# Supplement to “Composite Coefficient of Determination and Its Application in Ultrahigh Dimensional Variable Screening”

EFANG KONG, YINGCUN XIA AND WEI ZHONG

August 17, 2018

## 1 Proofs of results in the paper

**Proof of Lemma 1**. (1) The ‘if’ part is obvious. For the ‘only if’ part, note that for any given $v$,

$$
\begin{aligned}
C_{V|U}(v) &= E\{I_V(v) - F_{V|U}(v|U)\}^2 = E_U\Big\{E\left[\{I_V(v) - F_{V|U}(v|U)\}^2|U\right]\Big\} \\
&= E\left[\mathrm{Var}\{I(V \le v)|U\}\right] \le \mathrm{Var}[I(V \le v)] = F_V(v)(1 - F_V(v)) \\
&= E\left[\mathrm{Var}\{I(V \le v)|U\}\right] + \mathrm{Var}[F_{V|U}(v|U)].
\end{aligned}
$$

Therefore, if $\mathrm{CCD}(U, V) = 0$, then we must have $\mathrm{Var}[F_{V|U}(v|U)] = 0$, for nearly all $v$. In other words, random variable $F_{V|U}(v|U)(= \Pr(V \le v|U))$ is independent of $U$, and this observation holds true for nearly all $v$. Thus $V$ and $U$ must be independent.

(2) The ‘if’ part again is obvious. For the ‘only if’ part, note that $\mathrm{CCD}(U, V) = 1 \iff C_{V|U}(v) = C_{U|V}(u) = 0$ for nearly all $u$ and $v$. In other words, that $I_V(v) = E(I_V(v)|U)(\equiv F_{V|U}(v|U))$ holds for nearly all $v$. Since $E(I_V(v)|U)$ is $\sigma(U)-$measurable, so is $I_V(v)$ and hence a function of $U$. Since such observation holds true for all $v$, it could be concluded that $V$ is a function of $U$. The same line of arguments could be applied to $C_{U|V}(u) = 0$, to deduce that $U$ is also a function of $V$. This completes the proof.

(3) Trivial.

(4) Without loss of generality, suppose $M(.)$ and $N(.)$ are both monotone increasing. Let $U'$ and $V'$ are independent copies of $U$ and $V$, respectively. Write $X = M(U)$, $Y = N(V)$, $X' = M(U')$, and $Y' = N(V')$, so that

$$
\begin{aligned}
&CCD(M(U), N(V)) = CCD(X, Y) = 1 - \frac{1}{2}\Big[\frac{E\{C_{Y|X}(Y)\}}{D_Y} + \frac{E\{C_{X|Y}(X)\}}{D_X}\Big] \\
&E\{C_{Y|X}(Y)\} = E_{Y'}\Big[E_{Y,X}\{I(Y \le Y') - F_{Y|X}(Y'|X)\}^2\Big] \\
&\qquad\qquad\quad = E_{N'}\Big[E_{V,U}\{I(V \le V') - F_{V|U}(V'|U)\}^2\Big] = E\{C_{V|U}(V)\} \\
&F_Y(Y')(1 - F_Y(Y')) = F_V(V')(1 - F_V(V')).
\end{aligned}
$$

That $E\{C_{Y|X}(Y)\} = E\{C_{V|U}(V)\}$ and $D_V = D_Y$ could be proved in a similar manner.

(5) Let $\phi(.)$ and $f(t)$ be the probability distribution and density functions of $N(0, 1)$, respectively. When $(U, V)$ are bivariate Gaussian with correlation coefficient $\rho$, we have

$$
E\{C_{V|U}(V)\} = \frac{1}{2} - E\Big[\phi^2\Big(\frac{V - \rho U}{\sqrt{1-\rho^2}}\Big)\Big]
$$

where $U$ and $V$ are independent $N(0, 1)$ random variables. Equivalently, we have

$$
\begin{aligned}
E\{C_{V|U}(V)\} &= \frac{1}{2} - \frac{1}{\sqrt{2\pi(1+\rho^2)}}\int \phi^2\Big(\frac{t}{\sqrt{1-\rho^2}}\Big)\exp\Big[-\frac{t^2}{2(1+\rho^2)}\Big]dt \\
&= \frac{1}{2} - \int \phi^2(st) f(t) dt, \quad s \equiv \{(1+\rho^2)/(1-\rho^2)\}^{1/2} \\
&= \frac{1}{2} - \int_{t>0} \{\phi^2(st) + \phi^2(-st)\} f(t) dt
\end{aligned}
$$

