Supplemental Material for Model Free Feature Screening for Ultrahigh-Dimensional Data Conditional on Some Variables

Yi Liu[a,b] and Qihua Wang[a,c, *]

[a] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

[b] *College of Science, China University of Petroleum Qingdao 266580, China*

[c] *Institute of Statistical Science, Shenzhen University, Shenzhen 518006, China*

1 Proof of Step 1

We denote the positive constants c and C as generic constants depending on the context, which can vary from line to line. For simplicity, we write $Z(w)$ with Z for short, where $Z(\cdot)$ is an arbitrary function.

Step 1: For some $0 \leq \kappa < 1/2$, we first prove

$$\max_{1\leq j\leq p} \sup_{w\in[a,b]} P(|\hat{\rho}^2(X_j, Y|W = w) - \rho^2(X_j, Y|W = w)| \geq cn^{-\kappa}) \leq C\exp(-\frac{n^{-\kappa}}{Ch}). \tag{1}$$

*Corresponding author. E-mail address: qhwang@amss.ac.cn (Q. Wang).

Define

$$
\begin{aligned}
\hat{Z}_1 &= \frac{1}{n^2}\sum_{k,l=1}^{n} K(\frac{W_k-w}{h})K(\frac{W_l-w}{h}),\\
\hat{Z}_2 &= \frac{1}{n^2}\sum_{k,l=1}^{n} d_{kl}^{Y} K(\frac{W_k-w}{h})K(\frac{W_l-w}{h}),\\
\hat{Z}_{3,j} &= \frac{1}{n^2}\sum_{k,l=1}^{n} d_{kl}^{X_j} K(\frac{W_k-w}{h})K(\frac{W_l-w}{h}),\\
\hat{Z}_{4,j} &= \frac{1}{n^2}\sum_{k,l=1}^{n} (d_{kl}^{X_j})^2 K(\frac{W_k-w}{h})K(\frac{W_l-w}{h}),\\
\hat{Z}_5 &= \frac{1}{n^2}\sum_{k,l=1}^{n} (d_{kl}^{Y})^2 K(\frac{W_k-w}{h})K(\frac{W_l-w}{h}),\\
\hat{Z}_{6,j} &= \frac{1}{n^2}\sum_{k,l=1}^{n} d_{kl}^{X_j} d_{kl}^{Y} K(\frac{W_k-w}{h})K(\frac{W_l-w}{h}),\\
\hat{Z}_7 &= \frac{1}{n^3}\sum_{k,l,m=1}^{n} K(\frac{W_k-w}{h})K(\frac{W_l-w}{h})K(\frac{W_m-w}{h}),\\
\hat{Z}_{8,j} &= \frac{1}{n^3}\sum_{k,l,m=1}^{n} d_{kl}^{X_j} d_{km}^{X_j} K(\frac{W_k-w}{h})K(\frac{W_l-w}{h})K(\frac{W_m-w}{h}),\\
\hat{Z}_9 &= \frac{1}{n^3}\sum_{k,l,m=1}^{n} d_{kl}^{Y} d_{km}^{Y} K(\frac{W_k-w}{h})K(\frac{W_l-w}{h})K(\frac{W_m-w}{h}),\\
\hat{Z}_{10,j} &= \frac{1}{n^3}\sum_{k,l,m=1}^{n} d_{kl}^{X_j} d_{km}^{Y} K(\frac{W_k-w}{h})K(\frac{W_l-w}{h})K(\frac{W_m-w}{h}).
\end{aligned}
$$

Thus $\hat{\rho}^2(X_j, Y|W=w)$ can be written as

$$
\hat{\rho}^2(X_j, Y|W=w) = \frac{\hat{Z}_1\hat{Z}_{6,j}\hat{Z}_7^2 + \hat{Z}_2\hat{Z}_{3,j}\hat{Z}_7^2 - 2\hat{Z}_1^2\hat{Z}_7\hat{Z}_{10,j}}{\sqrt{\hat{Z}_1\hat{Z}_{4,j}\hat{Z}_7^2 + \hat{Z}_{3,j}^2\hat{Z}_7^2 - 2\hat{Z}_1^2\hat{Z}_7\hat{Z}_{8,j}}\sqrt{\hat{Z}_1\hat{Z}_5\hat{Z}_7^2 + \hat{Z}_2^2\hat{Z}_7^2 - 2\hat{Z}_1^2\hat{Z}_7\hat{Z}_9}}. \quad (2)
$$

To prove (1), according to Lemma S5 of Liu et al. (2014), we first show

$$
\max_{1\le j\le p}\sup_{w\in[a,b]} P(|\hat{Z}_{6,j} - h^2 f^2(w) E(d_{12}^{X_j} d_{12}^{Y}|W_1=w, W_2=w)| \ge 8cn^{-\kappa}) \le C\exp(-\frac{cn^{-\kappa}}{h}),
$$

and the other desired inequalities can be obtained analogously.

