag{4.10}$$

and

$$g_2(\mathbf{x}) = \lim_{n\to\infty} \frac{p_{2n}-2}{\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}} \left[(1-c)\left(2 - \frac{\mathbf{x}'((p_{2n}+2)\mathbf{d}_{2n}\mathbf{d}'_{2n})\mathbf{x}}{s^2_{2n}\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}}\right)\right] I(\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x} \geqslant p_{2n}-2) \\ + \lim_{n\to\infty} \left[(2 - s^{-2}_{2n}\mathbf{x}'\mathbf{d}_{2n}\mathbf{d}'_{2n}\mathbf{x})(1-c)\right] I\left(\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x} \leqslant p_{2n}-2\right), \tag{4.11}$$

with $I(\cdot)$ being an indicator function.

Theorem 3 lists the analytic expressions of the asymptotic risk of all above estimators. From Theorem 3, we can obtain the following risk comparisons.

*Corollary* 3
Under assumptions in Theorem 3, we have

(i) $\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{PSE}}_{1n}\right) \leqslant \mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{SE}}_{1n}\right) \leqslant \mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{1n}\right)$ holds for $0 < \|\boldsymbol{\delta}\|^2 \leqslant 1$;

(ii) Inequalities in (i) also hold for $\|\boldsymbol{\delta}\|^2 \leqslant 1 + \imath$ for some $\imath > 0$ if $\Delta_n = \imath p_{2n}$.

(iii) If $\|\boldsymbol{\delta}\| = o(1)$, then $\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{1n}\right) \leqslant \mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{PSE}}_{1n}\right) < \mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{1n}\right)\}$ holds for $\boldsymbol{\delta} = 0$, where the '=' holds when $p_{2n} \to \infty$.

Corollary 3 indicates that the performance of the PSE is closely related to the post selection least squares estimator. On one hand, if $\hat{S}_1 \subset S_1 \cup S_2$ and $(S_1 \cup S_2) \cap \hat{S}^c_1$ are large, then the post selection PSE tends to dominate the RE. Thus, the post selection PSE can improve the performance of the post selection least squares estimators in [13] and [14], especially when $p_n \gg n$ and an under-fitted submodel is selected by a large penalty parameter. On the other hand, if a variable selection approach almost generates the right submodel and $\|\boldsymbol{\delta}\| = o(1)$, that is, $\lim_{n\to\infty} \hat{S}_1 = S_1 \cup S_2$, then a post selection LSE ($\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{1n}$) is the most efficient one compared with all other post selection estimates.

*Remark* 6
In the high-dimensional setting where $p \gg n$, we do need to assume the true model to be sparse in the sense that most coefficients goes to 0 when $n \to \infty$. However, we still permit some $\beta_j$ to be small but not exactly 0. Such covariates with a small amount of influence on the response variable are often ignored incorrectly in high-dimensional variable selection methods. If we borrow information from those covariates using the proposed shrinkage methods, the prediction performance based on selected submodel can be improved substantially.

## 5. Simulation studies

In this section, we use some simulation studies to examine the quadratic risk performance of the proposed estimators. Our simulation is based on the linear regression model in (1.1).

*True model setting.* In all experiments, $\epsilon_i$'s are simulated from independent and identically distributed standard normal random variables, $x_{is} = \left(\xi^1_{(is)}\right)^2 + \xi^2_{(is)}$, where $\xi^1_{(is)}$ and $\xi^2_{(is)}$, $i = 1, \cdots, n$, $s = 1, \cdots, p_n$ are also independent copies of the standard normal distribution. In all experiments, we let $n = 200$ and $p_n = n^\tau$ for different sample size $n$, where $\tau$ changes from 1 to 1.2 with an increment of 0.02. Three different coefficient configurations are considered as follows:

Case 1: $\boldsymbol{\beta}^* = (5, 5, 5, \underbrace{0.5, \ldots, 0.5}_{10}, \mathbf{0}'_{p_3})'$;

Case 2: $\boldsymbol{\beta}^* = (10, 10, 10, \underbrace{0.1, \ldots, 0.1)'}_{50}, \mathbf{0}'_{p_3})'$;

Case 3: $\boldsymbol{\beta}^* = (10, 10, 10, \underbrace{0.1, \ldots, 0.1}_{p_2}, \mathbf{0}'_{20})'$.

All nonzero coefficients are randomly assigned to be either positive or negative. Both zero and weak signals coexist in the aforementioned three settings. In Case 1, most covariates are noises. Compared with Case 1, the weak signals become weaker, and the strong signals become stronger in Case 2. In addition, the number of weak signals is larger but also fixed. In Case 3, only $p_{3n} = 20$ zero signals, large amount of weak signals contribute simultaneously, and the number of weak signals grows with the number of covariates such that $p_{2n} \gg n$. Notice that the signal strength setting in this case is different from that considered in our post selection shrinkage analysis, where $p_{2n} < n$ and $p_{3n} \gg n$.

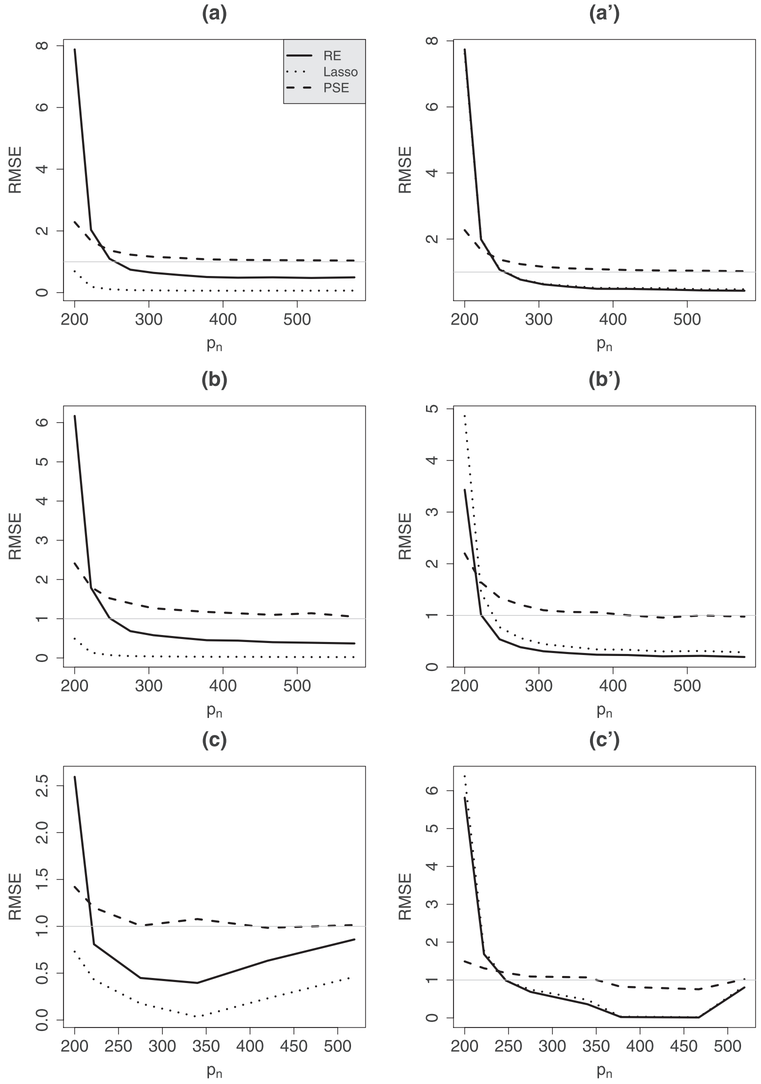

**Figure 1.** Relative mean squared errors (RMSEs) of post selection relative mean squared errora (PSEs) compared with one from Lasso or adaptive Lasso (ALasso) from simulation examples in Cases 1–3. The top (a or a'), middle (b or b'), and bottom (c or c') panels are for Cases 1, 2, and 3, respectively. The left (a–c) and right panels (a'–c') are comparisons when the candidate submodels are chosen from the Lasso and adaptive Lasso methods, respectively.

*Subset selection.* Because the adaptive Lasso, smoothly clipped absolute deviation, and minimax concave penalty perform closely under certain conditions, we only adopt the adaptive Lasso, and Lasso in selecting a subset before applying the post selection shrinkage strategy. All tuning parameters in variable selection approaches are chosen using the BIC.

*Tuning parameters and simulation Setting.* As we know, $a_n$ and $r_n$ are two important tuning parameters affecting $\hat{S}_2$ and $\hat{S}_3$. We choose those two tuning parameters based upon the asymptotic investigations in Theorem 2 for all our numerical studies. In particular, the post selection PSEs are computed for $r_n = c_2 a_n^{-2}(\log\log n)^3 \log(n \vee p_n)$ with $a_n = c_1 n^{-1/8}$. Corresponding coefficients $c_1$ and $c_2$ are chosen using cross validation.

*Evaluation.* Each design is repeated 1000 times, as a further increase in the number of realizations did not significantly change the result. Let $\boldsymbol{\beta}^{\diamond}_{1n}$ be either $\hat{\boldsymbol{\beta}}^{\mathrm{PSE}}_{1n}$ or $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{1n}$ after the variable selection. The performance of $\boldsymbol{\beta}^{\diamond}_{1n}$ is evaluated by the relative mean squared error (RMSE) criterion with respect to $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{1n}$ as follows:

$$\mathrm{RMSE}(\boldsymbol{\beta}^{\diamond}_{1n}) = \frac{E\|\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{1n} - \boldsymbol{\beta}^{*}_{1}\|^2}{E\|\boldsymbol{\beta}^{\diamond}_{1n} - \boldsymbol{\beta}^{*}_{1}\|^2}. \tag{5.1}$$

Therefore, RMSE $(\boldsymbol{\beta}^{\diamond}_{1n}) > 1$ means the superiority of $\boldsymbol{\beta}^{\diamond}_{1n}$ over $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{1n}$.

Copyright © 2016 John Wiley & Sons, Ltd.

**Table I.** Simulated RMSEs from simulation examples in Case 1–3.

| | | Lasso | | | | Adaptive Lasso | | | |
|---|---|---|---|---|---|---|---|---|---|
| Case | $p_n$ | $\lvert\hat{S}_1\rvert$ | $\hat{\beta}_{1n}^{\text{Lasso}}$ | $\hat{\beta}_{1n}^{\text{RE}}$ | $\hat{\beta}_{1n}^{\text{PSE}}$ | $\lvert\hat{S}_1\rvert$ | $\hat{\beta}_{1n}^{\text{ALasso}}$ | $\hat{\beta}_{1n}^{\text{RE}}$ | $\hat{\beta}_{1n}^{\text{PSE}}$ |
| 1 | 200 | 10.920 | 0.690 | 7.880 | 2.285 | 10.537 | 7.611 | 7.739 | 2.269 |
| | 222 | 10.785 | 0.190 | 2.035 | 1.680 | 10.434 | 2.001 | 1.991 | 1.667 |
| | 275 | 10.655 | 0.082 | 0.744 | 1.231 | 10.250 | 0.783 | 0.773 | 1.242 |
| | 340 | 10.491 | 0.066 | 0.574 | 1.126 | 10.137 | 0.585 | 0.558 | 1.114 |
| | 420 | 10.416 | 0.061 | 0.485 | 1.061 | 9.906 | 0.514 | 0.491 | 1.062 |
| | 519 | 10.293 | 0.063 | 0.476 | 1.047 | 9.781 | 0.480 | 0.446 | 1.042 |
| 2 | 200 | 3.112 | 0.491 | 6.169 | 2.409 | 3.170 | 4.859 | 3.431 | 2.199 |
| | 222 | 3.078 | 0.137 | 1.790 | 1.807 | 3.149 | 1.447 | 1.012 | 1.640 |
| | 275 | 3.041 | 0.048 | 0.684 | 1.393 | 3.083 | 0.561 | 0.384 | 1.205 |
| | 340 | 3.036 | 0.035 | 0.517 | 1.222 | 3.051 | 0.395 | 0.270 | 1.066 |
| | 420 | 3.000 | 0.029 | 0.442 | 1.138 | 3.025 | 0.335 | 0.233 | 1.003 |
| | 519 | 3.000 | 0.023 | 0.388 | 1.140 | 3.000 | 0.312 | 0.217 | 0.998 |
| 3 | 200 | 4.020 | 0.730 | 2.594 | 1.420 | 7.379 | 6.380 | 5.815 | 1.491 |
| | 222 | 6.109 | 0.430 | 0.809 | 1.200 | 10.005 | 1.778 | 1.684 | 1.310 |
| | 275 | 5.277 | 0.176 | 0.449 | 1.007 | 8.159 | 0.747 | 0.687 | 1.092 |
| | 340 | 3.046 | 0.034 | 0.396 | 1.077 | 3.783 | 0.476 | 0.361 | 1.070 |
| | 420 | 5.325 | 0.231 | 0.633 | 0.984 | 7.390 | 0.762 | 0.710 | 1.025 |
| | 519 | 7.213 | 0.461 | 0.860 | 1.014 | 9.114 | 0.844 | 0.804 | 1.020 |

$|\hat{S}_1|$is the average size of produced submodel; RMSEs, relative mean squared errors; PSE, post selection shrinkage estimator; RE, restricted estimator.

*Result:* We plot the mean RMSEs from 1000 iterations along $p_n$ in Figure 1. Some selected results are also reported in Table I. To check the behavior of Lasso or adaptive Lasso for subset selection, we also report the average number of selected important covariates as $|\hat{S}_1|$ in Table I. It is not surprising to see that RE post the adaptive Lasso is comparable with the adaptive Lasso itself, while RE post the Lasso behaves much better than Lasso [13, 14]. We summarize the simulation results as follows:

- Figure 1(a')–(c') lists results when the adaptive Lasso is used to generate the submodel. (i) When $p_n$ is closer to $n$, both post selection RE and adaptive Lasso perform better than the post selection PSE and WR (RMSE>1). (ii) When $p_n$ grows bigger, both RE and adaptive Lasso become worse than the post selection WR (RMSE<1). However, the post selection PSE still performs better than the post selection WR. Therefore, the post selection PSE provides a protection of the adaptive Lasso in the case that the adaptive Lasso loses its efficiency.
- Figure 1(a)–(c) lists results when the Lasso is used to generate the submodel. The advantage of the post selection PSE over the Lasso is more obvious than the earlier. This is because the adaptive Lasso tends to produce a more efficient estimator than the Lasso does.
- When $p_n$ grows, the post selection PSE is much more robust and at least as good as the WR estimator (RMSE is approaching to 1). When $p_n$ grows bigger, the improvement of the post selection PSE from adaptive Lasso or Lasso becomes more obvious. See Table I.
- In Case 3, the post selection PSE may lose its superiority to the post selection RE and adaptive Lasso, especially when $p_n$ grows quickly with $n$. One explanation is that the selected model size varies dramatically because the number of weak coefficients grows. However, if we still follow the model parsimony spirit and decide to use an aggressive tuning parameter to obtain a relatively consistent submodel size $\hat{S}_1$, the superiority of post selection PSEs follows the same pattern as in Cases 1 and 2.

## 6. Real-data example

In this section, we apply the proposed post selection shrinkage strategy to the growth data for the years 1960–1985 [30]. Table II lists the detailed descriptions of the dependent variable and 45 covariates related to education and its interaction with lgdp60$_i$, market efficiency, political stability, market openness, and demographic characteristics.