where the first equality follows from the fact that $V - \rho U$ is a $N(0, 1 + \rho^2)$ random variable, and the second equality is a result of change of variables. Since $s$ is an increasing function of $|\rho|$, that $E\{C_{V|U}(V)\}$ is monotone decreasing in $|\rho|$, is equivalent to that

$$
\int_{t>0} \{\phi^2(st) + \phi^2(-st)\} f(t) dt \tag{S.1}
$$

is an increasing function of $s$. Seeing that the derivative of (S.1) with respect to $s$,

$$2\int_{t>0}\{t\phi(st)f(st)-t\phi(-st)f(-st)\}f(t)dt=2\int_{t>0}tf(st)\{\phi(st)-\phi(-st)\}f(t)dt,$$

is always positive since $s>0$, the proof is thus complete. $\square$

**Proof of Lemma 2** : The proof of the first assertion is trivial, while the second follows from the Kac's Theorem of characteristic functions. In fact $MCCD(\mathbf{u},\mathbf{v})_\alpha=0$ implies that, for any $\mathbf{a}\in R^p$, $\mathbf{b}\in R^q$, $\mathbf{a}^\top\mathbf{u}$ and $\mathbf{b}^\top\mathbf{v}$ are independent, which coud be expressed in terms of characteristic functions

$$E\{\exp[i(t_1\mathbf{a}^\top\mathbf{u}+t_2\mathbf{b}^\top\mathbf{v})]\}=E\{\exp(it_1\mathbf{a}^\top\mathbf{u})\}E\{\exp(it_2\mathbf{b}^\top\mathbf{v})\}.$$

Since this holds for any $t_1,t_2\in R$ and $\mathbf{a}\in R^p$, $\mathbf{b}\in R^q$, Kac's Theorem suggests that $\mathbf{u}$ and $\mathbf{u}$ are independent. $\square$

Next, we make some notations for ease of presentation. For (random) sequences $\{s_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, write $s_n=O(b_n)$ if $s_n/b_n\,(n\to\infty)$ is bounded; $s_n=O_p(b_n)$ means $s_n/b_n\,(n\to\infty)$ is bounded in probability; $s_n=O_{wp1}(b_n)$ means $s_n/b_n\,(n\to\infty)$ is bounded with probability one.

Let $\xi=(U,V)$ denote any given $(X_k,Y)$, $k=1,\cdots,p$, with domain $\mathcal{X}=\mathcal{X}_1\otimes\mathcal{X}_2$. Observations of $\xi=(U,V)$ are thus written as $\xi_i:=(U_i,V_i)$, $i=1,\cdots,n$. Denote the joint probability density function of $U$ and $V$ by $f(U,V)$, while their respective marginal density functions by $f_U(.)$ and $f_V(.)$. For any $(u,v)\in\mathcal{X}$, define

$$F(u,v)=f_U(u)F_{V|U}(v|u)=\int_{-\infty}^v f(u,V)dV.$$

Let $f_U^{(2)}(.)$ and $F_U^{(2)}(U,V)$ denote the second order derivatives (with respect to $U$) of $f_U(.)$ and $F(U,V)$, respectively. With $n\to\infty$, $h_n\to0$, let $\delta_n=(nh_n/\log n)^{-1/2}$, and with any $(u,v)\in\mathcal{X}$,

$$a_n(u)=\int K(t)f_U(u+h_nt)dt=f_U(u)+\frac{1}{2}h_n^2f_U^{(2)}(u)+O(h_n^4),$$

$$c_n(u,v)=\int_{-\infty}^v\Big\{\int_{-\infty}^\infty K(t)f_U(u+h_nt,V)dt\Big\}dV=F(u,v)+\frac{1}{2}h_n^2F_U^{(2)}(u,v)+O(h_n^4).$$

