

Addendum *

Jianqing Fan and Jinchi Lv

Princeton University and University of Southern California

December 1, 2008

This note is an addendum to “Sure independence screening for ultrahigh dimensional feature space (with discussion)”, *Journal of the Royal Statistical Society Series B* (2008) 70 (5), 849–911. At the request by some readers, we would like to clarify the implementation of our iterative SIS (ISIS) introduced in Section 4 of the paper. At each step, the correlation screening applies to the correlation of all the remaining variables with the residual conditional on the selected variables, i.e., the partial correlation of all the remaining variables with the response given the variables selected in previous steps. While producing additional numerical results upon the requests by some readers, we found an implementation error that affects some simulation results reported in Tables 5 and 6 of the paper. We sincerely apologize for the mistake. The numbers in bold face in Tables 5 and 6 below are the corrected ones. We also add Tables 5* and 6* as extensions to Tables 5 and 6, respectively. Tables 5* reports the percentages that SIS, the lasso and ISIS miss at most one variable in simulated example II, and Tables 6* reports the percentages that the three methods miss at most one or two variables in simulated example III.

Table 5: Results of simulated example II: accuracy of SIS, the lasso and ISIS in including the true model $\{X_1, X_2, X_3, X_4\}$ ($\rho = 0.5$)

p	Method	$n = 20$	$n = 50$	$n = 70$
100	SIS	0.025	0.490	0.740
	Lasso	0.000	0.360	0.915
	ISIS	0.425	0.925	0.990
1000	SIS	0.000	0.000	0.000
	Lasso	0.000	0.000	0.000
	ISIS	0.030	0.990	0.995

Similar conclusions continue to hold for ISIS. From Tables 5 and 6, we see that ISIS can generally improve the performance of SIS in these scenarios. The performance of ISIS is better than that of Lasso in the scenario of example II, and is comparable with that of Lasso in the scenario of example III. When the dimensionality is too high compared to sample size and there

*This is an addendum to “Sure independence screening for ultrahigh dimensional feature space (with discussion)” by the authors, *Journal of the Royal Statistical Society Series B* (2008) 70 (5), 849–911.

Table 6: Results of simulated example III: accuracy of SIS, the lasso and ISIS in including the true model $\{X_1, X_2, X_3, X_4, X_5\}$ ($\rho = 0.5$)

p	Method	$n = 20$	$n = 50$	$n = 70$
100	SIS	0.000	0.285	0.645
	Lasso	0.000	0.310	0.890
	ISIS	0.000	0.430	0.850
1000	SIS	0.000	0.000	0.000
	Lasso	0.000	0.000	0.000
	ISIS	0.000	0.000	0.000

is some true variable which is very weakly correlated with the response, both SIS and ISIS can also fail to select all true variables, and Lasso can fail as well. Example III for $p = 1000$ is especially challenging given the fact that the variable X_5 was introduced in a way such that it has the same proportion of contribution to the response as the noise ε does.

Table 5*: Results of simulated example II: the percentages that SIS, the lasso and ISIS miss at most one variable ($\rho = 0.5$)

p	Method	$n = 20$	$n = 50$	$n = 70$
100	SIS	0.590	0.985	0.995
	Lasso	0.730	1	1
	ISIS	0.725	1	1
1000	SIS	0.090	0.785	0.915
	Lasso	0.100	0.995	1
	ISIS	0.065	1	0.995

Table 6*: Results of simulated example III: the percentages (labeled as ≥ 4 and ≥ 3) that SIS, the lasso and ISIS miss at most one or two variables ($\rho = 0.5$)

p	Method	≥ 4			≥ 3		
		$n = 20$	$n = 50$	$n = 70$	$n = 20$	$n = 50$	$n = 70$
100	SIS	0.160	0.840	0.950	0.695	0.985	1
	Lasso	0.305	0.985	1	0.850	1	1
	ISIS	0.205	0.995	1	0.620	0.995	1
1000	SIS	0.005	0.245	0.350	0.115	0.780	0.925
	Lasso	0.005	0.755	0.970	0.150	1	1
	ISIS	0.005	0.640	0.830	0.065	0.985	0.985