The growth regression model has been applied to test the negative relationship between the long-run growth rate and the initial GDP given other covariates. See [31] and [32] for literature reviews. Very recently, [33] took into account the

Copyright © 2016 John Wiley & Sons, Ltd.

possible discrepancy among the aforementioned negative relationship using a growth regression model with threshold. In particular, they consider a threshold variable in the following regression model,

$$\mathrm{gr}_i = \beta_0 + \beta_1 lgdp60_i + \mathbf{z}_i'\boldsymbol{\beta}_2 + I(Q_i < \tau)(\delta_0 + \delta_1 lgdp60_i + \mathbf{z}_i'\boldsymbol{\delta}_2) + \varepsilon_i, \tag{6.1}$$

where $\mathrm{gr}_i$ is the annualized GDP growth rate of country $i$ from 1960 to 1985, $\mathrm{lgdp60}_i$ is the log GDP in 1960, $\mathbf{z}_i$ includes all 45 covariates listed in Table II, and $Q_i$ is a threshold variable, where we use the initial GDP in 1960. Because the estimation of the threshold parameter $\tau$ is not our target, we consider five different $\tau$'s in our analysis: 1655, 2073, 2898, 3268, and 6030. Among them, $\tau = 2898$ is a threshold value suggested by [33], and the other four threshold values are $k$th percentiles

**Table II.** List of variable.

| Variable | Description |
|---|---|
| Dependent variable | |
| gr | Annualized GDP growth rate in the period of 1960–85 |
| Threshold variables | |
| gdp60 | Real GDP per capita in 1960 (1985 price) |
| Covariates | |
| lgdp60 | log GDP per capita in 1960 (1985 price) |
| lsk | Log(Investment/Output) annualized over 1960–85; a proxy for the log physical savings rate |
| lgrpop | Log population growth rate annualized over 1960–1985 |
| pyrm60 | Log average years of primary schooling in the male population in 1960 |
| pyrf60 | Log average years of primary schooling in the female population in 1960 |
| syrm60 | Log average years of secondary schooling in the male population in 1960 |
| syrf60 | Log average years of secondary schooling in the female population in 1960 |
| hyrm60 | Log average years of higher schooling in the male population in 1960 |
| hyrf60 | Log average years of higher schooling in the female population in 1960 |
| nom60 | Percentage of no schooling in the male population in 1960 |
| nof60 | Percentage of no schooling in the female population in 1960 |
| prim60 | Percentage of primary schooling attained in the male population in 1960 |
| prif60 | Percentage of primary schooling attained in the female population in 1960 |
| pricm60 | Percentage of primary schooling complete in the male population in 1960 |
| pricf60 | Percentage of primary schooling complete in the female population in 1960 |
| secm60 | Percentage of secondary schooling attained in the male population in 1960 |
| secf60 | Percentage of secondary schooling attained in the female population in 1960 |
| seccm60 | Percentage of secondary schooling complete in the male population in 1960 |
| seccf60 | Percentage of secondary schooling complete in the female population in 1960 |
| llife | Log of life expectancy at age 0 averaged over 1960–1985 |
| lfert | Log of fertility rate (children per woman) averaged over 1960–1985 |
| edu/gdp | Government expenditure on eduction per GDP averaged over 1960–1985 |
| gcon/gdp | Government consumption expenditure net of defence and education per GDP averaged over 1960–85 |
| revol | The number of revolutions per year over 1960–84 |
| revcoup | The number of revolutions and coups per year over 1960–84 |
| wardum | Dummy for countries that participated in at least one external war over 1960–84 |
| wartime | The fraction of time over 1960–1985 involved in external war |
| lbmp | Log(1+black market premium averaged over 1960–85) |
| tot | The term of trade shock |
| lgdp60 | 'educ' Product of two covariates (interaction of lgdp60 and education variables from pyrm60 to seccf60); total 16 variables |

**Table III.** Sizes of selected submodel.

| $\tau$ | 6030 | 3268 | 2898 | 2073 | 1655 |
|---|---|---|---|---|---|
| Lasso | 15 | 18 | 18 | 19 | 11 |
| adaptive Lasso | 19 | 13 | 20 | 19 | 11 |

for $k = 60, 70, 80, 90$, respectively. After removing all missing data, each setting includes $n = 82$ observations and $p = 90$ covariates besides two intercepts.

Before applying the post selection shrinkage strategy, we first obtain candidate subsets from two variable selection techniques: Lasso and adaptive Lasso, respectively. All tuning parameters are selected from fivefold cross validation. In Table III, we list the numbers of selected important variables, $\hat{s}_1 = |\hat{S}_1|$, and also the sizes of candidate submodels, under five different $\tau$'s. In Table IV, we list the frequency of each variable being selected among all five settings. We observe

**Table IV.** Frequency of selected variables (based upon either $\beta_j \neq 0$ or $\delta_j \neq 0$) among All 5 $\tau$'s.

| | Lasso | | ALasso | |
|---|---|---|---|---|
| Variable | $\#(\beta_j \neq 0)$ | $\#(\delta_j \neq 0)$ | $\#(\beta_j \neq 0)$ | $\#(\delta_j \neq 0)$ |
| lgdp60 | 5 | 0 | 5 | 0 |
| lsk | 5 | 0 | 5 | 0 |
| nom60 | 0 | 1 | 0 | 1 |
| prim60 | 3 | 0 | 3 | 0 |
| pricm60 | 3 | 3 | 3 | 3 |
| seccm60 | 0 | 5 | 0 | 5 |
| seccf60 | 1 | 0 | 1 | 0 |
| llife | 5 | 0 | 5 | 0 |
| lfert | 5 | 0 | 5 | 0 |
| edugdp | 3 | 0 | 4 | 0 |
| gcongdp | 5 | 0 | 5 | 0 |
| revol | 2 | 0 | 3 | 0 |
| wardum | 2 | 3 | 2 | 3 |
| wartime | 4 | 4 | 3 | 3 |
| lbmp | 5 | 0 | 5 | 0 |
| tot | 0 | 5 | 0 | 5 |
| lgdpsyrm60 | 2 | 0 | 2 | 0 |
| lgdphyrm60 | 3 | 0 | 1 | 0 |
| lgdphyrf60 | 0 | 1 | 1 | 0 |
| lgdpnof60 | 0 | 3 | 0 | 3 |
| lgdpprim60 | 2 | 0 | 2 | 1 |
| lgdpprif60 | 0 | 1 | 0 | 2 |
| lgdpseccf60 | 1 | 0 | 0 | 0 |

**Table V.** Estimation results under $\tau = 2898$ (Candidate submodel from ALasso).

| Variable | $\hat{\beta}^{(ALasso)}$ | $\hat{\delta}^{(ALasso)}$ | $\hat{\beta}^{(PSE)}$ | $\hat{\delta}^{(PSE)}$ |
|---|---|---|---|---|
| lgdp60 | $-9.253 \times 10^{-3}$ | — | $-1.287 \times 10^{-2}$ | — |
| lsk | $6.121 \times 10^{-4}$ | — | $3.942 \times 10^{-4}$ | — |
| nom60 | — | $1.400 \times 10^{-2}$ | — | $3.481 \times 10^{-2}$ |
| prim60 | $-4.579 \times 10^{-2}$ | — | $-7.472 \times 10^{-2}$ | — |
| pricm60 | $1.934 \times 10^{-2}$ | $1.974 \times 10^{-3}$ | $4.129 \times 10^{-2}$ | $7.058 \times 10^{-3}$ |
| seccm60 | — | $4.903 \times 10^{-4}$ | — | $4.324 \times 10^{-4}$ |
| llife | $1.200 \times 10^{-3}$ | — | $2.212 \times 10^{-3}$ | — |
| lfert | $-1.659 \times 10^{-3}$ | — | $-1.507 \times 10^{-3}$ | — |
| edugdp | $2.228 \times 10^{-5}$ | — | $2.309 \times 10^{-5}$ | — |
| gcongdp | $-2.351 \times 10^{-4}$ | — | $-2.610 \times 10^{-4}$ | — |
| revol | $-1.020 \times 10^{-6}$ | — | $-1.158 \times 10^{-4}$ | — |
| wardum | — | $-1.417 \times 10^{-4}$ | — | $-3.336 \times 10^{-4}$ |
| wartime | $-1.655 \times 10^{-4}$ | — | $-5.081 \times 10^{-5}$ | — |
| lbmp | $-1.580 \times 10^{-3}$ | — | $-1.595 \times 10^{-3}$ | — |
| tot | — | $5.202 \times 10^{-6}$ | — | $6.318 \times 10^{-6}$ |
| lgdphyrm60 | $1.122 \times 10^{-2}$ | — | $4.291 \times 10^{-2}$ | — |
| lgdphyrf60 | $-7.585 \times 10^{-3}$ | — | $-4.143 \times 10^{-2}$ | — |
| lgdpnof60 | — | $6.392 \times 10^{-2}$ | — | 0.189 |
| lgdpprif60 | — | $-3.130 \times 10^{-2}$ | — | −0.127 |

Copyright © 2016 John Wiley & Sons, Ltd.

that Lasso and adaptive Lasso variable selection results in Table IV are quite close for this data set. However, the selected candidate subset model can be quite different among all five different $\tau$'s.

After the variable selection, post selection PSE is applied based upon the selected candidate subsets in all settings. Tables V and VI give the estimation results for $\tau = 2898$ and $\tau = 1655$, where both candidate subsets are selected by the adaptive Lasso. We omit results under other settings to save the space.

Becuase we do not know what the true model is in the real-data analysis, we first evaluate the prediction improvement from variable selection estimates to post selection PSEs by computing the relative residual sum of squares (RRSS) of the estimator $\boldsymbol{\beta}^{\diamond}_{\mathcal{J}}$ over the WR estimator $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\mathcal{J}}$ as follows:

$$\mathrm{RRSS}(\boldsymbol{\beta}^{\diamond}_{\mathcal{J}}) = \frac{\sum_{i=1}^{n} \|\mathbf{y} - \sum_{j\in\mathcal{J}} \mathbf{X}_{\mathcal{J}} \hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\mathcal{J}}\|^2}{\sum_{i=1}^{n} \|\mathbf{y} - \sum_{j\in\mathcal{J}} \mathbf{X}_{\mathcal{J}} \boldsymbol{\beta}^{\diamond}_{\mathcal{J}}\|^2}, \tag{6.2}$$

where $\mathcal{J}$ is the index of the submodel chosen by corresponding variable selection methods, and $\boldsymbol{\beta}^{\diamond}_{\mathcal{J}}$ can be (adaptive) Lasso and the corresponding generated post selection SEs and post selection PSEs. Similar to the simulation studies, RRSS > 1 indicates the superiority of $\boldsymbol{\beta}^{\diamond}_{\mathcal{J}}$ over $\hat{\boldsymbol{\beta}}_{\mathcal{J}}$. The results on RRSS for different $\tau$'s are reported in Figure 2, where the left and right panels are based upon Lasso and adaptive Lasso submodels, respectively. Those RRSS values of post selection REs give the highest value in both cases. This is not surprising because we assume the selected submodel is the right one and does not account for any bias. In both cases, the post selection PSEs dominate the corresponding variable selection estimation in terms of the RRSS regardless of whether Lasso or adaptive Lasso is used for generating the candidate submodel. This is

**Table VI.** Estimation results under $\tau = 1655$ (Candidate submodel from ALasso).

| Variable | $\hat{\beta}^{(ALasso)}$ | $\hat{\delta}^{(ALasso)}$ | $\hat{\beta}^{(PSE)}$ | $\hat{\delta}^{(PSE)}$ |
|---|---|---|---|---|
| lgdp60 | $-2.841 \times 10^{-3}$ | — | $-1.306 \times 10^{-2}$ | — |
| lsk | $1.319 \times 10^{-3}$ | — | $1.284 \times 10^{-3}$ | — |
| seccm60 | — | $3.652 \times 10^{-4}$ | — | $5.873 \times 10^{-4}$ |
| llife | $3.532 \times 10^{-4}$ | — | $1.633 \times 10^{-3}$ | — |
| lfert | $-2.552 \times 10^{-4}$ | — | $-2.250 \times 10^{-3}$ | — |
| gcongdp | $-1.554 \times 10^{-4}$ | — | $-3.033 \times 10^{-4}$ | — |
| revol | $-3.715 \times 10^{-5}$ | — | $-9.248 \times 10^{-4}$ | — |
| wartime | $-4.965 \times 10^{-5}$ | $-1.120 \times 10^{-5}$ | $2.731 \times 10^{-4}$ | $-3.958 \times 10^{-5}$ |
| lbmp | $-1.428 \times 10^{-3}$ | — | $-5.887 \times 10^{-4}$ | — |
| tot | — | $5.175 \times 10^{-7}$ | — | $8.476 \times 10^{-6}$ |

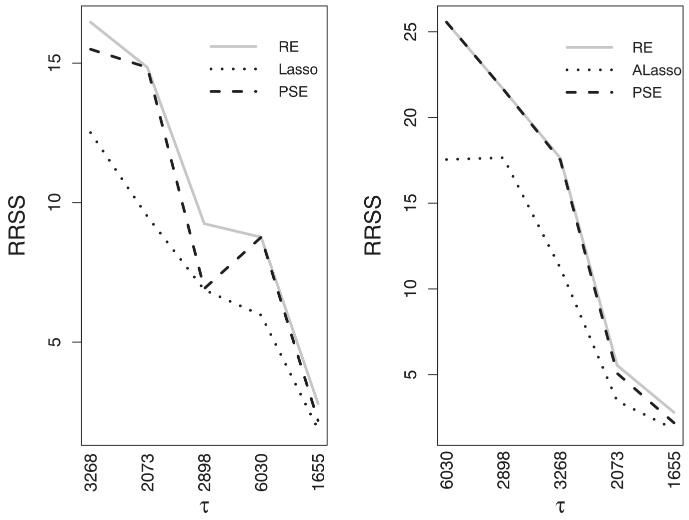

**Figure 2.** Relative residual sum of squares (RRSS) from (6.2) from post selection post selection shrinkage estimator (PSE) and the Lasso-type estimators: Lasso (left panel) or adaptive Lasso (ALasso) (right panel). The curve is plotted based upon a decreasing order of RRSS for better visibility, with corresponding values of $\tau$ plotted in $x$-axis.

Copyright © 2016 John Wiley & Sons, Ltd.

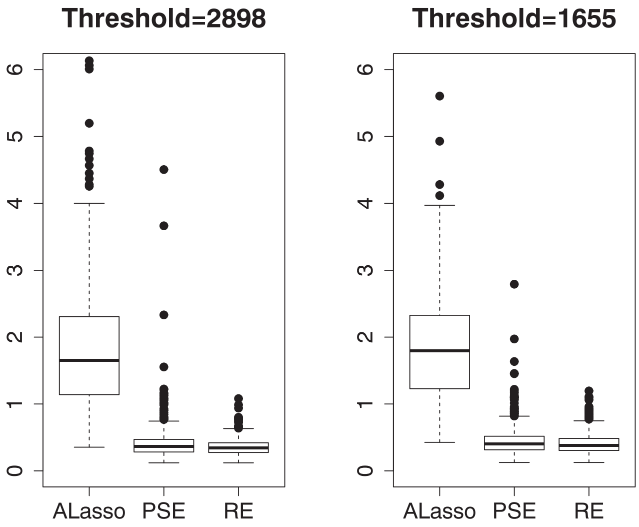

**Figure 3.** Prediction errors from post selection post selection shrinkage estimator (PSE), restricted estimator (RE), and adaptive Lasso (ALasso). Left: $\tau = 2898$; Right: $\tau = 1655$. All prediction errors are computed using cross validation following 500 random partitions of the data set. In each partition, the training set consists of 2/3 observations, and the test set consists of the remaining 1/3 observations.

because shrinkage estimation provides a better trade-off between bias and variance when selected submodels underfit the true model.

In addition, we also obtain prediction errors using cross validation following 500 random partitions of the data set. In each partition, the training set consists of 2/3 observations (size 55), and the test set consists of the remaining 1/3 observations (size 28). Corresponding results for $\tau = 2898$ and 1655 are reported in Figure 3, where the post selection PSEs are compared with the adaptive Lasso. The comparisons between the post selection PSEs and (adaptive) Lasso for other $\tau$'s follow the similar pattern and thus are omitted. It is observed that post selection PSEs produce much smaller prediction errors than the Lasso-type estimation.

## 7. Conclusion and discussions

In this paper, we generalize the shrinkage estimation to a high-dimensional sparse regression model. We propose a post selection shrinkage estimation strategy by shrinking a WR estimator in the direction of a candidate submodel obtained by existing PLSs variable selection methods.

When $p_n$ grows with $n$ quickly, it is reasonable to assume that the model sparsity exists in the sense that most covariates do not contribute. However, at the same time, some covariates may still make some small but jointly non-trivial contribution to the response. Existing penalized regularization approaches usually lead to a sparse model but tend to miss the possible small contributions from some covariates, resulting in excessive prediction errors or inefficient estimation. Our proposed post selection shrinkage strategy, taking into account possible contributions of covariates with weak and/or moderate signals, has dominant prediction performances over candidate submodel estimates generated from Lasso-type methods.

