Discussion of "Post Selection Shrinkage Estimation for High Dimensional Data Analysis"

Peihua Qiu, Kai Yang and Lu You
Department of Biostatistics
University of Florida
Gainesville, FL 32610, USA

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 has 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", "machine learning", and so forth. 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 who 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 to 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.

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, and i.i.d. and additive random noise with mean 0 and constant variance σ^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 who provide useful information for describing the response variable y, although their contributions might be relatively small, compared to 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 mentioned above, 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. 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 dimention reduction (e.g., Mai and Zou 2015). 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 2013).

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 β_j^* is the jth 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 1,000 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.

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.

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

Extra References

Mai, Q., and Zou, H. (2015), "Nonparametric Variable Transformation in Sufficient Dimension Reduction," *Technometrics*, **57**, 1–10.

Mukherjee, P.S., and Qiu, P. (2013), "Efficient Bias Correction For MRI Image Denoising," *Statistics in Medicine*, **32**, 2079–2096.