Define

$$
\begin{aligned}
Z_{6,j} &= E\hat{Z}_{6,j} = E d_{kl}^{X_j} d_{kl}^{Y} K(\frac{W_k-w}{h})K(\frac{W_l-w}{h}),\\
\hat{Z}_{6,j}^{*} &= \frac{1}{n(n-1)}\sum_{k\neq l}^{n} d_{kl}^{X_j} d_{kl}^{Y} K(\frac{W_k-w}{h})K(\frac{W_l-w}{h}).
\end{aligned}
$$

By condition (C1) and the assumptions in Theorem 1, it can be shown that $\max_{1\le j\le p}\sup_{w\in[a,b]}|Z_{6,j}-h^2f^2(w)E(d_{12}^{X_j}d_{12}^{Y}|W_1=w,W_2=w)|\le Ch^3<4cn^{-\kappa}$ for large enough n. Consequently,

$$\begin{aligned}
& \max_{1\le j\le p}\sup_{w\in[a,b]} P(|\hat{Z}_{6,j}-h^2f^2(w)E(d_{12}^{X_j}d_{12}^{Y}|W_1=w,W_2=w)|\ge 8cn^{-\kappa}) \\
\le & \max_{1\le j\le p}\sup_{w\in[a,b]} P(|\hat{Z}_{6,j}-Z_{6,j}| \\
& \qquad\qquad + |Z_{6,j}-h^2f^2(w)E(d_{12}^{X_j}d_{12}^{Y}|W_1=w,W_2=w)|\ge 8cn^{-\kappa}) \\
\le & \max_{1\le j\le p}\sup_{w\in[a,b]} P(|\hat{Z}_{6,j}-Z_{6,j}|\ge 4cn^{-\kappa}). \qquad (3)
\end{aligned}$$

By using the Cauchy-Schwartz inequality and condition (C2), it follows

$$Z_{6,j}\le M_0^2E|d_{kl}^{X_j}d_{kl}^{Y}|\le M_0^2(E(d_{kl}^{X_j})^2E(d_{kl}^{Y})^2)^{1/2},$$

where $|K(t)|\le M_0<\infty$ hold uniformly over the support of $K(\cdot)$. This, together with condition (C3), implies that $Z_{6,j}$ is uniformly bounded in p. That is, $\max_{1\le j\le p}Z_{6,j}<\infty$. Take n large enough such that $Z_{6,j}/n<2cn^{-\kappa}$. Then, it can be easily shown that

$$\begin{aligned}
& \max_{1\le j\le p}\sup_{w\in[a,b]} P(|\hat{Z}_{6,j}-Z_{6,j}|\ge 4cn^{-\kappa}) \\
= & \max_{1\le j\le p}\sup_{w\in[a,b]} P(|\hat{Z}^*_{6,j}\frac{n-1}{n}-Z_{6,j}\frac{n-1}{n}-Z_{6,j}\frac{1}{n}|\ge 4cn^{-\kappa}) \\
\le & \max_{1\le j\le p}\sup_{w\in[a,b]} P(|\hat{Z}^*_{6,j}-Z_{6,j}|\frac{n-1}{n}\ge 4cn^{-\kappa}-Z_{6,j}\frac{1}{n}) \\
\le & \max_{1\le j\le p}\sup_{w\in[a,b]} P(|\hat{Z}^*_{6,j}-Z_{6,j}|\ge 2cn^{-\kappa}). \qquad (4)
\end{aligned}$$