Before obtaining a shrinkage estimator, one key step is to generate a full estimation of $\boldsymbol{\beta}_n$ when $p \gg n$. We suggest a post selection WR estimator which is able to separate small coefficients from zero coefficients. The advantages of proposed post selection PSE are studied both theoretically and numerically. In theory, we established the asymptotic normality of the post selection WR estimator when $p_n$ grows with $n$ at an almost exponential rate such that $\log(p_n) = O(n^{\nu})$ for some $0 < \nu < 1$. Those novel asymptotic properties are used for investigating the asymptotic efficiency of the proposed post selection PSE analytically. In numerical studies, we chose tuning parameters $c_1$ and $c_2$ from cross validation but cannot guarantee their optimality for post selection PSE. The choice of tuning parameters is an important but challenging issue in high-dimensional data analysis that could potentially create very important future work. Although the proposed post selection PSE was presented based on a WR method, other methods can also be used to generate the shrinkage estimator.

Finally, we acknowledge the importance of Lasso-type variable selection methods, but at the same time, and do not depend completely on them, especially when many weak coefficients jointly affect the response variable. The Lasso is the start but not the end. We could potentially still make some significant prediction improvements. We hope this work will shed some more light on the investigation of the post variable selection shrinkage analysis in high-dimensional data analysis.

Copyright © 2016 John Wiley & Sons, Ltd.

## Appendix

All technical proofs are given in this section.

*Proof of Theorem 1*

After solving (3.1), we obtain

$$\tilde{\boldsymbol{\beta}}_{\hat{S}_1}(r_n) = \left(\mathbf{X}'_{\hat{S}_1} \boldsymbol{M}_{\hat{S}_1^c}(r_n) \mathbf{X}_{\hat{S}_1}\right)^{-1} \mathbf{X}'_{\hat{S}_1} \boldsymbol{M}_{\hat{S}_1^c}(r_n) \mathbf{y} \tag{A1}$$

and

$$\tilde{\boldsymbol{\beta}}_{\hat{S}_1^c}(r_n) = \left(r_n \mathbf{I}_{p_{2n}} + \mathbf{X}'_{\hat{S}_1^c} \boldsymbol{M}_{\hat{S}_1} \mathbf{X}_{\hat{S}_1^c}\right)^{-1} \mathbf{X}'_{\hat{S}_1^c} \boldsymbol{M}_{\hat{S}_1} \mathbf{y}, \tag{A2}$$

where $\boldsymbol{M}_{\hat{S}_1^c}(r_n) = \mathbf{I}_n - \mathbf{X}_{\hat{S}_1^c} \left(r_n \mathbf{I}_{p_{2n}} + \mathbf{X}'_{\hat{S}_1^c} \mathbf{X}_{\hat{S}_1^c}\right)^{-1} \mathbf{X}'_{\hat{S}_1^c}$ and $\boldsymbol{M}_{\hat{S}_1} = \mathbf{I}_n - \mathbf{X}_{\hat{S}_1} (\mathbf{X}'_{\hat{S}_1} \mathbf{X}_{\hat{S}_1})^{-1} \mathbf{X}'_{\hat{S}_1}$.

We only need to prove the result under the condition $\hat{S}_1 = S_1$, and then all matrices, vectors indexed by $\hat{S}_1$ can be replaced by $S_1$ or 1 without causing of any confusion. For example, $\boldsymbol{M}_{\hat{S}_1} = \boldsymbol{M}_{S_1} = \boldsymbol{M}_1$ under the condition.

First, we check the bias of $\hat{\boldsymbol{\beta}}^{\text{WR}}_{S_1^c}$. Because $\boldsymbol{M}_1$ is an idempotent matrix, $\boldsymbol{M}_1 \mathbf{X}_{1n} = 0$. Denote $q_n = p_{2n} + p_{3n}$. Then,

$$(\mathbf{X}'_{S_1^c} \boldsymbol{M}_1 \mathbf{X}_{S_1^c} + r_n \mathbf{I}_{q_n})^{-1} \mathbf{X}'_{S_1^c} \boldsymbol{M}_1 \mathbf{X}_{1n} \boldsymbol{\beta}_{10} = \mathbf{0}.$$

Let $\mathbf{Q}$ be a $q_n \times q_n$ orthogonal matrix such that

$$\mathbf{U}' \boldsymbol{M}_1 \mathbf{U} = \mathbf{X}'_{S_1^c} \boldsymbol{M}_1 \mathbf{X}_{S_1^c} = \mathbf{Q} \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q}',$$

where $\mathbf{D} = \text{diag}\{\varrho_{1n}, \dots, \varrho_{k_n n}\}$. Then, we have

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}^{\text{WR}}_{S_1^c}\right) - \boldsymbol{\beta}^*_{S_1^c} &= \left(\mathbf{X}'_{S_1^c} \boldsymbol{M}_1 \mathbf{X}_{S_1^c} + r_n \mathbf{I}_{q_n}\right)^{-1} \mathbf{X}'_{S_1^c} \boldsymbol{M}_1 \mathbf{y} - \boldsymbol{\beta}^*_{S_1^c} \Big] \\ &= \left(\mathbf{X}'_{S_1^c} \boldsymbol{M}_1 \mathbf{X}_{S_1^c} + r_n \mathbf{I}_{q_n}\right)^{-1} \mathbf{X}'_{S_1^c} \boldsymbol{M}_1 \mathbf{X}_{S_1^c} \boldsymbol{\beta}^*_{S_1^c} - \boldsymbol{\beta}^*_{S_1^c} \\ &= -r_n \left(\mathbf{X}'_{S_1^c} \boldsymbol{M}_1 \mathbf{X}_{S_1^c} + r_n \mathbf{I}_{q_n}\right)^{-1} \boldsymbol{\beta}^*_{S_1^c} \\ &= -\mathbf{Q} \begin{pmatrix} \left(\mathbf{I}_{k_n} + r_n^{-1} \mathbf{D}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{q_n - k_n} \end{pmatrix} \mathbf{Q}' \boldsymbol{\beta}^*_{S_1^c}. \end{aligned} \tag{A3}$$

Suppose that $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ and $\mathbf{Q}_1$ is a $q_n \times k_n$ matrix. Notice that $\mathbf{Q}\mathbf{Q}' = \mathbf{Q}'\mathbf{Q} = \mathbf{I}_{q_n}$. Then, $\mathbf{Q}'_1 \mathbf{Q}_1 = \mathbf{I}_{k_n}$, $\mathbf{Q}'_1 \mathbf{Q}_2 = \mathbf{0}$, and $\mathbf{Q}_2 \mathbf{Q}'_2$ is a projection matrix. Let $\theta^* = \mathbf{Q}_1 \mathbf{Q}'_1 \boldsymbol{\beta}^*_{S_1^c}$. Then,

$$\boldsymbol{\beta}^*_{S_1^c} = \mathbf{Q}_1 \mathbf{Q}'_1 \theta^* = \mathbf{Q}_1 \mathbf{Q}'_1 \mathbf{Q}_1 \mathbf{Q}'_1 \boldsymbol{\beta}^*_{S_1^c} = \mathbf{Q}_1 \mathbf{Q}'_1 \boldsymbol{\beta}^*_{S_1^c}. \tag{A4}$$

Replace $\boldsymbol{\beta}^*_{S_1^c}$ in (A3) by $\mathbf{Q}_1 \mathbf{Q}'_1 \boldsymbol{\beta}^*_{S_1^c}$, we have

$$E\left(\hat{\boldsymbol{\beta}}^{\text{WR}}_{S_1^c}\right) - \boldsymbol{\beta}_{0S_1^c} = -\mathbf{Q}_1 \left(\mathbf{I}_{k_n} + r_n^{-1} \mathbf{D}\right)^{-1} \mathbf{Q}'_1 \boldsymbol{\beta}^*_{S_1^c}.$$

Thus,

$$\left\| E\left(\hat{\boldsymbol{\beta}}^{\text{WR}}_{S_1^c}\right) - \boldsymbol{\beta} *_{S_1^c} \right\|^2 = \theta'_0 \mathbf{Q}_1 \left(\mathbf{I}_{k_n} + r_n^{-1} \mathbf{D}\right)^{-2} \mathbf{Q}_1 \boldsymbol{\theta}_0 \leqslant (1 + \varrho_{1n}/r_n)^{-2} \|\boldsymbol{\beta}_{0S_1^c}\|^2$$

Copyright © 2016 John Wiley & Sons, Ltd.

For every $j \notin S_1$, $\left|\text{bias}\left(\hat{\beta}_j^{\text{WR}}\right)\right| \leqslant \left\|E\left(\hat{\boldsymbol{\beta}}_{S_1^c}^{\text{WR}}\right) - \boldsymbol{\beta}_{0S_1^c}\right)\right\|$, and thus,

$$\left|\text{bias}\left(\hat{\beta}_j^{\text{WR}}\right)\right| \leqslant (1 + \varrho_{1n}/r_n)^{-1}\|\boldsymbol{\beta}_{0S_1^c}\| \leqslant (r_n/\varrho_{1n})O(n^{\tau}) \leqslant O(r_n n^{\tau-\eta}).$$

The rest of the proof just mimics the proof of Theorem 2 in [28]. We will provide some outlines of the proof. If we let $r_n = c_2 a_n^{-2}(\log\log n)^3 \log(n \vee p)$ and $\log(p_n) = O(n^{\nu})$ in (B3), then for $u_n = 1 + (\log\log n)^{-1}$, we have

$$\frac{\left|\text{bias}\left(\hat{\beta}_j^{\text{WR}}\right)\right|}{a_n(u_n - 1)} \leqslant \frac{r_n n^{\tau-\eta}}{a_n(u_n - 1)} \leqslant \frac{c_2(\log\log n)^4}{a_n^3 n^{\eta-\tau-\nu}} \leqslant \frac{c_2(\log\log n)^4}{c_1^3 n^{\eta-\tau-\nu-3\alpha}} \to 0 \quad \text{if } 3\alpha < \eta - \nu - \tau,$$

where the last '$\leqslant$' is from (3.5) and $c_1$ is defined there. From the normal assumption of $\varepsilon_i$ and the solution in linear expression in (A2), we know $\hat{\boldsymbol{\beta}}_{S_1^c}^{\text{WR}}$ is normally distributed and

$$\begin{aligned}
\text{Var}\left(\hat{\boldsymbol{\beta}}_{S_1^c}^{\text{WR}}\right) &= \sigma^2\left(\mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{S_1^c} + r_n\mathbf{I}_{q_n}\right)^{-1}\mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{S_1^c}\left(\mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{S_1^c} + r_n\mathbf{I}_{q_n}\right)^{-1} \\
&\preceq \sigma^2\left(\mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{S_1^c} + r_n\mathbf{I}_{q_n}\right)^{-1} \\
&\preceq \sigma^2 r_n^{-1}\mathbf{I}_{q_n},
\end{aligned}$$

where '$\mathbf{A} \preceq \mathbf{B}$' means $\mathbf{B} - \mathbf{A}$ is a non-negative definite matrix. Thus, for any $j \notin S_1$, $\text{Var}\left(\hat{\beta}_j^{\text{WR}}\right) = O(1/r_n)$. Notice that $\sqrt{r_n}a_n(u_n - 1) = O\left((\log\log n)^{1/2}\right) \to \infty$. We have

$$\frac{a_n(u_n - 1)}{\sqrt{\text{Var}\left(\hat{\beta}_j^{\text{WR}}\right)}} \geqslant a_n(u_n - 1)\sqrt{r_n} \to \infty.$$

$$\begin{aligned}
P\left(|\hat{\beta}_j^{\text{WR}} - \beta_j^*| > a_n(u_n - 1)\right) &\leqslant P\left(|N(0,1)| > \frac{a_n(u_n - 1)}{\sqrt{\text{Var}(\hat{\beta}_j^{\text{WR}})}} - \frac{|\text{bias}(\hat{\beta}_j^{\text{WR}})|}{\sqrt{\text{Var}(\hat{\beta}_j^{\text{WR}})}}\right) \\
&= 2\Phi\left(\frac{|\text{bias}(\hat{\beta}_j^{\text{WR}})| - a_n(u_n - 1)}{\sqrt{\text{Var}(\hat{\beta}_j^{\text{WR}})}}\right) \\
&\leqslant 2\Phi\left(-c_0\sqrt{r_n}a_n/(\log\log n)\right) \\
&\leqslant \exp\{-c_0^2 r_n a_n^2/(\log\log n)^2\},
\end{aligned}$$

where $\Phi$ is the cumulative distribution function of a standard normal random variable, $c_0 > 0$ is a constant, '$\leqslant$' is the tail probability of a normal random variable. Thus,

$$\begin{aligned}
&P\left(\{j \notin S_1 : |\beta_j^*| > a_n u_n\} \subset \{j \notin S_1 : |\hat{\beta}_j^{\text{WR}}| > a_n\}\right) \\
&\geqslant 1 - P\left(\bigcup_{j:|\beta_j^*|>a_n u_n}\{|\hat{\beta}_j^{\text{WR}}| \leqslant a_n\}\right) \\
&\geqslant 1 - P\left(\bigcup_{j:|\beta_j^*|>a_n u_n}\{|\hat{\beta}_j^{\text{WR}} - \beta_j^*| \leqslant a_n(u_n - 1)\}\right) \\
&\geqslant 1 - \sum_{j\notin S_1} P\left(|\hat{\beta}_j^{\text{WR}} - \beta_j^*| > a_n(u_n - 1)\right) \\
&\geqslant 1 - q_n \exp\{-c_0^2 r_n a_n^2/(\log\log n)^2\} \\
&\geqslant 1 - \exp\{-\left(c_0^2 r_n a_n^2/(\log\log n)^2 - \log(p_n)\right)\} \\
&\geqslant 1 - \exp\{-(c_0^2 \log\log n - 1)\log(p_n \vee n)\}.
\end{aligned}$$

When $n$ is large enough, there exists $c_0^2 \log\log n - 1 > t > 0$ for some $t > 0$. Thus,

$$\lim_{n\to\infty} P\left(\{j \notin S_1 : |\beta_j^*| > a_n u_n\} \subset \{j \notin S_1 : |\hat{\beta}_j^{\mathrm{WR}}| > a_n\}\right) \geqslant 1 - (p_n \vee n)^{-t} \to 1.$$

Similarly, we have

$$\lim_{n\to\infty} P\left(\{j \notin S_1 : |\beta_j^*| > a_n/u_n\} \supset \{j \notin S_1 : |\hat{\beta}_j^{\mathrm{WR}}| > a_n\}\right) \geqslant 1 - (p_n \vee n)^{-t} \to 1.$$

Because of the continuity of $\hat{\beta}_j^{\mathrm{WR}}$ and $\lim_{n\to\infty} u_n = 1$, we have

$$\lim_{n\to\infty} P(\hat{S}_2 | \hat{S}_1 = S_1) = 1.$$

□

*Proof of Corollary 1*

Because $S_1 \subset \hat{S}_1$, a weighted ridge estimator $\hat{\boldsymbol{\beta}}_{\hat{S}_1}$ aims to find some weak signals from $\hat{S}_1^c \cap S_2$. Because $\hat{S}_1^c \subset S_1^c$, the smallest positive eigenvalues of $\mathbf{X}'_{\hat{S}_1^c} \boldsymbol{M}_{\hat{S}_1} \mathbf{X}_{\hat{S}_1}$ must be larger than $\lambda_{1n}$, and $\|\boldsymbol{\beta}^*_{\hat{S}_1^c}\|_2 \leqslant \|\boldsymbol{\beta}^*_{S_1^c}\|_2$. Thus, we can borrow the proof of Theorem 1 here, by treating $\hat{S}_1$ and $S_2 \cap \hat{S}_1^c$ as the new $S_1$ and $S_2$. □

