Harrison H. Zhou (*Yale University*) **and Xihong Lin** (*Harvard University*)

We congratulate the authors on a stimulating paper that provide theoretical justifications for using simple marginal regression (or correlation learning) to pre-screen variables followed by a joint variable selection procedure to perform variable selection for ultra-high dimensional data. Such an univariate screening method has been commonly used in practice, such as genome-wide association studies, and microarrays and proteomics, where the number of genes (proteins) is often much higher than the number of subjects. It is of substantial interest to understand the theoretical properties of these simple screening methods. The attractive theoretical results in this paper are of considerable practice interest. The proposed methods can be extended to other regression settings, such as generalized linear models and general greedy algorithms.

An important finding of this paper is that the proposed method can identify the true model with a high probability even for ultrahigh dimensional variable selection settings like $p = \exp\left(n^{\xi}\right)$, with $\xi > 0$ arbitrarily large. To understand when this result can be applied in practice, e.g., the assumption on the magnitude of p and n and the required amount of signal relative to noise, we consider a special linear model. Denote by the $n \times p$ design matrix by $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_p)$. Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with (i) $\mathbf{X}_i \sim N(0, I_{p\times p})$ independent; (ii) $p = \exp\left(n^{\xi}\right)$ with $\xi > 0$; (iii) $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_{n\times n})$; and (iv) $\beta_1 = n^{-\kappa}$ for some $\kappa < 1/2$, and $\beta_j = 0$ for all $j \geq 2$. Let $\lambda = \infty$, and write

$$\begin{aligned} \hat{\beta}_1 &= \frac{1}{n}\mathbf{X}_1^T\mathbf{y} = \frac{1}{n}\left\|\mathbf{X}_1\right\|^2\beta_1 + \frac{1}{n}\mathbf{X}_1^T\boldsymbol{\varepsilon} \\ \hat{\beta}_j &= \frac{1}{n}\mathbf{X}_j^T\mathbf{y} = \frac{1}{n}\mathbf{X}_j^T\left(\mathbf{X}_1\beta_1 + \boldsymbol{\varepsilon}\right) \quad j \geq 2. \end{aligned}$$

Note that $\hat{\beta}_1 = (1 + o(1))\beta_1 = (1 + o(1))n^{-\kappa}$. For given X_1 and $\boldsymbol{\varepsilon}$, $\hat{\beta}_j$ are i.i.d. normal, then the maximum noise

$$max_{2\leq j\leq p}\left(\hat{\beta}_j\right) = \{1 + o(1)\}\sqrt{2\log p}\sqrt{\frac{1+\beta_1^2}{n}} = \{1 + o(1)\}\sqrt{2\left(1+\beta_1^2\right)}n^{(\xi-1)/2}.$$

This implies the true model can be identified with a high probability when $\kappa < (1-\xi)/2$, i.e., $\xi < 1 - 2\kappa$. It is difficult to identify the true model when $\xi > 1 - 2\kappa$, as the maximum noise dominates the true signal. We would be interested in learning whether the authors' method can be applied to identify the true model when $\xi > 1 - 2\kappa$.

The example above is related to a scenario that some predictors are highly correlated. For example, when p is large, it is expected that there is a predictor X_j with $j \geq 2$ such that the sample correlation coefficient between X_j and true predictor X_1 is very close to 1. The authors proposed a useful iterative sure independence screening procedure to deal with the correlated X

case. To help effectively apply the proposed procedure in practice, we would be interested in any guideline the authors could provide on how to choose k_i in each step and when to stop to ensure the true model can be identified with a high probability, or β can be estimated with some nice risk property. Further, it seems that if a variable is selected in the previous steps, it cannot be deleted in later steps. Intuitively, it would be desirable to let variables in and out in each step. Can the authors' procedure be modified to allow for this? We do realize that the problem might get more complicated for a general covariance matrix of $\mathbf{x}$, e.g., when the covariance is nonstationary, such as a constant exchangeable correlation among the x's. The concentration property may not hold in the case. We would be interested in learning whether the proposed method is applicable to such correlated X cases? What would be the required assumptions about the signals relative to the noises?

We would like to make one minor comment. We think under *Condition 3* of the paper the term $\log d$ for the risk of method SIS–DS in Theorem 4 may not be necessary. Hence the result can be made more interesting.

We would like to congratulate the authors again for a stimulating paper that opens new opportunities for more research in variable selection for ultrahigh dimensional data.