Comment on "Sure Independence Screening for Ultra-high Dimensional Feature Space" by Jianqing Fan and Jinchi Lv

Elizaveta Levina
Assistant Professor
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Ji Zhu
Assistant Professor
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

We congratulate the authors on developing an attractive and practical method for high-dimensional variable selection with many potential applications. One area where this method may have applications is genome-wide association studies, where one is interested in identifying common genetic factors that influence health and disease. For complex diseases and traits, the genetic contribution of a true association is often expected to be moderate. In terms of the model in the paper, this would result in a low signal-to-noise ratio (SNR) $\mathrm{Var}(X\beta)/\mathrm{Var}(\epsilon)$, but in the simulations in Section 4, the signal to noise ratios are all relatively high, ranging from 40 to 200. Thus we decided to briefly investigate the behavior of ISIS under lower SNR levels.

To illustrate the point, we mimicked the simple simulated example I from Section 4.2.1. Specifically, we considered the linear model

$$Y = 5X_1 + 5X_2 + 5X_3 + \epsilon,$$

where $X_1, \ldots, X_p$ ($p = 1000$) have a multivariate normal distribution $N(0, \Sigma)$ and $\epsilon \sim N(0, \sigma^2)$ is independent of the predictors. The covariance matrix Σ has diagonal elements equal to 1 and off-diagonal elements equal to $\rho = 0.5$. Instead of setting $\sigma = 1$, which corresponds to SNR of 150, we considered several different values of σ, i.e., $\sigma = 1, 2, 4, 8, 12, 16$. The last scenario $\sigma = 16$ corresponds to SNR of 0.6.

We considered $n = 25, 50$ and 100, and ran ISIS exactly as described in Section 4.1.1 using Lasso at the second stage of variable selection, with the tuning parameter selected by BIC. For each simulation, we recorded how many important variables (out of X_1, X_2 and X_3) were selected. The results over 100 replications are summarized in Table 1.

Not surprisingly, as the signal to noise ratio decreases, the performance of ISIS degrades. For example, when $n = 50$, $p = 1000$ and $\sigma = 12$, which corresponds to SNR about 1 (considered relatively high in genome-wide association studies), the ISIS was still able to identify some important variables (identifying at least one important variable in 61 out of 100 simulations), but was not able to identify all three important variables.

A theoretical question to consider is whether the asymptotic results in the paper could be extended to incorporate SNR or σ^2 into the rate explicitly. Modifications of the method that would allow it to be applied to low SNR large-scale problems may also be an interesting topic for further investigation.

Table 1: We set $p = 1000$ and $\rho = 0.5$. The true model contains three important variables. #3 corresponds to the number of times that all three important variables were selected by ISIS out of 100 simulations, similarly for #2, #1 and #0.

	$n = 25$				$n = 50$				$n = 100$			
σ	#0	#1	#2	#3	#0	#1	#2	#3	#0	#1	#2	#3
1	16	32	18	34	0	0	0	100	0	0	0	100
2	21	35	20	24	1	1	2	96	0	0	0	100
4	23	37	25	15	2	4	22	72	0	0	1	99
8	62	35	2	1	9	42	38	11	1	0	26	73
12	73	25	2	0	39	39	22	0	12	20	45	23
16	84	16	0	0	63	32	5	0	25	51	17	7