*Proof of Theorem 2*

Similar to the proof in Theorem 1, we assume $\hat{S}_1 = S_1$. Then, the penalized quadratic loss function in (3.1) becomes

$$L(\boldsymbol{\beta}_n; S_1) = \left\{ \|\mathbf{y} - \mathbf{X}_n \boldsymbol{\beta}\|^2 + r_n \|\boldsymbol{\beta}_{S_1^c}\|^2 \right\}.$$

Therefore, $\hat{\boldsymbol{\beta}}_n^{\mathrm{WR}} = \arg\ \min\{L(\boldsymbol{\beta}_n; S_1)\}$ satisfies,

$$\frac{\partial L\left(\hat{\boldsymbol{\beta}}_n^{\mathrm{WR}}\right)}{\partial \boldsymbol{\beta}_{S_3^c}} = \mathbf{0}.$$

From the notation $\mathbf{X}'_{S_3^c} = \mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$. If we write $\mathbf{X}'_{S_3} = (\mathbf{w}_1, \dots, \mathbf{w}_n)$, then

$$-\sum_{i=1}^n \left(y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}_{S_3^c}^{\mathrm{WR}} - \mathbf{w}_i'\hat{\boldsymbol{\beta}}_{S_3}^{\mathrm{WR}}\right)\mathbf{z}_i + r_n \begin{pmatrix} \mathbf{0}_{p_{1n}} \\ \hat{\boldsymbol{\beta}}_{S_2}^{\mathrm{WR}} \end{pmatrix} = \mathbf{0}_{p_{1n}+p_{2n}}.$$

Replacing $y_i$ by $\mathbf{z}_i'\boldsymbol{\beta}_{0S_3^c} + \mathbf{w}_i'\boldsymbol{\beta}_{0S_3} + \varepsilon_i$, we have

$$-\sum_{i=1}^n \left(\varepsilon_i - \mathbf{z}_i'\left(\hat{\boldsymbol{\beta}}_{S_3^c}^{\mathrm{WR}} - \boldsymbol{\beta}_{0S_3^c}\right) - \mathbf{w}_i'\left(\hat{\boldsymbol{\beta}}_{S_3}^{\mathrm{WR}} - \boldsymbol{\beta}_{0S_3}\right)\right)\mathbf{z}_i + r_n \begin{pmatrix} \mathbf{0}_{p_{1n}} \\ \hat{\boldsymbol{\beta}}_{S_2}^{\mathrm{WR}} \end{pmatrix} = \mathbf{0}.$$

Notice that $\boldsymbol{\Sigma}_n = n^{-1}\sum_{i=1} \mathbf{z}_i\mathbf{z}_i' \to \boldsymbol{\Sigma}$. Thus,

$$\begin{aligned} n^{1/2}\mathbf{d}_n'\left(\hat{\boldsymbol{\beta}}_{S_3^c}^{\mathrm{WR}} - \boldsymbol{\beta}_{S_3^c}^*\right) = {} & n^{-1/2}\sum_{i=1}^n \mathbf{d}_n'\varepsilon_i\boldsymbol{\Sigma}_n^{-1}\mathbf{z}_i - n^{-1/2}r_n\mathbf{d}_n'\boldsymbol{\Sigma}_n^{-1}\begin{pmatrix} \mathbf{0}_{p_{1n}} \\ \hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}} \end{pmatrix} \\ & - n^{-1/2}\sum_{i=1}^n \mathbf{d}_n'\mathbf{w}_i'\left(\hat{\boldsymbol{\beta}}_{3n}^{\mathrm{WR}} - \boldsymbol{\beta}_{S_3}^*\right)\boldsymbol{\Sigma}_n^{-1}\mathbf{z}_i \end{aligned} \qquad \text{(H5)}$$

Copyright © 2016 John Wiley & Sons, Ltd.

Under conditions (B1–B3), with probability 1, $\hat{\boldsymbol{\beta}}_{S_3}^{\text{WR}} = \mathbf{0}$ from Theorem 1. Therefore, the third term in (H5) is zero. By abusing the notation, if we rewrite $\mathbf{d}_n = (\mathbf{d}_{1n}', \mathbf{d}_{2n}')'$, then

$$\begin{aligned} n^{-1/2} r_n \mathbf{d}_n' \boldsymbol{\Sigma}_n^{-1} \begin{pmatrix} \mathbf{0}_{p_{1n}} \\ \hat{\boldsymbol{\beta}}_{S_2}^{\text{WR}} \end{pmatrix} &\leqslant \rho_1^{-1} n^{-1/2} r_n \mathbf{d}_{2n}' \hat{\boldsymbol{\beta}}_{S_2}^{\text{WR}} \\ &= O_P(\rho_1^{-1} n^{-1/2} r_n \mathbf{d}_{2n}' \boldsymbol{\beta}_{S_2}^{*}) \\ &\leqslant O_P\left(\rho_1^{-1} r_n n^{-1/2} \|\mathbf{d}_{2n}\| \|\boldsymbol{\beta}_{S_2}^{*}\|\right) \\ &\leqslant O_P\left(\rho_1^{-1} r_n n^{-(1/2-\tau)}\right) = o_P(1), \end{aligned}$$

where the first '$\leqslant$' is from (B4), the first '=' is from (A2) and (B1), the second '$\leqslant$' is from the Cauchy–Schwarz inequality, the third '$\leqslant$' is from (A2). The last '=' holds because $r_n = o(n^{1/2-\tau})$ if we choose $r_n = c_2 a_n^{-2} (\log\log n)^3 \log(n \vee p_n)$ with $a_n = c_1 n^{-\alpha}$ for $\alpha < 1/4 - \tau/2$ for $0 < \tau < 1/2$. Therefore,

$$\lim_{n\to\infty} n^{1/2} s_n^{-1} \mathbf{d}_n' \left(\hat{\boldsymbol{\beta}}_{S_3^c}^{\text{WR}} - \boldsymbol{\beta}_{S_3^c}^{*}\right) = \lim_{n\to\infty} n^{-1/2} s_n^{-1} \sum_{i=1}^{n} \mathbf{d}_n' \varepsilon_i \boldsymbol{\Sigma}_n^{-1} \mathbf{z}_i \tag{H6}$$

Define $u_i = n^{-1/2} s_n^{-1} \mathbf{d}_n' \boldsymbol{\Sigma}_n^{-1} \mathbf{z}_i,\ 1 \leqslant i \leqslant n$. From (B1), we know that $\sum_{i=1}^{n} u_i \varepsilon_i$ is normal with variance,

$$\operatorname{Var}\left(\sum_{i=1}^{n} (u_i \varepsilon_i)\right) = \sigma^2 n^{-1} s_n^{-2} \mathbf{d}_n' \boldsymbol{\Sigma}_n^{-1} \left(\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i\right) \boldsymbol{\Sigma}_n^{-1} \mathbf{d}_n = 1.$$

□

*Proof of Theorem 3*

First, (4.9a) holds because we have

$$\lim_{n\to\infty} E\left[n^{1/2} s_{1n}^{-1} \mathbf{d}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \boldsymbol{\beta}_1^{*}\right)\right]^2 = E\left\{\lim_{n\to\infty} \left[n^{1/2} s_{1n}^{-1} \mathbf{d}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \boldsymbol{\beta}_1^{*}\right)\right]^2\right\} = E[Z^2] = 1,$$

where $Z \sim N(0, 1)$. We now verify (4.9b). Let $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}_{2n} \hat{\boldsymbol{\beta}}_{2n}^{\text{WR}} - \mathbf{X}_{3n} \hat{\boldsymbol{\beta}}_{3n}^{\text{WR}}$. Then,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} &= \arg\ \min\{\|\tilde{\mathbf{y}} - \mathbf{X}_{1n} \boldsymbol{\beta}_{1n}\|^2\} \\ &= (\mathbf{X}_{1n}' \mathbf{X}_{1n})^{-1} \mathbf{X}_{1n}' \tilde{\mathbf{y}} \\ &= \hat{\boldsymbol{\beta}}_{1n}^{\text{RE}} - (\mathbf{X}_{1n}' \mathbf{X}_{1n})^{-1} \mathbf{X}_{1n}' \mathbf{X}_{2n} \hat{\boldsymbol{\beta}}_{2n}^{\text{WR}} - (\mathbf{X}_{1n}' \mathbf{X}_{1n})^{-1} \mathbf{X}_{1n}' \mathbf{X}_{3n} \hat{\boldsymbol{\beta}}_{3n}^{\text{WR}} \\ &= \hat{\boldsymbol{\beta}}_{1n}^{\text{RE}} - (\mathbf{X}_{1n}' \mathbf{X}_{1n})^{-1} \mathbf{X}_{1n}' \mathbf{X}_{2n} \hat{\boldsymbol{\beta}}_{2n}^{\text{WR}}. \end{aligned} \tag{H7}$$

From the definition,

$$\begin{aligned} R\left(\mathbf{d}_{1n}' \hat{\boldsymbol{\beta}}_{1n}^{\text{RE}}\right) &= \lim\nolimits_{n\to\infty} E\left[n^{1/2} s_{1n}^{-1} \mathbf{d}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\text{RE}} - \boldsymbol{\beta}_1^{*}\right)\right]^2 \\ &= \lim\nolimits_{n\to\infty} s_{1n}^{-2} E\left\{n^{1/2} \mathbf{d}_{1n}' \left[\left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \boldsymbol{\beta}_1^{*}\right) - \left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\text{RE}}\right)\right]\right\}^2 \\ &= \lim\nolimits_{n\to\infty} E\left\{n^{1/2} s_{1n}^{-2} \mathbf{d}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \boldsymbol{\beta}_1^{*}\right)\right\}^2 + \lim\nolimits_{n\to\infty} E\left\{n^{1/2} s_{1n}^{-2} \mathbf{d}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\text{RE}}\right)\right\}^2 \\ &\quad - 2 \lim\nolimits_{n\to\infty} E\left\{n s_{1n}^{-2} \mathbf{d}_{1n}' \left(\mathbf{X}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\text{RE}}\right) \left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \boldsymbol{\beta}_1^{*}\right)' \mathbf{d}_{1n}\right\} \\ &= I_1 + I_2 + I_3. \end{aligned}$$

From (4.9a), we know $I_1 = \lim_{n\to\infty} E\left\{n^{1/2}s_{1n}^{-1}\mathbf{d}'_{1n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^*\right)\right\}^2 = 1$. From (H7),

$$\begin{aligned}
I_2 &= \lim_{n\to\infty} s_{1n}^{-2}E\left\{n^{1/2}\mathbf{d}'_{1n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right)\right\}^2 \\
&= \lim_{n\to\infty} s_{1n}^{-2}E\left\{n^{1/2}\mathbf{d}'_{1n}\boldsymbol{\Sigma}_{n11}^{-1}\boldsymbol{\Sigma}_{n12}\hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}}\right\}^2 \\
&= \lim_{n\to\infty} \left(s_{2n}^2/s_{1n}^2\right) E\left\{n^{1/2}s_{2n}^{-1}\mathbf{d}'_{2n}\hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}}\right\}^2,
\end{aligned}$$

where $\mathbf{d}_{2n} = \boldsymbol{\Sigma}_{n21}\boldsymbol{\Sigma}_{n11}^{-1}\mathbf{d}_{1n}$ and $s_{2n}^2 = \mathbf{d}'_{2n}\boldsymbol{\Sigma}_{n22.1}^{-1}\mathbf{d}_{2n}$. From Ouellette (1981) Equation (1.12), we obtain

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22.1}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} = \boldsymbol{\Sigma}_{11.2}^{-1} - \boldsymbol{\Sigma}_{11}^{-1}. \tag{H8}$$

Therefore,

$$s_{2n}^2 = \sigma^2\mathbf{d}'_{1n}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22.1}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{d}_{1n} = \sigma^2\mathbf{d}'_{1n}\boldsymbol{\Sigma}_{11.2}^{-1}\mathbf{d}_{1n} - \sigma^2\mathbf{d}'_{1n}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{d}_{1n}.$$

Because $s_{2n}^2/s_{1n}^2 \to 1 - c$,

$$I_2 = (1-c)\lim_{n\to\infty} E\left[\chi_1^2(\Delta_{\mathbf{d}_{1n}})\right] = (1-c)(1+\Delta_{\mathbf{d}_{1n}}),$$

where $\chi_v^2(t)$ is a $\chi^2$ distribution with degrees of freedom $v$ and noncentral parameter $t$. Here, $\Delta_{\mathbf{d}_{1n}}$ is given in (4.8). From the Cauchy–Schwarz inequality,

$$\Delta_{\mathbf{d}_{1n}} = s_{2n}^{-2}(\mathbf{d}'_{2n}\boldsymbol{\delta})^2 \leqslant \boldsymbol{\delta}'\boldsymbol{\Sigma}_{n22.1}\boldsymbol{\delta}.$$

Furthermore,

$$\begin{aligned}
I_3 &= -2\lim_{n\to\infty} E\left\{nS_{1n}^{-2}\mathbf{d}'_{1n}\left(\mathbf{X}'_{1n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right)\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^*\right)'\mathbf{d}_{1n}\right\} \\
&= 2\lim_{n\to\infty}\left[\mathbf{d}'_{1n}(\mathbf{I}_{p_{1n}}\ \ \mathbf{0}_{p_{1n}\times p_{2n}})\boldsymbol{\Sigma}_n^{-1}(\mathbf{0}'_{p_{1n}\times p_{2n}}\ \ \mathbf{I}_{p_{2n}})'\boldsymbol{\Sigma}_{n21}\boldsymbol{\Sigma}_{n11}^{-1}\right] \\
&= -2\lim_{n\to\infty}(s_{2n}/s_{1n})^2 = -2(1-c)
\end{aligned}$$

Thus, $R\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right) = I_1 + I_2 + I_3 = 1 - (1-\Delta_{\mathbf{d}_{1n}})(1-c)$. Thus, (4.9b) holds.

We now investigate (4.9c). First from the definition,

$$\begin{aligned}
R\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{SE}}\right) &= \lim_{n\to\infty} E[n^{1/2}s_{1n}^{-1}\mathbf{d}'_{1n}(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{SE}} - \boldsymbol{\beta}_1^*)]^2 \\
&= \lim_{n\to\infty} s_{1n}^{-2}E\left\{n^{1/2}\mathbf{d}'_{1n}\left[\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^*\right) - (\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}})((p_{2n}-2)/T_n)\right]\right\}^2 \\
&= \lim_{n\to\infty} E\left\{n^{1/2}s_{1n}^{-2}\mathbf{d}'_{1n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^*\right)\right\}^2 \\
&\quad - \Bigg(\lim_{n\to\infty} 2E\left\{ns_{1n}^{-2}((p_{2n}-2))/T_n)\mathbf{d}'_{1n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right)\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^*\right)'\mathbf{d}_{1n}\right\} \\
&\quad - \lim_{n\to\infty} E\left\{n^{1/2}s_{1n}^{-2}\mathbf{d}'_{1n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right)((p_{2n}-2)/T_n)\right\}^2\Bigg) \\
&= J_1 - (J_2 - J_3).
\end{aligned}$$

Again, $J_1 = \lim_{n\to\infty} E\left\{n^{1/2}s_{1n}^{-2}\mathbf{d}'_{1n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^*\right)\right\}^2 = 1$. From (H7),

$$\mathbf{d}'_{1n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right) = -\mathbf{d}'_{1n}\boldsymbol{\Sigma}_{n11}^{-1}\boldsymbol{\Sigma}_{n12}\hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}} = \mathbf{d}'_{2n}\hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}}.$$

Then, we have

$$\begin{aligned} J_2 - J_3 &= \lim_{n\to\infty} 2E\left\{ ns_{1n}^{-2}((p_{2n}-2)/T_n)\mathbf{d}'_{1n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\text{RE}}\right)\left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \boldsymbol{\beta}_1^*\right)' \mathbf{d}_{1n} \right\} \\ &\quad - \lim_{n\to\infty} E\left\{ n^{1/2}s_{1n}^{-2}\mathbf{d}'_{1n}(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\text{RE}})((p_{2n}-2)/T_n)\right\}^2 \\ &= -\lim_{n\to\infty} 2s_{1n}^{-2}E\left\{ ((p_{2n}-2)/T_n)\sqrt{n}\mathbf{d}'_{2n}\hat{\boldsymbol{\beta}}_{2n}^{\text{WR}}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \boldsymbol{\beta}_1^*\right)'\mathbf{d}_{1n}\right\} \\ &\quad - \lim_{n\to\infty} s_{1n}^{-2}E\left\{\left[((p_{2n}-2)/T_n)\sqrt{n}\mathbf{d}'_{2n}\hat{\boldsymbol{\beta}}_{2n}^{\text{WR}}\right]^2\right\} \end{aligned}$$