Some standard results from empirical processes (e.g., Pollard, 1984) are such that

$$T_{1n}(v) \triangleq \frac{1}{n}\sum_{i=1}^{n} I(V_i < v) - F_V(v) = O_{wp1}((\log n/n)^{1/2}),$$
$$T_{2n}(u) \triangleq \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(U_i - u) - a_n(u) = O_{wp1}(\delta_n),$$
$$T_{3n}(u,v) \triangleq \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(U_i - u)I(V_i < v) - c_n(u,v) = O_{wp1}(\delta_n),$$

uniformly in $(u,v) \in \mathcal{X}$. In addition, we have

$$\{\hat{f}_U(u)\}^{-1} = \{f_U(u)\}^{-1} - \frac{1}{2}h_n^2\{f_U(u)\}^{-2}f_U^{(2)}(u) - \{f(u)\}^{-2}T_{2n}(u) + O_{wp1}(h_n^4 + \delta_n^2)$$
$$\hat{F}(v|u) = F_{V|U}(v|u) + \frac{1}{2}h_n^2 a(u,v) + \{f_U(u)\}^{-1}\frac{1}{n}\sum_{i=1}^{n} g_n(\xi_i, u, v) + O_{wp1}(h_n^4 + \delta_n^2) \quad \text{(S.2)}$$

uniformly in $(u,v) \in \mathcal{X}$, where

$$a(u,v) = \{f_U(u)\}^{-1}\{F_U^{(2)}(u,v) - F_{V|U}(v|u)f_U^{(2)}(u)\}$$
$$g_n(\xi_i, u, v) = K_{h_n}(U_i - u)\{I(V_i < v) - F_{V|U}(v|u)\} - \{c_n(u,v) - a_n(u)F_{V|U}(v|u)\}.$$

**Proof of Theorem 1.** For root-$n$ consistency, it suffices to show that

$$\frac{1}{n}\sum_{k=1}^{n} \hat{C}_{V|U}(V_k) = E\{C_{V|U}(V)\} + O_p(n^{-1/2}), \quad \hat{D}_V = D_V + O_p(n^{-1/2}), \quad \text{(S.3)}$$
$$\frac{1}{n}\sum_{k=1}^{n} \hat{C}_{U|V}(U_k) = E\{C_{U|V}(V)\} + O_p(n^{-1/2}), \quad \hat{D}_U = D_U + O_p(n^{-1/2}).$$

We only prove (S.3) to illustrate. To begin with, note that

$$\hat{D}_V = \frac{1}{n}\sum_{k=1}^{n} F_V(V_k)\{1 - F_V(V_k)\} + D_n + O_{wp1}(\log n/n), \quad \text{(S.4)}$$

where

$$D_n \triangleq \frac{1}{n^2}\sum_{i,k=1}^{n} \{1 - 2F_V(V_k)\}\{I(V_i < V_k) - F_V(V_k)\},$$

a $U$-statistics of zero mean with a finite variance. Root-$n$ consistency of $\hat{D}_V$ thus holds. We now move on to prove the first identity in (S.3). In view of (S.2), we have

$$\begin{aligned}\hat{C}_{V|U}(v) &= \frac{1}{n}\sum_{i=1}^{n}\{\hat{F}(v|U_i) - I(V_i < v)\}^2 = h_n^2\frac{1}{n}\sum_{i=1}^{n} a(U_i, v)\{F(v|U_i) - I(V_i < v)\} \\ &\quad + \frac{2}{n^2}\sum_{i<j}\Phi(\xi_i, \xi_j|v) + \frac{1}{n}\sum_{i=1}^{n}\{F_{V|U}(v|U_i) - I(V_i < v)\}^2 + O_{wp1}(h_n^4 + \delta_n^2), \quad \text{(S.5)}\end{aligned}$$

where

$$\begin{aligned}\Phi(\xi_i, \xi_j|v) = \{f_U(U_i)\}^{-1}\{F_{V|U}(v|U_i) - I(V_i < v)\}g_n(\xi_j, U_i, v) \\ +\{f_U(U_j)\}^{-1}\{F_{V|U}(v|U_j) - I(V_j < v)\}g_n(\xi_i, U_j, v),\end{aligned}$$

which is both symmetric and degenerate, since

$$E_{\xi_i}\{\Phi(\xi_i, \xi_j|v)\} = E\Big[\{f_U(U_i)\}^{-1}\{F_{V|U}(v|U_i) - I(V_i < v)\}E_{\xi_i}\{g_n(\xi_j, U_i, v)\}\Big] = 0.$$