To establish the uniform consistency of $\hat{Z}_{6,j}$ in j, it suffices to show the uniform consistency of $\hat{Z}^*_{6,j}$ in j. Let $h(X_{jk},Y_k,W_k;X_{jl},Y_l,,W_l;w)=d_{kl}^{X_j}d_{kl}^{Y}K(\frac{W_k-w}{h})K(\frac{W_l-w}{h}):=hI(h>M)+hI(h\le M)$ be the kernel of the U statistics $\hat{Z}^*_{6,j}$, where M will be specified later. Then

$$\begin{aligned}
\hat{Z}^*_{6,j} &= \frac{1}{n(n-1)}\sum_{k\ne l}^{n}hI(h\le M)+\frac{1}{n(n-1)}\sum_{k\ne l}^{n}hI(h>M) \\
&:= \hat{Z}^*_{61,j}+\hat{Z}^*_{62,j}.
\end{aligned}$$

Accordingly, we decompose the $Z_{6,j}$ into

$$Z_{6,j} = EhI(h\le M)+EhI(h>M):=Z_{61,j}+Z_{62,j}.$$

We prove the uniform consistency of $\hat{Z}^*_{61,j}$ in j, first, with the Markov's inequality, for any $t > 0$, we have

$$P(\hat{Z}^*_{61,j} - Z_{61,j} \geq cn^{-\kappa}) \leq \exp(-tcn^{-\kappa}) \exp(-tZ_{61,j}) E \exp(t\hat{Z}^*_{61,j}). \tag{5}$$

With a similar discussion about U statistics in Zhu et al. (2011), the U statistics $\hat{Z}^*_{61,j}$ can be represented as an average of iid random variables. That is,

$$\hat{Z}^*_{61,j} = (n!)^{-1} \sum_{n!} \Omega(X_{j1}, Y_1, W_1; \cdots; X_{jn}, Y_n, W_n, w),$$

where $\sum_{n!}$ denotes the summation over all possible permutations of $(1, 2, \cdots, n)$, and each $\Omega(X_{j1}, Y_1; \cdots; X_{jn}, Y_n)$ is an average of $m = [n/2]$ iid random variables. Thus, it follows from the Jensen's inequality that, for $j = 1, 2, \cdots, p$,

$$\begin{aligned} E \exp(t\hat{Z}^*_{61,j}) &= E \exp(t(n!)^{-1} \sum_{n!} \Omega(X_{j1}, Y_1, W_1; \cdots; X_{jn}, Y_n, W_n, w)) \\ &\leq (n!)^{-1} \sum_{n!} E \exp(t\Omega(X_{j1}, Y_1, W_1; \cdots; X_{jn}, Y_n, W_n, w)) \\ &= E^m \exp(tm^{-1} hI(h \leq M)). \end{aligned} \tag{6}$$

According to Lemma 1 in Li et al. (2012), for $j = 1, 2, \cdots, p$, by (5), (6), we have

$$\begin{aligned} P(\hat{Z}^*_{61,j} - Z_{61,j} \geq cn^{-\kappa}) &\leq \exp(-tcn^{-\kappa}) E^m \exp(m^{-1}t[hI(h \leq M) - Z_{61,j}]) \\ &\leq \exp(-tcn^{-\kappa} + M^2 t^2 / 8m). \end{aligned}$$

By choosing $t = 4cn^{-\kappa} m / M^2$ such that $P(\hat{Z}^*_{61,j} - Z_{61,j} \geq cn^{-\kappa}) \leq \exp(-2c^2 n^{-2\kappa} m / M^2)$. Similarly, we can prove that $P(\hat{Z}^*_{61,j} - Z_{61,j} \leq -cn^{-\kappa}) \leq \exp(-2c^2 n^{-2\kappa} m / M^2)$. Therefore, for $j = 1, 2, \cdots, p$, we have

$$P(|\hat{Z}^*_{61,j} - Z_{61,j}| \geq cn^{-\kappa}) \leq 2 \exp(-2c^2 n^{-2\kappa} m / M^2) = 2 \exp(-\frac{cn^{-\kappa}}{h} \frac{2cn^{-\kappa} mh}{M^2}). \tag{7}$$

Since $h = O(n^{-\gamma})$, we take $M = O(n^\tau)$, where $\gamma - \kappa < \tau < (1 - \gamma - \kappa)/2$, such that

$$\frac{2cn^{-\kappa} mh}{M^2} = Cn^{1-\gamma-\kappa-2\tau} > 1,$$

for sufficiently large n. (7) is then simplified as

$$P(|\hat{Z}^*_{61,j} - Z_{61,j}| \geq cn^{-\kappa}) \leq C \exp(-\frac{cn^{-\kappa}}{h}). \tag{8}$$