From Theorem 1, $\hat{s}_2 = p_{2n} + o_p(1)$ and

$$T_n = \left(\sqrt{n}\hat{\boldsymbol{\beta}}_{n2}^{\text{WR}}\right)'(\boldsymbol{\Sigma}_{n22.1})\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{n2}^{\text{WR}}\right)/\hat{\sigma}^2 + o_p(1).$$

We now define $\mathbf{a}' = \left(\mathbf{d}'_{1n}\ \mathbf{0}_{1\times p_{2n}}\right)$, $\mathbf{b}' = \left(\mathbf{0}_{p_{1n}\times 1}\ -\mathbf{d}_{2n}\right)$, and $\eta(\mathbf{x}) = ((p_{2n}-2)/(\mathbf{x}'\mathbf{W}\mathbf{x}))\mathbf{b}'\mathbf{x}$, where $\mathbf{W} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{n22.1}\end{pmatrix}$. Then, from the asymptotic normality,

$$J_2 - J_3 = \lim_{n\to\infty} s_{1n}^{-2}E[2\eta(\mathbf{z}+\boldsymbol{\zeta})\mathbf{a}'\mathbf{z} - (\eta(\mathbf{z}+\boldsymbol{\zeta}))^2],$$

where $\boldsymbol{\zeta}' = \left(\mathbf{0}_{p_{1n}\times 1}\ -\boldsymbol{\beta}'_{20}\right)$ and $\mathbf{z}$ satisfy that

$$\left(\mathbf{d}'_n\boldsymbol{\Sigma}_n^{-1/2}\mathbf{d}_n\right)^{-1}\mathbf{d}'_n\mathbf{z} \to N(0,1)$$

and

$$\lim_{n\to\infty}\left(\mathbf{d}'_n\boldsymbol{\Sigma}_n^{-1}\mathbf{d}_n\right)^{-1}(\mathbf{d}'_n\boldsymbol{\zeta})^2 = \lim_{n\to\infty}\Delta_{d_{1n}}.$$

From Stein's lemma, we have

$$\begin{aligned} E(\eta(\mathbf{z}+\boldsymbol{\zeta})\mathbf{a}'\mathbf{z}) &= \mathbf{a}'\boldsymbol{\Sigma}_n^{-1}(\partial\eta(\mathbf{z}+\boldsymbol{\zeta})/\partial\mathbf{z}) \\ &= \frac{(p_{2n}-2)\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{b}}{(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})} - \frac{2(p_{2n}-2)\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{b}}{((\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta}))^2} \end{aligned}$$

So we have

$$\begin{aligned} J_2 - J_3 &= \lim\nolimits_{n\to\infty} s_{1n}^{-2}E[2\eta(\mathbf{z}+\boldsymbol{\zeta})\mathbf{a}'\mathbf{z} - (\eta(\mathbf{z}+\boldsymbol{\zeta}))^2] \\ &= \lim\nolimits_{n\to\infty} s_{1n}^{-2}E\left\{\left[2\frac{(p_{2n}-2)\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{b}}{(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})} - 4\frac{(p_{2n}-2)\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{b}}{((\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta}))^2}\right]\right\} \\ &\quad - \lim\nolimits_{n\to\infty} s_{1n}^{-2}E\left\{\frac{(p_{2n}-2)^2\mathbf{b}'(\mathbf{z}+\boldsymbol{\zeta})(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{b}}{((\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta}))^2}\right\} \\ &= \lim\nolimits_{n\to\infty} E\left\{\frac{(p_{2n}-2)}{(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})}f\right\}, \end{aligned}$$

where

$$f = \frac{2\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{b}}{s_{1n}^2} - \frac{4(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}\boldsymbol{\Sigma}_n^{-1}\mathbf{a}\mathbf{b}'(\mathbf{z}+\boldsymbol{\zeta})}{s_{1n}^2(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})} - \frac{(p_{2n}-2)(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{b}\mathbf{b}'(\mathbf{z}+\boldsymbol{\zeta})}{s_{1n}^2(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})}.$$

Notice that $\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{b} = \mathbf{d}_{2n}'\boldsymbol{\Sigma}_{n22.1}^{-1}\mathbf{d}_{2n} = s_{2n}^2$ and $\mathbf{W}\boldsymbol{\Sigma}_n^{-1}\mathbf{ab}' = \mathbf{bb}'$. Therefore,

$$f = 2\frac{s_{2n}^2}{s_{1n}^2} - \frac{(p_{2n}+2)(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{bb}'(\mathbf{z}+\boldsymbol{\zeta})}{s_{1n}^2(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})}.$$

Thus,

$$\begin{aligned} J_2 - J_3 &= \lim_{n\to\infty} E\left\{\frac{(p_{2n}-2)}{(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})}\left[\frac{2s_{2n}^2}{s_{1n}^2} - \frac{(p_{2n}+2)(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{d}_{2n}\mathbf{d}_{2n}'(\mathbf{z}+\boldsymbol{\zeta})}{s_{1n}^2(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})}\right]\right\} \\ &= \lim_{n\to\infty}\frac{s_{2n}^2}{s_{1n}^2}E\left\{\frac{(p_{2n}-2)}{(\mathbf{z}_2+\boldsymbol{\delta})'\boldsymbol{\Sigma}_{n22.1}(\mathbf{z}_2+\boldsymbol{\delta})}\left[2 - \frac{(p_{2n}+2)(\mathbf{z}_2+\boldsymbol{\delta})'\mathbf{d}_{2n}\mathbf{d}_{2n}'(\mathbf{z}_2+\boldsymbol{\delta})}{s_{2n}^2(\mathbf{z}_2+\boldsymbol{\delta})'\boldsymbol{\Sigma}_{n22.1}(\mathbf{z}_2+\boldsymbol{\delta})}\right]\right\}, \end{aligned}$$

where $\mathbf{z}_2$ satisfies that $s_{2n}^{-1}\mathbf{d}_{2n}'\mathbf{z}_2 \to N(0,1)$. Thus (4.9c) holds. Similarly, we can obtain (4.9d). □

*Proof of Corollary 3*

We first verify (i). Define $\tilde{\mathbf{z}}_2 = \sigma^{-2}\boldsymbol{\Sigma}_{n22.1}^{1/2}(\mathbf{z}_2+\boldsymbol{\delta})$ and $\mathbf{B} = (\sigma^2/s_{2n}^2)\boldsymbol{\Sigma}_{n22.1}^{-1/2}\mathbf{d}_{2n}\mathbf{d}_{2n}'\boldsymbol{\Sigma}_{n22.1}^{-1/2}$. From the Cramér–Wold device, we have

$$\begin{aligned} J_2 - J_3 &= (1-c)\lim_{n\to\infty}\left\{E\left[\frac{2(p_{2n}-2)}{\tilde{\mathbf{z}}_2'\tilde{\mathbf{z}}_2}\right] - E\left[\frac{(p_{2n}-2)(p_{2n}+2)\tilde{\mathbf{z}}_2'\mathbf{B}\tilde{\mathbf{z}}_2}{(\tilde{\mathbf{z}}_2'\tilde{\mathbf{z}}_2)^2}\right]\right\} \\ &= (1-c)\lim_{n\to\infty}\left\{E\left[\frac{p_{2n}-2}{\chi_{p_{2n}}^2(\Delta_n)}\right] - \mathrm{Tr}(\mathbf{B})E\left[\frac{(p_{2n}-2)(p_{2n}+2)}{\chi_{p_{2n}+2}^4(\Delta_n)}\right]\right\} \\ &\quad + (1-c)\lim_{n\to\infty}\left\{E\left[\frac{p_{2n}-2}{\chi_{p_{2n}}^2(\Delta_n)}\right] - (\boldsymbol{\delta}'\mathbf{B}\boldsymbol{\delta})E\left[\frac{(p_{2n}-2)(p_{2n}+2)}{\chi_{p_{2n}+4}^2(\Delta_n)}\right]\right\}, \end{aligned}$$

where $\Delta_n = \boldsymbol{\delta}'\boldsymbol{\Sigma}_{n22.1}\boldsymbol{\delta}$ and 'Tr($\mathbf{B}$)' is the trace of matrix $\mathbf{B}$. Here, the second '=' is from Theorem 8 in Chapter 2 in [29]. Notice that $\mathrm{Tr}(\mathbf{B}) = 1$. Using the relationship between the chi-square distribution and Poisson distribution,

$$\begin{aligned} J_2 - J_3 &= (1-c)\lim_{n\to\infty}\left\{E_\kappa\left[\frac{p_{2n}-2}{p_{2n}-2+2\kappa}\left(1 - \frac{p_{2n}+2}{p_{2n}+2\kappa}\right)\right]\right\} \\ &\quad + (1-c)\lim_{n\to\infty}\left\{E_\kappa\left[\frac{p_{2n}-2}{p_{2n}-2+2\kappa}\left(1 - \frac{\boldsymbol{\delta}'\mathbf{B}\boldsymbol{\delta}(p_{2n}-2+2\kappa)(p_{2n}+2)}{(p_{2n}+2+2\kappa)(p_{2n}+2\kappa)}\right)\right]\right\}, \end{aligned}$$

where $\kappa$ is a Poisson distribution with mean $\Delta_n/2$ and $E_\kappa$ means the expectation is taken for the Poisson random variable $\kappa$. Because $P(\kappa \geqslant 1) \to 1$ when $p_{2n} \to \infty$. With almost probability 1, we have

$$0 \leqslant \frac{p_{2n}-2}{p_{2n}-2+2\kappa}\left(1 - \frac{p_{2n}+2}{p_{2n}+2\kappa}\right) \leqslant 1.$$

If $\|\boldsymbol{\delta}\|^2 \leqslant 1$, then $\boldsymbol{\delta}'\mathbf{B}\boldsymbol{\delta} = \left(\boldsymbol{\delta}'\boldsymbol{\Sigma}_{n22.1}^{-1/2}\mathbf{d}_{2n}\right)^2 / \left(\mathbf{d}_{2n}'\boldsymbol{\Sigma}_{n22.1}^{-1}\mathbf{d}_{2n}\right) \leqslant \boldsymbol{\delta}'\boldsymbol{\delta} \leqslant 1$. Then, $E[g_1(\mathbf{z}_2+\boldsymbol{\delta})] \geqslant 0$. Furthermore, when $\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x} \leqslant p_{2n} - 2$, we have

$$2 - s_{2n}^{-2}\mathbf{x}\mathbf{d}_{2n}\mathbf{d}_{2n}'\mathbf{x}' \geqslant 2 - \frac{(p_{2n}-2)\mathbf{x}\mathbf{d}_{2n}\mathbf{d}_{2n}'\mathbf{x}'}{s_{2n}^2\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}} > 2 - \frac{(p_{2n}+2)\mathbf{x}\mathbf{d}_{2n}\mathbf{d}_{2n}'\mathbf{x}'}{s_{2n}^2\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}}.$$

Therefore, $g_2(\mathbf{x}) \geqslant g_1(\mathbf{x})$. Thus, (i) holds.

Copyright © 2016 John Wiley & Sons, Ltd.

In fact, the inequalities in (i) also hold even though $\|\boldsymbol{\delta}\|^2 > 1$. For example, suppose $\Delta_n = \iota p_{2n}$ for some constant $\iota > 0$. Then, $p_{2n}^{-1/2}(2\kappa - \Delta_n) \rightsquigarrow N(0, \iota^{-1})$. Therefore, if $\|\boldsymbol{\delta}\|^2 \leqslant 1 + \iota$, with probability 1, we have

$$1 - \frac{\boldsymbol{\delta}'\mathbf{B}\boldsymbol{\delta}(p_{2n} - 2 + 2\kappa)(p_{2n} + 2)}{(p_{2n} + 2 + 2\kappa)(p_{2n} + 2\kappa)} \to 1 - \frac{\|\boldsymbol{\delta}\|^2}{1 + \iota} > 0.$$

Thus, (ii) holds.

We now verify (iii). If $\boldsymbol{\delta} = \mathbf{0}$, then $\Delta_{\mathbf{d}_{1n}} = 0$. Thus, $\mathrm{ADR}(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}) = c < \mathrm{ADR}(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}})$. We now compare $\mathrm{ADR}(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{SE}})$ with $\mathrm{ADR}(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}})$. Denote $\mathbf{A} = (p_{2n} + 2)s_{2n}^2\boldsymbol{\Sigma}_{n22.1}^{-1/2}\mathbf{d}_{2n}\mathbf{d}'_{2n}\boldsymbol{\Sigma}_{n22.1}^{-1/2}$. If $\boldsymbol{\delta} = \mathbf{0}$, then we have

$$(1 - c)^{-1}g_1(\mathbf{z}_2) = \lim_{n\to\infty}\frac{(p_{2n} - 2)}{\mathbf{z}_2'\boldsymbol{\Sigma}_{n22.1}\mathbf{z}_2} - \lim_{n\to\infty}\frac{(p_{2n} - 2)\left(\boldsymbol{\Sigma}_{n22.1}^{1/2}\mathbf{z}_2\right)'\mathbf{A}\left(\boldsymbol{\Sigma}_{n22.1}^{1/2}\mathbf{z}_2\right)}{\left(\mathbf{z}_2'\boldsymbol{\Sigma}_{n22.1}\mathbf{z}_2\right)^2}.$$

From Theorem 2.1.8 in [29] and moment of inverse chi-squares distribution, we have

$$\lim_{n\to\infty} E\left[\frac{(p_{2n} - 2)}{\mathbf{z}_2'\boldsymbol{\Sigma}_{n22.1}\mathbf{z}_2}\right] = 2$$

and

$$\lim_{n\to\infty} E\left[\frac{(p_{2n} - 2)\left(\boldsymbol{\Sigma}_{n22.1}^{1/2}\mathbf{z}_2\right)'\mathbf{A}\left(\boldsymbol{\Sigma}_{n22.1}^{1/2}\mathbf{z}_2\right)}{\left(\mathbf{z}_2'\boldsymbol{\Sigma}_{n22.1}\mathbf{z}_2\right)^2}\right] = \lim_{n\to\infty} 1 + 2/p_{2n}.$$

Thus, if $p_{2n} = p_2$ is fixed, $E[g_1(\mathbf{z}_2)] = (1 - c)(1 - 2/p_2) < 1 - c$. Therefore,

$$\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{SE}}\right) > \mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right).$$

Similarly, we can verify $\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right) < \mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{PSE}}\right)$. When $p_{2n} \to \infty$, $\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{PSE}}\right) \to 1$. □

## References

1. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 1996; **58**:267–288.
2. Leng C, Lin Y, Wahba G. A note on the Lasso and related procedures in model selection. *Statistica Sinica* 2006; **16**:1273–1284.
3. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 2001; **96**:1348–1360.
4. Fan J, Lv J. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory* 2011; **57**(8):5467–5484.
5. Zou H. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* 2006; **101**:1418–1429.
6. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 2010; **38**(2):894–942.
7. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 2010; **20**(1):101.
8. Zhang CH, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B* 2014; **76**(1):217–242.
9. Zhao P, Yu B. On model selection consistency of LASSO. *Journal of Machine Learning Research* 2006; **7**:2541–2563.
10. Huang J, Ma SG, Zhang CH. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* 2008; **18**:1603–1618.
11. Bickel P, Ritov Y, Tsybakov A. Simultaneous analysis of Lasso and dantzig selector. *Annals of Statistics* 2009; **37**:1705–1732.
12. Hansen BE. *The risk of James-Stein and Lasso shrinkage*, 2013. http://www.ssc.wisc.edu/bhansen/papers/lasso.pdf [accessed 20 December 2015].
13. Belloni A, Chernozhukov V. Least squares after model selection in high-dimensional sparse models. *Bernoulli* 2009; **19**:521–547.
14. Liu H, Yu B. Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics* 2013; **7**:3124–3169.
15. Stein C. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955*, vol. I. University of California Press: Berkeley and Los Angeles, 1956; 187–195.
16. James W, Stein C. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1961; **1**:361–379.
17. Ahmed SE. *Penalty, Shrinkage and Pretest Strategies: Variable Selection and Estimation*. Springer: New York, 2014.