Define $\mathcal{F}_n \triangleq \{\Phi(\xi_i, \xi_j|v) : v \in [0, 1]\}$, a class of real-valued symmetric functions on $\mathcal{X}^{\otimes 2}$. The fact that $\Phi(., .|v)$ is formed through multiplication and addition between functions (indexed by $v$) such as $I(. \le v)$, $F_{V|U}(v|.)$, and $c_n(., v)$ (the latter two are both monotone in $v$), which together with results such as Lemma 2.6.18 (van der Varrt and Wellner, 1996, pp. 147) and Problem 3 of van der Vaart and Wellner (1996) (pp. 165) renders $\mathcal{F}$ as a polynomial (Vapnik-Chervonenkis) class. Theorem 9 of Nolan and Pollard (1987) are thus all met with $W(n, x) = (1 + nx^{\iota})^{-1}$, where $\iota > 0$ such that $nh_n^{\iota/2} < \infty$. This leads to

$$\sup_v \frac{1}{n^2}\Big|\sum_{i<j}\Phi(\xi_i, \xi_j|v)\Big| = o_{wp1}((nh_n)^{-1}) \quad \text{(S.6)}$$

As for the first term on the RHS of (S.5), we only need to note that

$$\begin{aligned}E\Big[\sum_{i,k} a(U_i, V_k)\{F_{V|U}(V_k|U_i) - I(V_i < V_k)\}\Big]^2 = O(n^3) \\ h_n^2\frac{1}{n}\sum_{i,k} a(U_i, V_k)\{F_{V|U}(V_k|U_i) - I(V_i < V_k)\} = O_p(h_n^2 n^{-1/2}). \quad \text{(S.7)}\end{aligned}$$

We could then conclude from (S.5), (S.6) and (S.7) that

$$\frac{1}{n}\sum_{k=1}^{n}\hat{C}_{V|U}(V_k)=\frac{1}{n^2}\sum_{i,k}\{F_{V|U}(V_k|U_i)-I(V_i<V_k)\}^2+o_p(n^{-1/2});\qquad (S.8)$$

the corresponding $U-$statistics with a bounded 'kernel'

$$\begin{aligned} h(\xi_i,\xi_k) &\triangleq \{F_{V|U}(V_k|U_i)-I(V_i<V_k)\}^2+\{F_{V|U}(V_i|U_k)-I(V_k<V_i)\}^2-E\{C_{U|V}(V)\} \\ &= 1+E\{C_{U|V}(V)\}-2I(V_i<V_k)-2F_{V|U}(V_i|U_k)I(V_k<V_i) \qquad (S.9) \\ &\quad +F^2_{V|U}(V_k|U_i)+F^2_{V|U}(V_i|U_k) \end{aligned}$$

is in general not degenerate, i.e. $\mathrm{Var}\{E[h(\xi_i,\xi_k)|\xi_i]\}\neq 0$. Therefore, according to Theorem A in Serfling (1980) (pp. 192), $n^{-3/2}\sum_{i<k}h(\xi_i,\xi_k)$ is asymptotically normal with zero mean and a positive variance. This together with (S.8) implies that

$$\frac{1}{n}\sum_{k=1}^{n}\hat{C}_{V|U}(V_k)=E\{C_{U|V}(V)\}+O_p(n^{-1/2})$$

and is also asymptotically normal. Since we already have $\hat{D}_V\rightarrow D_V(\equiv 1/6)$ in probability, the root-$n$ consistency and asymptotic normality of $\frac{1}{n}\sum_{k=1}^{n}\hat{C}_{V|U}(V_k)/\hat{D}_V$ is obvious.

We now study the asymptotics related to when $U$ and $V$ are independent. First of all, consider an asymptotic expansion of $\hat{F}(v|u)$ similar to (S.2) but of higher order; since $F(u,v)=f_U(u)F_{V|U}(v|u)=f_U(u)F_V(v|u)$ and $c_n(u,v)=a_n(u)F_{V|U}(v|u)$, the bias term of order $O(h_n^2)$ vanishes and the stochastic terms also admit simpler form and as a result