Next, we consider the uniform consistency of $\hat{Z}^*_{62,j}$ in j, with Cauchy-Schwartz inequality and Markov's inequality, we have

$$Z^2_{62,j} \leq E(h^2)P(h > M) \leq E(h^2)\exp(-s'M)E(\exp(s'h)) \quad \text{for} \ \ 0 < s' < s_0.$$

By some simple inequalities, we have

$$\begin{aligned}
h &= \{(X_{jk} - X_{jl})^2(Y_k - Y_l)^\top(Y_k - Y_l)K^2(\frac{w_k - w}{h})K^2(\frac{w_l - w}{h})\}^{1/2} \\
&\leq 2M_0^2\{(X_{jk}^2 + X_{jl}^2)(Y_k^2 + Y_l^2)\}^{1/2} \\
&\leq M_0^2\{(X_{jk}^2 + X_{jl}^2 + Y_k^2 + Y_l^2)^2\}^{1/2} \\
&\leq M_0^2(X_{jk}^2 + X_{jl}^2 + Y_k^2 + Y_l^2),
\end{aligned}$$

which yields that

$$E(\exp(s'h)) \leq E(\exp(s'M_0^2(X_{jk}^2 + X_{jl}^2 + Y_k^2 + Y_l^2))).$$

By condition (C3), and notice that $M = O(n^\tau)$, then $\max_{1\leq j\leq p} Z_{62,j} < cn^{-\kappa}/2$ when n is sufficiently large. Thus, for $j = 1, 2, \cdots, p$, we get

$$\begin{aligned}
P(|\hat{Z}^*_{62,j} - Z_{62,j}| > cn^{-\kappa}) &\leq P(|\hat{Z}^*_{62,j}| > cn^{-\kappa}/2) \\
&\leq P(X_{ji}^2 + Y_i^2 > M/(2M_0^2), \ \ \text{for some } 1 \leq i \leq n) \\
&\leq nP(X_{ji}^2 > M/(4M_0^2)) + nP(Y_i^2 > M/(4M_0^2)) \\
&\leq 2nC\exp(-sM/(4M_0^2)) \\
&= 2C\exp(-\frac{cn^{-\kappa}}{h}\frac{h(sM/(4M_0^2) - \log n)}{cn^{-\kappa}}).
\end{aligned}$$

The second line is because that if $X_{ji}^2 + Y_i^2 \leq M/(2M_0^2)$ for all $1 \leq i \leq n$, we then have $h \leq M_0^2(X_{jk}^2 + X_{jl}^2 + Y_k^2 + Y_l^2) \leq M$. This leads to $|\hat{Z}^*_{62,j}| = 0$, which contradict with $|\hat{Z}^*_{62,j}| > cn^{-\kappa}/2$. The fourth line is due to Markov's inequality and condition (C3). Since $\frac{h(sM/(4M_0^2)-\log n)}{cn^{-\kappa}} = C(n^{\kappa+\tau-\gamma} - n^{\kappa-\gamma}\log n) > 1$, then, for $j = 1, 2, \cdots, p$, we have

$$P(|\hat{Z}^*_{62,j} - Z_{62,j}| > cn^{-\kappa}) \leq C\exp(-\frac{cn^{-\kappa}}{h}). \tag{9}$$

Summarizing equations (3), (4), (8), (9), we obtain

$$\max_{1\leq j\leq p}\sup_{w\in[a,b]} P(|\hat{Z}_{6,j} - h^2f^2(w)E(d_{12}^{X_j}d_{12}^Y|W_1 = w, W_2 = w)| \geq 8cn^{-\kappa}) \leq C\exp(-\frac{cn^{-\kappa}}{h}).$$

□

2 Additional simulation

Following Example 3 in Fan et al. (2014), we further consider a linear varying coefficient model setup. Let $\{Z_1, \cdots, Z_p\}$ be iid standard normal, $\{U_1, U_2\}$ be iid standard unifromly distributed random variables, and the noise ϵ follows the standard normal distribution. The model is specified as:

$$Y = 2X_1 + 3WX_2 + (W+1)^2 X_3 + \frac{4\sin(2\pi W)}{2 - \sin 2\pi W} X_4 + \epsilon,$$

where $X_j = (Z_j + t_1 U_1)/(1 + t_1)$, $j = 1, \cdots, p$ and $W = (U_2 + t_2 U_1)/(1 + t_2)$. t_1, t_2 controls the correlation among the predictors $\mathbf{X}$ and the correlation between $\mathbf{X}$ and W respectively. We study one setting: $t_1 = t_2 = 0$, which results in uncorrelated case.