Copyright © 2016 John Wiley & Sons, Ltd.

18. Ahmed SE, Fallahpour S. Shrinkage estimation strategy in quasi-likelihood models. *Statistics & Probability Letters* 2012; **82**:2170?2179.
19. Ahmed SE, Hossain S, Doksum KA. Lasso and shrinkage estimation in Weibull censored regression models. *Journal of Statistical Inference and Planning* 2012; **12**:1273–1284.
20. Ahmed SE, Doksum KA, Hossain S, You J. Shrinkage, pretest and absolute penalty estimators in partially linear models. *Australian & New Zealand Journal of Statistics* 2007; **49**:435–454.
21. Marsaglia G, Styan GPH. Equalities and inequalities for ranks of matrices. *Linear and Multilinear Algebra* 1974; **2**:269–292.
22. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools (with discussion). *Technometrics* 1993; **35**:109–148.
23. Zhang CH, Huang J. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics* 2008; **36**: 1567–1594.
24. Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* 1998; **20**(1):3361.
25. Wainwright MJ. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* 2009; **55**(5):2183–2202.
26. Zheng Z, Fan Y, Lv J. High dimensional threshold regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B* 2014; **76**:627–649.
27. Weng H, Feng Y, Qiao X. *Regularization after retention in ultrahigh dimensional linear regression models*, 2013. arXiv preprint arXiv:1311.5625.
28. Shao J, Deng X. Estimation in high-dimensional linear models with deterministic design matrices. *Annals of Statistics* 2012; **40**:812–831.
29. Saleh A. KME. *Theory of Preliminary Test and Stein-Type Estimation with Applications*. Wiley: New York, 2006.
30. Barro R, Lee J. *Data set for a panel of 139 countries*, 1994. http://admin.nber.org/pub/barro.lee/ [accessed 20 December 2015].
31. Barro R, Sala-i Martin X. *Economic Growth*. McGraw-Hill: New York, 1995.
32. Durlauf S, Johnson PA, Temple JRW. Growth econometrics. *Handbook of Economic Growth* 2005; **1**:555–677.
33. Lee S, Seo MH, Y S. The LASSO for high-dimensional regression with a possible change-poin. *Journal of the Royal Statistical Society: Series B* 2016; **78**:193–210.

Copyright © 2016 John Wiley & Sons, Ltd.

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2224 Published online in Wiley Online Library

# Discussion of ‘Post selection shrinkage estimation for high-dimensional data analysis’

I wholeheartedly congratulate Gao, Ahmed, and Feng on the stimulating paper, which provides a novel idea to estimate weak but non-vanishing coefficients that can be missed in the first stage of model selection. The basic idea is to conduct the shrinkage estimator on the variables that are not selected in the first stage, penalizing the $L_2$-norm of the coefficients of remaining variables without penalty on those that survived the first stage, and then to recruit additionally those variables with big estimated ridge regression coefficients. This basically assesses the conditional contributions of the remaining variables, given those already selected in the first stage, using a ridge regression technique to avoid the curse of dimensionality in implementation. Furthermore, such a post selection shrinkage estimator is further combined with the post-lasso estimator to improve further the estimated coefficients for the strong signal components. The advantages of including these additional variables are shown unambiguously by using both asymptotic and empirical studies. I welcome the opportunity to make a few comments.

There are a couple of assumptions that is worth discussing. First of all, the authors assume that there are three well-separated regimes of regression coefficients, whose indices are denoted by $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}_3$ by the authors. This can be too ideal in various applications. In addition, while the conditional model selection consistency for $\mathbf{S}_2$ is given in Theorem 1, I can not find the condition on the minimal signal strength in $\mathbf{S}_2$ to separate $\mathbf{S}_2$ from $\mathbf{S}_3$. Assumption **(A2)** is probably inadequate.

Secondly, it is assumed that $\mathbf{S}_1$ is consistently estimated in the preliminary stage. While technical conditions can be posed to make model selection consistency, in many applications, this is very hard to achieve. What happens to the procedure if there are some missed variables in the first stage? Can it be recovered in the second stage? How about irrelevant variables recruited in the first stage?

Finally, the assumption on the design matrix is rather strong. It excludes reasonably correlated design matrix. Are assumptions **(B2)** and **(B4)** typically compatible? A showcase example is welcome.

The capstone of the paper is to define the weighted ridge estimator and post-selection shrinkage estimator. High-dimensional ridge estimation can be tricky: from numerical stability to biases and variances in each component. One tuning ridge parameter is not flexible enough. For weighted ridge estimator, instead of using shrinkage and thresholding, one can also use penalized least-squares:

$$\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \sum_{j\in\hat{S}_1} |\beta_j| + \lambda_2 \sum_{j\notin\hat{S}_1} |\beta_j|$$

where $\lambda_1$ and $\lambda_2$ are penalization parameters. Following the idea of the authors, one can take $\lambda_1 = 0$. This corresponds to $p'_\lambda(\hat{\beta}_j)$ for a sufficiently large $|\hat{\beta}_j|$ such as those $j \in \hat{S}_1$ for SCAD and MCP penalties. In that sense, the given penalized least-squares corresponds to the one-step implementation of the folded-concave penalized least-squares, namely, adaptive Lasso (Zou, 2006; Zou and Li, [1]; Fan, Xue, and Zou, [2]). Fan, Liu, Sun, and Zhang unveil the iteration effects of iteratively reweighted lasso implementation of the folded-concave penalized least-squares and demonstrate the advantages of such iterations. Can we have similar iterative versions of weighted ridge estimators that allow the sets of $\hat{S}_1$, $\hat{S}_2$, and $\hat{S}_3$ to change?

Two versions of post-selection shrinkages are given in (3.6) and (3.8). The authors are welcome to provide intuitions on these estimators and why they expect to improve the sampling properties. The simulation results on the selection consistency of $\hat{S}_1$ and $\hat{S}_2$ are also welcome. In addition, comparisons with other existing procedures will provide additional insights.

Jianqing Fan
*Department of Operations Research and Financial Engineering*
*Princeton University*
E-mail: jqfan@princeton.edu

Copyright © 2017 John Wiley & Sons, Ltd.

## References

1. Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics* 2008; **36**: 1509–1533.
2. Fan J, Xue L, Zou H. Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics* 2014; **42**:819–849.

Copyright © 2017 John Wiley & Sons, Ltd.

Discussion

Applied Stochastic
Models in Business
and Industry

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2220

Published online in Wiley Online Library

# Discussion of ‘Post selection shrinkage estimation for high-dimensional data analysis’

The authors are congratulated on this interesting paper about high dimensional (HD) data analysis. Because of rapid progress in data acquisition techniques, more and more applications have HD data involved. Thus, statistical modeling and analysis of HD data have become a popular research area in the past 10–15 years. Many related methodologies have been developed in the literature, most of which use the terminologies such as ‘variable selection’, ‘dimension reduction’, and ‘machine learning’. As pointed out in the paper, many existing methods discuss the HD problem under the sparsity assumption, and try to select a small number of important covariates to be kept in a model and remove all other covariates from regression modeling. In reality, however, it can happen that there are many covariates that all provide useful information about the response variable $y$, although the amount of such useful information in a single covariate might be small. One novelty of the proposed methods in the current paper is that the authors try to properly accommodate such relatively small contributions from these covariates by (i) selecting the important covariates using the conventional LASSO or adaptive LASSO algorithm and (ii) suggesting a post-selection shrinkage estimation strategy to properly accommodate the contribution of some less important covariates. Both theoretical arguments and numerical examples show that the proposed methods have some advantages, compared with certain existing variable selection methods (e.g., LASSO), for HD model estimation. Next, we comment on certain aspects of the proposed methods and provide some suggestions for future research on the related topics.

## 1. Model and model assumptions

The paper focuses on the linear regression model (1.1), as did in most papers on HD data modeling. This model has many model assumptions, including the linearity, i.i.d. and additive random noise with mean 0 and constant variance $\sigma^2$. Although it is not mentioned immediately after (1.1), the paper also assumes that the noise is normally distributed (cf., the assumption (B1) in Section 4). The authors pointed out that many existing methods on variable selection require the sparsity assumption that only a small number of model coefficients in model (1.1) are non-zero and that this assumption may not be valid in many applications. In practice, a more realistic scenario is that there could be a quite number of covariates that provide useful information for describing the response variable $y$, although their contributions might be relatively small, compared with a small number of important covariates. So, the paper focuses on this scenario and suggests some new methodologies to handle it properly.

We agree with the authors completely that the sparsity assumption may not be valid in most applications. Instead, the scenario with a small number of important covariates and a relatively large number of helpful but less important covariates might be more realistic. To describe this scenario, the authors introduce three *signal strength assumptions* (A1)–(A3) to define three sets of covariates according to their signal strength levels: $S_1$ includes covariates with strong signal strength, $S_2$ includes covariates with weak signal strength, and $S_3$ includes covariates with zero signal strength. The three levels of signal strength are defined based on the magnitudes of the true regression coefficients. For instance, the assumption (A2) specifies the covariates in $S_2$ to be those whose regression coefficients, denoted as $\boldsymbol{\beta}^*_{S_2}$, satisfy $\|\boldsymbol{\beta}^*_{S_2}\| = O(n^\tau)$, where $\tau \in (0, 1)$ is a constant and $\|\cdot\|$ is the $L_2$ norm. We would like to point out that the definitions of $S_1 - S_3$ in (A1)–(A3) may not be rigorous enough. For instance, by the current definition, $S_2$ should also contain all covariates in $S_3$ because a sequence $a_n = O(n^\tau)$ can include the case when $a_n = 0$, for all $n$, by the definition of the big O notation. So, we would suggest that you change ‘$\|\boldsymbol{\beta}^*_{S_2}\| = O(n^\tau)$’ to ‘$\|\boldsymbol{\beta}^*_{S_2}\| \sim O(n^\tau)$’ in (A2) and specify that all components of $\boldsymbol{\beta}^*_{S_2}$ are non-zero. Similarly, by the current definition, $S_1$ and $S_2$ may not be disjoint. For instance, if $p_n = \exp(n^{2(\tau-0.5)+1})$ and $\tau > 0.5$, then some covariates in $S_2$ can also belong to $S_1$. So, it requires much effort on the definitions of $S_1$-$S_3$ so that they are three disjoint sets of covariates and really represent the covariates at the high, low, and zero signal strength levels.

As aforementioned, one important contribution of the current paper is to generalize the sparsity assumption that divides all covariates into two sets (i.e., useful and non-useful covariates) to cases with three sets (i.e., useful, less useful, and non-useful covariates). It tries to accommodate certain covariates with relatively weak signal strength in the modeling.

Copyright © 2017 John Wiley & Sons, Ltd.

We agree with the authors that this is an important step forward in the variable selection research. However, in practice, it is always challenging to divide all covariates into two or three categories, because the signal strength might be a continuous quantity and it is quite subjective to divide its values into two or three categories. For instance, in Case I of your simulation example, the components of $\boldsymbol{\beta}^*$ are chosen to be 5, 0.5, or 0. If the components of $\boldsymbol{\beta}^*$ can take the values of 5, 4, 3, 2, 1, 0.5, 0.2, and 0, how can we divide them into three categories? Should we consider the three groups $\{5, 4, 3, 2\}$, $\{1, 0.5, 0.2\}$ and $\{0\}$? or, an alternative grouping $\{5, 4, 3, 2, 1\}$, $\{0.5, 0.2\}$ and $\{0\}$? Of course, we can also consider four or more groups. It may require some future research effort to address this kind of arbitrariness involved in the grouping of the covariates.

The three-step post selection shrinkage estimation strategy discussed in Section 3 is a creative one. From the definitions of $\widehat{\boldsymbol{\beta}}^{WR}(r_n, a_n)$ in (3.2) and $\widehat{\boldsymbol{\beta}}^{PSE}_{\widehat{S}_1}$ in (3.8), selection of the parameters $r_n$ and $a_n$ is critically important to their performance. In the simulation study, the authors suggest choosing $a_n = c_1 n^{-1/8}$ and $r_n = c_2 a_n^{-2}(\log\log n)^3 \log(n \bigvee p_n)$, where the constants $c_1$ and $c_2$ are determined by cross-validation. However, it is still unknown whether this parameter selection scheme will work well in general cases. Much research is needed to provide practical guidelines for choosing these parameters in different scenarios.

As mentioned in the first paragraph of this part, the sparsity assumption is only one of many assumptions of model (1.1). In cases when there are a large number of covariates involved, it is difficult to imagine that the regression function is still linear. Recently, there is some research on nonparametric transformation of covariates in the context of dimension reduction (e.g., Mai and Zou [1]). Also, in image or other spatial data, the random noise could be spatially correlated. In MRI or fMRI image data, the random noise may not be additive and the noise variability could change over location (e.g., Mukherjee and Qiu [2]).

## 2. Evaluation of different methods

In the simulation study in Section 5, the authors use the relative mean squared error (RMSE) criterion defined in (5.1) for comparing the three different methods RE, ALASSO, and PSE. While this criterion is good for evaluating the overall performance of the coefficient estimators, it has its limitations. For instance, in Case 1 of your simulation example, 3 coefficients have their true values of 5, 10 coefficients have their true values of 0.5, and the remaining coefficients are all 0. The coefficient values are dramatically different in such a case. So, the criterion RMSE is mainly for evaluating the performance of the estimates of the first three coefficients in $\boldsymbol{\beta}^*$. An alternative criterion is the average or sum of $(\widehat{\beta}_j^* - \beta_j^*)/\beta_j^*$, over all $j$, where $\beta_j^*$ is the $j$th component of $\boldsymbol{\beta}^*$. This alternative criterion will not be dominated by certain coefficients whose values are much larger than the other coefficients. Also, the scale of a covariate can be changed in practice. For instance, in your real-data example discussed in Section 6, the covariate gdp60 can be in the unit of dollars, or in the unit of 1000 dollars. If the unit of a covariate changes, then its coefficient value will also change. Consequently, some less important covariates become important ones in your definitions of $S_1$-$S_3$, and vice versa. Your suggested methods and the criterion RMSE depend on the specific unit of each covariate, while the suggested alternative criterion does not. Another alternative criterion is the mean square error of the entire regression function, defined as

$$E(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^* - \boldsymbol{X}\boldsymbol{\beta}^*)^2.$$

This criterion does not depend on the covariate scale either.

## 3. Model diagnoses and applications

One major contribution of the paper is to make certain variable selection methods (e.g., LASSO) more practical by loosening the sparsity assumption and accommodating certain covariates whose contribution in describing the response variable $y$ is less important than the major covariates that are likely to be selected by the conventional variable selection methods. This is definitely a welcome research effort. However, to make a method relevant to applications and compare different methods about their adequacy and goodness-of-fit in a specific application, some proper diagnosis tools and goodness-of-fit tests are necessary, which could be good topics for future research. For instance, in the real-data example discussed in Section 6, why is the model (6.1) adequate for describing the GDP growth data? Are the random errors $\{\varepsilon_i\}$ i.i.d. and normally distributed? If some of these assumptions are violated, will the related variable selection methods still perform well? For a specific variable selection method, after the model (6.1) is estimated, how do the residual plots look like? Can we perform a formal goodness-of-fit test about the estimated model? And so on and so forth. Thus, a great future research effort is still needed to answer all these questions. Definitely, the research effort in the current paper is a first step towards that direction.

We will close by thanking the authors for a thought-provoking paper and a novel variable selection method that has its potential to be used in a wide range of applications.

Copyright © 2017 John Wiley & Sons, Ltd.

PEIHUA QIU, KAI YANG AND LU YOU
*Department of Biostatistics*
*University of Florida*
*Gainesville, FL 32610, USA*
E-mail: *pqiu@ufl.edu*

## Acknowledgement

This work was partially supported by the National Science Foundation under the grant DMS-1405698.