$$\begin{aligned} \hat{F}(v|u) &= F_V(v)+\{f_U(u)\}^{-1}\frac{1}{n}\sum_{i=1}^{n}K_{h_n}(U_i-u)\{I(V_i<v)-F_V(v)\} \\ &\quad -\frac{1}{2}h_n^2\{f_U(u)\}^{-2}\frac{1}{n}\sum_{i=1}^{n}K_{h_n}(U_i-u)\{I(V_i<v)-F_V(v)\} \\ &\quad -\frac{1}{2}h_n^2\{f_U(u)\}^{-2}\frac{1}{n}\sum_{i\neq j}K_{h_n}(U_i-u)\{I(V_i<v)-F_V(v)\}\{K_{h_n}(U_j-u)-a_n(u)\} \\ &\quad +O_{wp1}(h_n^4\delta_n+h_n^2\delta_n^2+\delta_n^3), \end{aligned}$$

so that

$$
\begin{aligned}
\hat{C}_{V|U}(v) &= \frac{1}{n}\sum_{i=1}^{n}\{\hat{F}_{V|U}(v|U_i) - I(V_i < v)\}^2 = \frac{1}{n}\sum_{i=1}^{n}\{F_V(v) - I(V_i < v)\}^2 \\
&+\frac{2}{n^2}\sum_{i,j}\Phi_1(\xi_i,\xi_j|v) + \frac{h_n^2}{n^2}\sum_{i,j}\{f_U(U_i)\}^{-1}\Phi_1(\xi_i,\xi_j|v) \qquad \text{(S.10)} \\
&+\frac{1}{n^3}\sum_{i,j,k}\Phi_2(\xi_i,\xi_j,\xi_k|v) + \frac{2}{n^3}\sum_{i,j,k}\Phi_3(\xi_i,\xi_j,\xi_k|v) + O_{wp1}(h_n^4\delta_n + h_n^2\delta_n^2 + \delta_n^3),
\end{aligned}
$$

where

$$
\begin{aligned}
\Phi_1(\xi_i,\xi_j|v) &= \{f_U(U_i)\}^{-1}K_{h_n}(U_i - U_j)\{F_V(v) - I(V_i < v)\}\{F_V(v) - I(V_j < v)\}, \\
\Phi_2(\xi_i,\xi_j,\xi_k|v) &= \{f_U(U_i)\}^{-2}K_{h_n}(U_j - U_i)K_{h_n}(U_k - U_i) \\
&\qquad \{F_V(v) - I(V_j < v)\}\{F_V(v) - I(V_k < v)\}, \\
\Phi_3(\xi_i,\xi_j,\xi_k|v) &= \{f_U(U_i)\}^{-2}K_{h_n}(U_j - U_i)\{K_{h_n}(U_k - U_i) - a_n(U_i)\} \\
&\qquad \{F_V(v) - I(V_j < v)\}\{F_V(v) - I(V_i < v)\}.
\end{aligned}
$$

For the first term on the RHS of (S.10), consider

$$
\begin{aligned}
&h(V_i,V_k) \triangleq \{I(V_i < V_k) - F_V(V_k)\}^2 + \{I(V_k < V_i) - F_V(V_i)\}^2 - \frac{1}{3} \qquad \text{(S.11)} \\
&= \frac{2}{3} + F_V^2(V_i) + F_V^2(V_k) - 2I(V_i < V_k)F_V(V_k) - 2I(V_k < V_i)F_V(V_i), \\
&E\{h(\xi_i,\xi_k)|\xi_i\} = \frac{2}{3} + F_V^2(V_i) + \frac{1}{3} - 2F_V^2(V_i) - \{1 - F_V^2(V_i)\} = 0.
\end{aligned}
$$

This, according to Theorem 5.5.2 of Serfling (1980) (pp. 194), implies that

$$
\frac{1}{n}\sum_{i\geq k}h(V_i,V_k) \xrightarrow{d} \begin{array}{c}\text{a weighted sum of independent and} \\ \text{centered } \chi_1^2 \text{ random variables}\end{array} \qquad \text{(S.12)}
$$

To quantify the three terms in (S.10) involving either $\Phi_1(.,.|.)$ or $\Phi_3(.,.,.|.)$, we make use of the following facts

