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_{MUJI}, and for simultaneously estimating $\boldsymbol{\beta}$ based on the three sets, S_1, S_2, and S_{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_{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.

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 τ_n. Set $\hat{\boldsymbol{\beta}}^{\text{MS}}_{\hat{S}_1} = (\mathbf{X}'_{\hat{S}_1}\mathbf{X}_{\hat{S}_1})^{-1}\mathbf{X}'_{\hat{S}_1}\mathbf{y}$. If the number of variables in $\hat{S}_1$ exceeds the sample size, a Lasso regression can be used instead. Here, $\hat{\boldsymbol{\beta}}^{\text{MS}}_{\hat{S}_1}$ plays the same role as $\hat{\boldsymbol{\beta}}^{\text{RE}}_{\hat{S}_1}$ 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 ν_n is a tuning parameter.
   The set of MUJI variables is selected by $\hat{S}_{\text{MUJI}} = \{j \in \hat{S}_1^c : |\mathbf{X}_j'\mathbf{X}_{j'}| > \rho_n \text{ for some } j' \in \hat{S}_1\}$, where ρ_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 $\hat{\boldsymbol{\beta}}^{\text{SWR}}_{\hat{S}_1}$ the components of $\hat{\boldsymbol{\beta}}^{\text{SWR}}$ corresponding to $\hat{S}_1$. Though $\hat{\boldsymbol{\beta}}^{\text{SWR}}_{\hat{S}_1}$ 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

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

   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_{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_{MUJI}.

Example 1
Assume that ϵ_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 6×6 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 σ 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

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

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

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 $\mathbf{X}_{i,S_1 \cup S_{\text{MUJI}}} \sim N(\mathbf{0}, \mathbf{\Sigma})$ and

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

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 τ_n, ν_n, ρ_n and α_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.

Applied Stochastic Models in Business and Industry

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*