## References

1. Mai Q, Zou H. Nonparametric variable transformation in sufficient dimension reduction. *Technometrics* 2015; **57**:1–10.
2. Mukherjee PS, Qiu P. Efficient bias correction For MRI image denoising. *Statistics in Medicine* 2013; **32**:2079–2096.

Copyright © 2017 John Wiley & Sons, Ltd.

Applied Stochastic
Models in Business
and Industry

Discussion

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2216 Published online in Wiley Online Library

# Discussion of ‘Post selection shrinkage estimation for high-dimensional data analysis’

We congratulate Guo, Ahmed, and Feng (referred to as GAF hereafter) on an interesting paper that advances theory and methodologies relevant to post selection estimators in high-dimensional data settings. As existing post estimators have often ignored contributions from weak signals, the key contribution of this paper is proposing a new post selection shrinkage estimator (PSE) that takes into account the joint impact of both strong and weak signals. Through intensive theoretical and empirical work, GAF have demonstrated that the PSE possesses improved prediction performance compared with the post selection estimators generated by Lasso-type methods. In this discussion, we re-consider the PSE estimator from two new perspectives.

First, we notice that GAF have only focused on detecting marginally strong and weak signals. However, variables that are regarded as ‘noise variables’ (or in $S_3$) but have non-ignorable impact on the outcome, together with some variables in $S_1$ or $S_2$, are also worth considering. These variables, termed marginally unimportant but jointly informative variables, have aroused much interest recently. We plan to explore the performance of PSE in the presence of marginally unimportant but jointly informative variables. Secondly, we are keen on investigating whether the PSE approach can be extended to encompass ultrahigh-dimensional data because the pre-determined important set $\hat{S}_1$, as defined by GAF, is obtained from the regularized regression method that is not feasible for ultrahigh-dimensional data analysis.

## 1. Existence of marginally unimportant but jointly informative variables

The performance of post selection estimators largely depends on how the submodel $S_1$ is selected. It is well known that Lasso-type penalized regularization approaches tend to select only one representative variable out of several highly correlated variables, and also tend to miss marginally weak signals. As marginally unimportant but jointly informative (MUJI) variables are highly correlated with some variables in $S_1$, they have low priorities to be selected using the regularization method, which will incur inefficient estimation and large prediction errors. Although the proposed post selection shrinkage estimator (PSE) takes into account covariates with marginally weak impact on the response, it fails to account for the effects of MUJI variables, which typically belong to $S_3$. The existence of MUJI variables can be easily identified by investigating the covariance structure. This naturally leads to a question on how to incorporate such a covariance structure into the construction of post selection estimators for identifying MUJI variables, denoted by $S_{\text{MUJI}}$, and for simultaneously estimating $\boldsymbol{\beta}$ based on the three sets, $S_1$, $S_2$, and $S_{\text{MUJI}}$.

## 2. Applicability to the ultrahigh-dimensional data

In an ultrahigh-dimensional data setting, where the number of covariates $p_n$ is in the exponential order of sample size $n$, solving a penalized regression problem is computationally infeasible as it involves inverting a $p_n \times p_n$ matrix. Moreover, the finite sample oracle bounds for selection and estimation errors are in the scale of $O(\log p_n/n)$, which are too wide for ultrahigh-dimensional settings. Therefore, the current PSE method may not be directly applicable to model the ultrahigh-dimensional data.

To address the challenge, we modify the PSE algorithm proposed by Guo, Ahmed, and Feng (GAF) and present a covariance insured screening-based PSE (CIS-PSE), which incorporates the correlation structure to identify $S_{\text{MUJI}}$ and facilitates variable selection in ultrahigh-dimensional settings.

## 3. Covariance insured screening-based post selection shrinkage estimator

Following GAF, we use the same definitions of $S_1$, $S_2$, $S_3$, representing strong, weak, and sparse signal set, respectively. Assuming that $\mathbf{X}$ has been standardized columnwise, we design the proposed CIS-PSE algorithm as follows.

Copyright © 2017 John Wiley & Sons, Ltd.

1. Select $\hat{S}_1$, $\hat{S}_2$, and $\hat{S}_{\text{MUJI}}$:

   Obtain the marginally strong set $\hat{S}_1$ using the selection criteria of $\hat{S}_1 = \{j : |\mathbf{X}_j'\mathbf{y}/(\mathbf{X}_j'\mathbf{X}_j)| > \tau_n\}$ for some tuning parameter $\tau_n$. Set $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{MS}} = (\mathbf{X}_{\hat{S}_1}'\mathbf{X}_{\hat{S}_1})^{-1}\mathbf{X}_{\hat{S}_1}'\mathbf{y}$. If the number of variables in $\hat{S}_1$ exceeds the sample size, a Lasso regression can be used instead. Here, $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{MS}}$ plays the same role as $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{RE}}$ in GAF except that $\hat{S}_1$ is obtained by a marginal screening, and thus is adaptive to the ultrahigh-dimensional data.

   Then, compute residuals from the fitted model based on $\hat{S}_1$, that is, $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}_{\hat{S}_1}\hat{\boldsymbol{\beta}}_{\hat{S}_1}$. Treating $\hat{\boldsymbol{\epsilon}}$ as the working response variable, we recruit new predictors by $\hat{S}_2 = \{j \in \hat{S}_1^c : |\mathbf{X}_j'\hat{\boldsymbol{\epsilon}}/(\mathbf{X}_j'\mathbf{X}_j)| > \nu_n\}$, where $\nu_n$ is a tuning parameter.

   The set of MUJI variables is selected by $\hat{S}_{\text{MUJI}} = \{j \in \hat{S}_1^c : |\mathbf{X}_j'\mathbf{X}_{j'}| > \rho_n$ for some $j' \in \hat{S}_1\}$, where $\rho_n$ is a tuning parameter.

2. Obtain an initial post selection least squares estimator with variables belonging to $\hat{S}_1 \cup \hat{S}_2 \cup \hat{S}_{\text{MUJI}}$. If the number of variables in $\hat{S}_1 \cup \hat{S}_2 \cup \hat{S}_{\text{MUJI}}$ exceeds the sample size, we use a ridge regression with a penalty only on coefficients in $\hat{S}_1^c \cap (\hat{S}_2 \cup \hat{S}_{\text{MUJI}})$. Denote the resulting estimates by $\hat{\boldsymbol{\beta}}^{\text{R}}$. Similar to GAF, we hard-threshold the parameters in $\hat{S}_1^c$ to obtain the post screening weighted ridge (SWR) estimator $\hat{\boldsymbol{\beta}}^{\text{SWR}}$ from

$$\hat{\beta}_j^{\text{SWR}} = \begin{cases} \hat{\beta}_j^{\text{R}}, & j \in \hat{S}_1 \\ \hat{\beta}_j^{\text{R}} I(\hat{\beta}_j^{\text{R}} > a_n), & j \in \hat{S}_1^c \cap (\hat{S}_2 \cup \hat{S}_{\text{MUJI}}) \\ 0, & \text{otherwise.} \end{cases}$$

   Denote by $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{SWR}}$ the components of $\hat{\boldsymbol{\beta}}^{\text{SWR}}$ corresponding to $\hat{S}_1$. Though $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{SWR}}$ is defined similarly as in GAF, it incorporates both $\hat{S}_2$ and $\hat{S}_{\text{MUJI}}$.

3. We obtain the CIS-PSE of $\boldsymbol{\beta}_1$ by

$$\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{CIS-PSE}} = \hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{SWR}} - \left(\frac{\hat{s}_2 - 2}{\hat{T}_n} \wedge 1\right)\left(\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{SWR}} - \hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\text{MS}}\right),$$

   where $\hat{s}_2 = |\hat{S}_2 \cup \hat{S}_{\text{MUJI}}|$ and $\hat{T}_n$ is as defined in GAF.

In summary, the proposed CIS-PSE estimator is different from the PSE in two aspects. First, it incorporates $S_{\text{MUJI}}$ that could be missed by the PSE because of high correlations with variables in $S_1$. Second, aided by a screening procedure, the CIS-PSE can accommodate ultrahigh-dimensional data.

## 4. Numerical examples

To evaluate the performance of our proposal, we consider two examples where non-ignorable signals come from either $S_2$ or $S_{\text{MUJI}}$.

*Example* 1

Assume that $\epsilon_i$ are i.i.d. from $N(0,1)$. $\mathbf{X}_{i,S_1 \cup S_{2;1:3}} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a $6 \times 6$ covariance matrix with unit marginal variances, $\text{cor}(X_1, X_4) = \text{cor}(X_2, X_5) = \text{cor}(X_3, X_6) = 0.8$ and all other covariances being zeros. For $s \notin \{1, \ldots, 6\}$, $x_{is}$ are simulated independently from $N(0, \sigma^2)$, where $\sigma$ is chosen such that the signal to noise ratios for the weak signals in $S_2$ are about 1. We set $n = 200$ and $p_n = 400$, 10,000, and 100,000. The absolute values of the true regression coefficients are set to be

$$|\boldsymbol{\beta}^*| = (\overbrace{10, 10, 10}^{S_1}, \overbrace{\underbrace{0.5, 0.5, 0.5}_{S_{2,1:3}}, \underbrace{0.5, \cdots, 0.5}_{10}}^{S_2}, \overbrace{0, \cdots, 0}^{S_3})'$$

with all nonzero coefficients randomly assigned to be either positive or negative.

Copyright © 2017 John Wiley & Sons, Ltd.

**Table I.** Numerical results.

| Example | | | $p_n = 400$ | $p_n = 10,000$ | $p_n = 100,000$ |
|---|---|---|---|---|---|
| Example 1 | PSE | MSE | 0.46 | NA | NA |
| | | RMSE | 1.02 | NA | NA |
| | | $\lvert\hat{S}_1\rvert$ | 3.0 | NA | NA |
| | | $\lvert\hat{S}_2\rvert$ | 8.6 | NA | NA |
| | CIS-PSE | MSE | 0.08 | 1.62 | 1.47 |
| | | RMSE | 22.75 | 10.87 | 8.76 |
| | | $\lvert\hat{S}_1\rvert$ | 3.0 | 2.9 | 3.0 |
| | | $\lvert\hat{S}_2\rvert$ | 11.1 | 10.1 | 10.6 |
| Example 2 | PSE | MSE | 0.05 | NA | NA |
| | | RMSE | 1.02 | NA | NA |
| | | $\lvert\hat{S}_1\rvert$ | 3.0 | NA | NA |
| | | $\lvert\hat{S}_2\rvert$ | 7.6 | NA | NA |
| | CIS-PSE | MSE | 0.09 | 0.56 | 0.42 |
| | | RMSE | 5.01 | 1.27 | 0.99 |
| | | $\lvert\hat{S}_1\rvert$ | 3.0 | 3.0 | 3.0 |
| | | $\lvert\hat{S}_2\rvert$ | 8.2 | 6.2 | 9.1 |

CIS-PSE, covariance insured screening-based post selection shrinkage estimator; MSE, mean squared error; NA, not applicable ; RMSE, relative mean squared error.

*Example* 2

Consider the same setting as Example 1 except that $\mathbf{X}_{i,S_1\cup S_{\text{MUJI}}} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and

$$|\boldsymbol{\beta}^*| = (\overbrace{10, 10, 10}^{S_1}, \overbrace{\underbrace{0.5, \cdots, 0.5}_{10}}^{S_2}, \overbrace{\underbrace{0, 0, 0}_{S_{\text{MUJI}}}, 0, \cdots, 0}^{S_3})'.$$

We obtained the estimation of $\boldsymbol{\beta}_{S_1}$ via PSE and CIS-PSE and compared their performance. We applied cross-validation for tuning parameters $\tau_n$, $\nu_n$, $\rho_n$ and $\alpha_n$. To evaluate the model performance we measured mean squared error (MSE)($\hat{\boldsymbol{\beta}}^{\diamond}_{S_1}$) $:= \|\hat{\boldsymbol{\beta}}^{\diamond}_{S_1} - \boldsymbol{\beta}^*_{S_1}\|_2^2$ with $\diamond$ being either PSE or CIS-PSE. For the PSE, we obtained the relative MSE (RMSE) with respect to $\hat{\boldsymbol{\beta}}^{\text{WR}}_{S_1}$ as in GAF, and for the CIS-PSE, RMSE is with respect to $\hat{\boldsymbol{\beta}}^{\text{SWR}}_{S_1}$. That is, $\text{RMSE}(\hat{\boldsymbol{\beta}}^{\text{PSE}}_{S_1}) = \text{E}\|\hat{\boldsymbol{\beta}}^{\text{WR}}_{S_1} - \boldsymbol{\beta}^*_{S_1}\|_2^2/\text{E}\|\hat{\boldsymbol{\beta}}^{PSE}_{S_1} - \boldsymbol{\beta}^*_{S_1}\|_2^2$ and $\text{RMSE}(\hat{\boldsymbol{\beta}}^{\text{CIS-PSE}}_{S_1}) = \text{E}\|\hat{\boldsymbol{\beta}}^{\text{SWR}}_{S_1} - \boldsymbol{\beta}^*_{S_1}\|_2^2/\text{E}\|\hat{\boldsymbol{\beta}}^{\text{CIS-PSE}}_{S_1} - \boldsymbol{\beta}^*_{S_1}\|_2^2$. We also report numbers of correctly identified variables in $S_1$ and $S_2$ (denoted as $|\hat{S}_1|$ and $|\hat{S}_2|$) to evaluate the screening performance.

The results are shown in Table I based on 400 independent replications. We observe that the CIS-PSE outperforms the original PSE in the low-dimensional setting. Its performance is satisfactory even in the ultrahigh-dimensional setting, which defies the original PSE procedure. Moreover, the results seem to hint that incorporating MUJI signals improves estimation accuracy.

## 5. Conclusions

Our discussion is meant to address two fundamental questions surrounding GAF's PSE procedure: (1) can PSE be adopted for modeling ultrahigh-dimensional data; (2) can PSE incorporate variables that are marginally weak but highly correlated with some variables in $S_1$, and thus have joint effects on the response together with variables from $S_1$? Based on GAF's work, we have proposed a simple but efficient modification of PSE to address these two intriguing issues. The limited simulations conducted by us lent support to the benefit of considering MUJI variables in estimation and the feasibility of applications in ultrahigh-dimensional cases. We hope that our brief exploration adds some new perspectives to the development of post selection estimators and will appreciate the feedback from the authors.

Copyright © 2017 John Wiley & Sons, Ltd.

YANMING LI
*Department of Biostatistics*
*University of Michigan*
*Ann Arbor, MI 48109, USA*

HYOKYOUNG GRACE HONG
*Department of Statistics and Probability*
*Michigan State University*
*East Lansing, MI 48824, USA*

YI LI
*Department of Biostatistics*
*University of Michigan*
*Ann Arbor, MI 48109, USA*
Email: *yili@med.umich.edu*

Copyright © 2017 John Wiley & Sons, Ltd.

Applied Stochastic Models in Business and Industry

Discussion

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2214 Published online in Wiley Online Library

# Discussion of 'Post selection shrinkage estimation for high-dimensional data analysis'

The ever increasing ability to collect and store large amounts of data creates the need for novel statistical methods that can be used to analyze large data sets. This article is an important contribution to the development of statistical methods for high dimensional data. A good example where such methods are needed is the pharmaceutical industry where the study of the effect of pharmaceutical products on health for people with certain bio-markers is at the root of the current move to personalized medicine. Individual bio-markers such as SNPs typically have a weak effect on health responses. However, it is thought that a collection of bio-markers jointly can have an important effect. The search for such collections, sometimes called networks or pathways, is a topic of great concern for those working on personalized medicine.

The current paper by Gao, Ahmed, and Feng provides methods for examining the joint effect of bio-markers that individually have weak signals. The methods are based on subtracting out the effect of strong predictors for health such as age and body mass index, thereby giving bio-marker networks made up of predictors with individual weak signals a chance to be discovered.

The proposed method by Gao, Ahmed, and Feng consists of first selecting predictors with strong signals using a statistical variable selection procedure such as the Least Absolute Shrinkage and Selection Operator (Lasso). The next step is to compute a ridge regression estimate of the regression coefficients for the variables not selected in the first step and to delete variables whose estimated coefficients fall below a threshold. The third step is to use the results from steps one and two to construct a shrinkage estimate of the regression coefficients for the variables with strong signals. Our suggestion is that in step two, the joint effect of the weak predictors on the response be studied using a partial regression coefficient. See Doksum and Samarov [1, Section 4] and Rao [2, Section 4g.2].