We take p to be 1000, and the sample size n is 200, the model size d is chosen to be $d_i = i[n^{4/5}/log(n^{4/5})]$, $i = 1, 2, 3$, where $[a]$ denotes the integer part of a. Let $0 = s_0 < s_1 < \cdots < s_{10} < s_{11} = 1$ be a partition of the interval. Using the s_i as knots, we construct 7 normalized cubic (order is 3) B-spline basis functions for NIS method. All the simulations are based on 500 replications. Similarly, we report the S, P_j and P_{All} in Table 1-2. We further present the boxplot of S in Figure 1.

Insert Table 1-2, Figure 1 here

Based on the results in Table 1-2 and Figure 1, we can see that in this uncorrelated case, CDC-SIS, CC-SIS, NIS perform well and behave better than the unconditional screening methods SIS, DC-SIS and DC-RoSIS. NIS behave comparable to the CDC-SIS and CC-SIS according to the top fifty percent quantiles of S. However, in terms of the 75% and 95% quantiles of S, NIS method need a larger model size to include all active predictors than the CDC-SIS and CC-SIS methods. Moreover, P_{All} of NIS is a little lower than those of CDC-SIS and CC-SIS, but all these three methods outperform the unconditional screening methods significantly.

References

Fan, J., Ma, Y., and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109(507):1270–1284.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 109(505):266–274.

Zhu, L. P., Li, L., Li, R. and Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475.

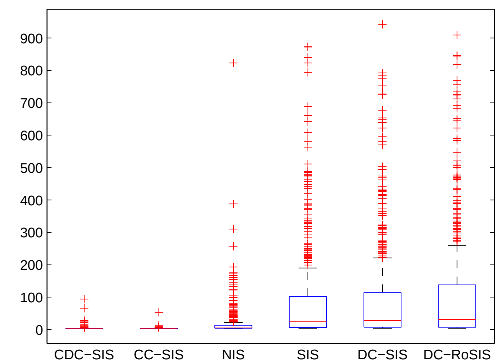
Table 1: **The quantile of S**

t_1, t_2	method	5%	25%	50%	75%	95%
$t_1 = t_2 = 0$	CDC-SIS	4.0000	4.0000	4.0000	4.0000	5.0000
	CC-SIS	4.0000	4.0000	4.0000	4.0000	5.0000
	NIS	4.0000	4.0000	5.0000	13.0000	74.0000
	SIS	4.0000	6.0000	25.5000	102.0000	410.5000
	DC-SIS	4.0000	7.0000	28.0000	114.0000	427.5000
	DC-RoSIS	4.0000	7.0000	31.0000	138.0000	472.5000

Table 2: **The proportion of P_j and P_{All}**

			P_j				P_{All}
t_1, t_2	method	size	X_1	X_2	X_3	X_4	All
$t_1 = t_2 = 0$	CDC-SIS	d_1	1.0000	0.9880	1.0000	1.0000	0.9880
		d_2	1.0000	0.9960	1.0000	1.0000	0.9960
		d_3	1.0000	0.9960	1.0000	1.0000	0.9960
	CC-SIS	d_1	1.0000	0.9980	1.0000	1.0000	0.9980
		d_2	1.0000	0.9980	1.0000	1.0000	0.9980
		d_3	1.0000	0.9980	1.0000	1.0000	0.9980
	NIS	d_1	0.9860	0.8120	1.0000	0.9920	0.7940
		d_2	0.9980	0.8860	1.0000	0.9980	0.8840
		d_3	1.0000	0.9260	1.0000	1.0000	0.9260
	SIS	d_1	1.0000	0.9980	1.0000	0.4160	0.4140
		d_2	1.0000	1.0000	1.0000	0.5340	0.5340
		d_3	1.0000	1.0000	1.0000	0.6180	0.6180
	DC-SIS	d_1	1.0000	0.9900	1.0000	0.4100	0.4060
		d_2	1.0000	0.9960	1.0000	0.5260	0.5240
		d_3	1.0000	0.9980	1.0000	0.5980	0.5960
	DC-RoSIS	d_1	1.0000	0.9800	1.0000	0.3840	0.3780
		d_2	1.0000	0.9920	1.0000	0.5080	0.5060
		d_3	1.0000	0.9940	1.0000	0.5760	0.5740

Figure 1: **Boxplot of minimum model size**



(a) $t_1 = t_2 = 0$