The emphasis in the article is on improving on the Lasso estimator of the regression coefficients for the variables that provide strong signals. Constructions of improvements to Lasso estimators are favored activities of statisticians. However, we think that a more important contribution of the current paper is to the development of techniques that can be used to detect networks of individually weak signals that jointly have an important effect. Theorem 2 gives the asymptotic normal distribution of the estimates of the coefficients for the weak signals, but further studies of the properties of these estimates are of interest using for instance the partial correlation of collections of such signals with the response.

Finally, it is important to note that for the results of this article and others like it to lead to useful applications, interdisciplinary research projects are desirable.

DOKSUM KJELL
*Statistics Department, University of Wisconsin, Madison, USA*
E-mail: *kdoksum@gmail.com*

JOAN FUJIMURA
*Sociology Department, University of Wisconsin, Madison, USA*

## References

1. Doksum K, Samarov A. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics* 1995; **22**:1443–1473.
2. Rao CR. *Linear Statistical Inference and Its Applications* (2nd edn). Wiley: New York, 1973.

Copyright © 2017 John Wiley & Sons, Ltd.

Rejoinder

Applied Stochastic
Models in Business
and Industry

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2245 Published online in Wiley Online Library

# Rejoinder to ‘Post-selection shrinkage estimation for high-dimensional data analysis’

We sincerely thank all the discussants Kjell Doksum and Joan Fujimura (DF); Jianqing Fan (Fan); Peihua Qiu, Kai Yang, and Lu You (QYY); and Yanming Li, Hyokyoung Grace Hong, and Yi Li (LHL) for the thought-provoking and insightful discussions on our paper. We would also like to thank the Editor Fabrizio Ruggeri for processing and organizing the discussion. Ahmed would like to specially thank him for his encouragement on this paper and patience.

## 1. Strong signal, weak signal, and noise

One fundamental ingredient of our work is to formally split the signals into strong and weak ones. The rationale is that the usual one-step method such as the least absolute shrinkage and selection operator (LASSO) may be very effective in detecting strong signals while failing to identify some weak ones, which in turn has a significant impact on the model fitting, as well as prediction. The discussions of both Fan and QYY contain very interesting comments on the separation of the three sets of variables. Regarding Assumption (A2) about the weak signal set $S_2$, we admit that the original version was not as rigorous as it could have been, as it could have contained the variables in S3. We now propose the following Assumption (A2’) that replaces (A2) in the original paper.

(A2’): The parameter vector $\boldsymbol{\beta}^*$ satisfies that $\|\boldsymbol{\beta}^*_{S_2}\| \sim n^{\tau}$ for some $0 < \tau < 1$, where $\|\cdot\|$ is the $\ell_2$ norm and $\beta^*_j \neq 0$ for any $j \in S_2$.

QYY mentioned that in practice, it is sometimes difficult to have a subjective separation of strong and weak signals. First of all, we would like to emphasize that the conditions imposed in the paper are from an asymptotic point of view, which demonstrate the great performance of the proposed estimators in the specified scalings and covariance structure. Second, we would like to argue that this separation is sometimes unnecessary in practice as the ultimate goal of high-dimensional regression is to provide accurate predictions for future data after variable selection and insightful interpretations on the importance of the predictors in terms of explaining the response. Third, the separation of strong and weak signals was mainly used to stimulate the post-selection shrinkage estimation (PSE) method, and the variables identified as ‘strong’ or ‘weak’ by PSE do not necessarily have a natural separation in terms of true regression coefficients, at least for a fixed sample size.

## 2. Conditions on designed matrix

We thank Fan for pointing out that the assumption on the design matrix could be strong. In fact, condition (B2) is mainly motivated from [1], and it requires the weak signals to be correlated to strong ones, in order for it to be detectable using the weighted ridge regression. On the other hand, condition (B4) requires that the eigenvalues of the design matrix corresponding to both strong and weak signals are bounded away from both 0 and infinity. Now, we describe one specific example. Consider an $n \times p$ design matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3]$. $\mathbf{X}_1$ and $\mathbf{X}_2$ correspond to strong and weak signals, and $\mathbf{X}_3$ includes noises. Suppose all signals in $\mathbf{Z} = [\mathbf{X}_1, \mathbf{X}_2]$ are correlated with constant correlation coefficient of $r$ and uncorrelated with noises in $\mathbf{X}_3$. Then, such a design matrix satisfies both conditions (B2) and (B4). We agree that some reasonably correlated design matrix for all variables could be excluded under those conditions.

## 3. MUJI variables

We thank LHL for bringing up the marginally unimportant but jointly informative (MUJI) variable set [2], namely, ‘marginally unimportant but jointly important’ variables. Indeed, the inclusion of MUJI variables could significantly improve the performance of the vanilla sure independence screening approach [3]. However, we would like to argue that

Copyright © 2017 John Wiley & Sons, Ltd.

**Table I.** Simulated results for example 1

| Method | MUJI (Y/N) | | $p_n = 400$ | $p_n = 100,000$ |
|---|---|---|---|---|
| LASSO-PSE | | $\lvert\hat{S}_1\rvert$ | 6 | 6 |
| | No | MSE | 0.0015 | 0.0046 |
| | | MPE | 0.0501 | 0.1661 |
| | | RMSE | 5.0701 | 3.1477 |
| | Yes | MSE | 0.0022 | 0.0129 |
| | | MPE | 0.0279 | 0.5374 |
| | | RMSE | 3.4907 | 1.0945 |
| MCP-PSE | | $\lvert\hat{S}_1\rvert$ | 3 | 3 |
| | No | MSE | 0.1284 | 0.0196 |
| | | MPE | 1.9041 | 0.2629 |
| | | RMSE | 0.3456 | 0.6933 |
| | Yes | MSE | 0.0154 | 0.0049 |
| | | MPE | 0.2447 | 0.0755 |
| | | RMSE | 2.8821 | 2.6005 |
| CIS-PSE | | $\lvert\hat{S}_1\rvert$ | 3 | 3 |
| | No | MSE | 1.4339 | 0.1215 |
| | | MPE | 20.5638 | 1.6907 |
| | | RMSE | 0.0754 | 0.5184 |
| | Yes | MSE | 0.0431 | 0.0151 |
| | | MPE | 0.5850 | 0.3049 |
| | | RMSE | 2.5070 | 1.9322 |

Larger RMSE, smaller MSE, and smaller MPE indicate better performance. CIS, Covariance Insured Screening; MPE, mean prediction error; MSE, mean squared error; PSE, post-selection shrinkage estimation; RMSE, relative mean squared error.

in our proposal, the estimation of $S_1$ could be done by any variable selection method that could identify the strong signals, for example, LASSO. As a result, $S_1$ could already contain the MUJI variables as it considers the joint regression on all predictors.

Motivated by the MUJI variables, LHL proposed a new shrinkage estimator called Covariance Insured Screening-based PSE (CIS-PSE), which uses two simulation examples to compare HD-PSE and CIS-PSE. They conclude that using MUJI can help to improve the risk performance of the shrinkage estimator. However, the comparison could be a little unfair because $S_1$ in CIS-PSE is generated by marginal correlation, while $S_1$ in HD-PSE is from LASSO. Thus, $S_1$ generated from two methods can be different. To ensure a fair comparison, we let $S_1$ in the first step from both CIS-PSE and HD-PSE be consistent. We consider different scenarios: (i) $S_1$ is selected by LASSO; (ii) $S_1$ is selected by the minimax concave penalty (MCP); (iii) $S_1$ is selected using the marginal strong set suggested by LHL in the first step while producing CIS-PSE. For each of those aforementioned three cases, we compute the MUJI set $\hat{S}_{\text{MUJI}}$ as suggested by LHL and then shrinking $\hat{S}_1 \cup \hat{S}_2 \cup \hat{S}_{\text{MUJI}}$ in the direction of $\hat{S}_1$. We define those three estimates as LASSO-PSE, MCP-PSE, and CIS-PSE, correspondingly. We then recheck those two examples, compare their performance, and report the results in Tables I and II [‡]. When $p_n = 100,000$, we apply ridge regression and keep the 500 variables with the largest absolute coefficients before applying our algorithm.

In the tables, we report mean squared error and relative mean squared error. We also report the mean prediction error based upon the selected subset, defined as

$$E(\mathbf{X}_{\hat{S}_1}\hat{\boldsymbol{\beta}}^*_{S_1} - \mathbf{X}_{\hat{S}_1}\boldsymbol{\beta}^*_{\hat{S}_1})^2.$$

In Example 1 in LHL, there is strong correlation among three covariates with weak signals and three covariates with strong signals. From the evaluation results reported in Table I, we observe that when using the MCP-PSE and CIS-PSE,

[‡] *The results of LASSO-PSE and MCP-PSE are identical when $p_n = 400$ because they always select the same $\hat{S}_1$.*

Copyright © 2017 John Wiley & Sons, Ltd.

**Table II.** Simulated results for example 2

| Method | MUJI(Y/N) | | $p_n = 400$ | $p_n = 100,000$ |
|---|---|---|---|---|
| LASSO-PSE | | $\|\widehat{S}_1\|$ | 3 | 3 |
| | No | MSE | 0.0018 | 0.0100 |
| | | MPE | 0.0225 | 0.1594 |
| | | RMSE | 8.4982 | 4.6249 |
| | Yes | MSE | 0.0067 | 0.0294 |
| | | MPE | 0.0917 | 0.4487 |
| | | RMSE | 2.3025 | 1.5091 |
| MCP-PSE | | $\|\widehat{S}_1\|$ | 3 | 3 |
| | No | MSE | 0.0018 | 0.0100 |
| | | MPE | 0.0225 | 0.1594 |
| | | RMSE | 8.4982 | 4.6249 |
| | Yes | MSE | 0.0067 | 0.0294 |
| | | MPE | 0.0917 | 0.4487 |
| | | RMSE | 2.2974 | 1.5091 |
| CIS-PSE | | $\|\widehat{S}_1\|$ | 3 | 3 |
| | No | MSE | 1.2953 | 0.0100 |
| | | MPE | 13.5683 | 0.1594 |
| | | RMSE | 0.0719 | 4.6249 |
| | Yes | MSE | 0.0262 | 0.0294 |
| | | MPE | 0.3166 | 0.4487 |
| | | RMSE | 2.5238 | 1.5091 |

Larger RMSE, smaller MSE, smaller MPE indicate better performance. CIS, Covariance Insured Screening; MPE, mean prediction error; MSE, mean squared error; PSE, post-selection shrinkage estimation; RMSE, relative mean squared error.

incorporating the MUJI variables improves the performance of the method as it can include additional signals from the MUJI set. However, when using LASSO-PSE, it is clear that using MUJI actually deteriorates the performance of the method by having larger mean squared errors and smaller relative mean squared errors. This is probably because LASSO already selects some weak signals in additional to the strong signals, which makes the MUJI detection step unnecessary. In Example 2 in LHL, there is strong correlation among three noise covariates and three covariates with strong signals. From the evaluation results reported in Table II, we observe that both Lasso and MCP only select strong signals with no weak signals. Incorporating MUJI variables deteriorates the performances of both MCP-PSE and LASSO-PSE in this case. This is because MUJI variables may pick up those noises in the second step. However, CIS-PSE with MUJI variables can help to improve the performance of the method.

From this preliminary numerical study, we can see that including MUJI variables may or may not improve the performance of the PSE, depending on the selected submodel.

The corresponding theoretical analysis regarding when the MUJI variables help the final estimation is an interesting open research question.

## 4. About the algorithm

DF suggested to use the partial least square method in the second step to select the weak signals, as opposed to the current weighted ridge regression. We appreciate the suggestion; however, one still needs to impose regularization on the estimates, which would lead to a different strategy and should be of interest for further research.

QYY posed the question about the selection of the tuning parameters $a_n$ and $r_n$ in the PSE strategy. We agree that the proposed cross-validation method, while effective in our limited numerical experience, may need further theoretical justification. Recently, [4, 5] conducted a systematic study on the cross-validation-based tuning parameter selection method for high-dimensional penalized regression problems. Some work along similar lines could be an interesting research project. In

Copyright © 2017 John Wiley & Sons, Ltd.

addition, it is also important to develop a certain adaptive tuning parameter selection method and demonstrate its robustness against model misspecification.

## 5. Future directions

This paper introduced the post-shrinkage estimation framework and used specific methods to select the strong and weak signals. The shrinkage estimation received a lot of attention since its inception decades ago. It strikes a balance between post-selected submodels and high-dimensional weighted ridge estimators and is proved to be an effective strategy.

There are a number of alternatives to mimick the ideas of the PSE. For example, Fan suggested a great idea involving using the penalized least square with different penalty levels, closely related to the folded concave penalties including the smoothly clipped absolute deviations penalty (SCAD) and MCP.

The current methodology can be extended in a host of directions, including nonparametric models (suggested by QYY), spatially corrected data, among others. We would like to remark here that shrinkage estimation strategies have already been applied to some nonparametric models in low-dimensional cases such as [6–8], among others that can be extended to high-dimensional cases.

Another interesting direction would be to study the shrinkage method in robust high-dimensional data analysis, such as M-estimation. Recently, [9, 10] proposed penalized weighted least squares and penalized weighted least absolute deviation methods to study robust high-dimensional regression. The methods unify the M-estimation in a penalized weighted least squares and least absolute deviation framework. Such a connection will enable us to extend the post-selection shrinkage strategy to robust high-dimensional regression models.

The scope of research in PSE is expanding. How to develop a system of diagnostic tools for the high-dimensional post-shrinkage estimators is an important direction for future research, as suggested by QYY.

## Acknowledgements

Ahmed is supported by grants from NSERC. Gao is supported by Simons grant no. 359337. Feng is supported by NSF CAREER grant DMS-1554804.

Xiaoli Gao
*Department of Mathematics and Statistics,*
*University of North Carolina at Greensboro,*
*Greensboro, North Carolina, USA*
Email: *x_gao2@uncg.edu*

S. Ejaz Ahmed
*Department of Mathematics and Statistics,*
*Brock University, Saint Catharines,*
*Ontario, Canada*

Yang Feng
*Department of Statistics,*
*Columbia University, New York, USA*

## References

1. Shao J, Deng X. Estimation in high-dimensional linear models with deterministic design matrices. *Annals of Statistics* 2012; **40**:812–831.
2. Li Y, Hong H, Kang J, He K, Zhu J, Li Y. Classification with ultrahigh-dimensional features. *arXiv preprint arXiv:1611.01541* 2016.
3. Fan J, Lv J. Sure independence screening for ultra-high-dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B* 2008; **70**:849–911.
4. Yu Y, Feng Y. Modified cross-validation for LASSO penalized high-dimensional linear models. *Journal of Computational and Graphical Statistics* 2014; **23**:1009–1027.
5. Feng Y, Yu Y. Restricted consistent cross-validation for tuning parameter selection in high-dimensional variable selection. *manuscript* 2016.
6. Hossain S, Ahmed SE, Yi GY, Chen B. Shrinkage and pretest estimators for longitudinal data analysis under partially linear models. *Journal of Nonparametric Statistics* 2016; **28**:531–549.
7. Buhamra S, Al-Kandarri S, Ahmed SE. Nonparametric inference strategies for the quantile functions under left truncation and right censoring. *Journal of Nonparametric Statistics* 2007; **19**:189–198.

Copyright © 2017 John Wiley & Sons, Ltd.

8. Ahmed SE, Hussein AA, Sen PK. Risk comparison of some shrinkage M-estimators in linear models. *Journal of Nonparametric Statistics* 2006; **18**:401–415.
9. Gao XL, Fang Y. Penalized weighted least squares for outlier detection and robust regression. https://arxiv.org/abs/1603.07427, 2016.
10. Gao XL, Feng Y. Penalize weighted least absolute deviation regression. *Statistics and Its Interface* 2017. Accepted for publication.

Copyright © 2017 John Wiley & Sons, Ltd.