$$
\begin{aligned}
&E\{\Phi_1(\xi_i,\xi_j|V_k)|\xi_i\} = E\{\Phi_1(\xi_i,\xi_j|V_k)|\xi_j\} = E\{\Phi_1(\xi_i,\xi_j|V_k)|\xi_k\} = 0 \\
&E\{\Phi_3(\xi_i,\xi_j,\xi_k|V_l)|\xi_i,\xi_j\} = E\{\Phi_3(\xi_i,\xi_j,\xi_k|V_l)|\xi_i,\xi_k\} = 0; \\
&E\{\Phi_3(\xi_i,\xi_j,\xi_k|V_l)|\xi_i,\xi_l\} = E\{\Phi_3(\xi_i,\xi_j,\xi_k|V_l)|\xi_j,\xi_k\} = 0; \\
&E\{\Phi_3(\xi_i,\xi_j,\xi_k|V_l)|\xi_j,\xi_l\} = E\{\Phi_3(\xi_i,\xi_j,\xi_k|V_l)|\xi_k,\xi_l\} = 0;
\end{aligned}
$$

whence according to Serfling (1980) (Sec 5.3.4) we have

$$\frac{1}{n^3}\sum_{i,j,k}\Phi_1(\xi_i,\xi_j|V_k) = O_p(n^{-1}h_n^{-1/2}), \tag{S.13}$$

$$\frac{1}{n^4}\sum_{i,j,k,l}\Phi_3(\xi_i,\xi_j,\xi_k|V_l) = O_p(n^{-2/3}h_n^{-1}). \tag{S.14}$$

It remains to deal with the term which involves $\Phi_2(.,.,.|.)$. First note that

$$E[\{f_U(U_i)\}^{-2}K_h^2(U_j-U_i)\{F_V(V_k)-I(V_j<V_k)\}^2] = \frac{1}{6}h_n^{-1}R(K)(1+o(1)). \tag{S.15}$$

Secondly, for any distinct quadruplet $1 \le i,j,k,l \le n$, let $\{(i,j,k,l)\}$ denote the set of all possible permutations of $(i,j,k,l)$; and consequently define

$$\tilde{\Phi}_2(\xi_i,\xi_j,\xi_k|V_l) = \sum_{(\tilde{i},\tilde{j},\tilde{k},\tilde{l})\in\{(i,j,k,l)\}}\Phi_2(\xi_{\tilde{i}},\xi_{\tilde{j}},\xi_{\tilde{k}}|V_{\tilde{l}}).$$

Write $U_{ij} = U_i - U_j$, $U_{ik} = U_i - U_k$, $\tilde{V} = F_V(V)$, a uniform [0,1] random variable; since $E\{\Phi_2(\xi_i,\xi_j,\xi_k|V_l)|\xi_i\} = E\{\Phi_2(\xi_i,\xi_j,\xi_k|V_l)|\xi_j\} = E\{\Phi_2(\xi_i,\xi_j,\xi_k|V_l)|\xi_k\} = E\{\Phi_2(\xi_i,\xi_j,\xi_k|V_l)|\xi_l\} = 0$, we have

$$\tilde{h}_n(\xi_j,\xi_k) \triangleq E\{\tilde{\Phi}_2(\xi_i,\xi_j,\xi_k|V_l)|\xi_i,\xi_j\} = \{f_U(U_k)\}^{-1}H_n(U_j,U_k)C_1(\tilde{V}_j,\tilde{V}_k),$$

where $C_1(\tilde{V}_j,\tilde{V}_k) \triangleq E_L[\{\tilde{V}_l - I(\tilde{V}_j<\tilde{V}_l)\}\{\tilde{V}_l - I(\tilde{V}_k<\tilde{V}_l)\}]$ and

$$\begin{aligned} H_n(U_j,U_k) &= \int f_U(U_k)[f_U(U_k+h_nt)]^{-1}K_{h_n}(U_{jk}+h_nt)K(t)dt \\ &= \int K_{h_n}(U_{jk}+h_nt)K(t)dt(1+O(h_n)) \end{aligned} \tag{S.16}$$

For any real valued square-integrable function $g(.) \in L_2(\mathcal{X}, F_U(.)\otimes F_V(.))$, consider the Hilbert-Schmidt operator defined as

$$E_{\xi_k}[\tilde{h}_n(\xi,\xi_k)g(\xi_k)|\xi] \equiv \int_{\mathcal{X}} H_n(U,U_k)g(U_k,V_k)C_1(\tilde{V},\tilde{V}_k)dU_kd\tilde{V}_k,$$

where the inequality follows because $U_k$ and $V_k$ are independent. As the form of function $C_1(.,.)$ is independent of the original distribution $F_V(.)$, the solutions $\lambda$

(eigenvalues) to the following characteristic equation in terms of real valued square-integrable function $g(.) \in L_2(\mathcal{X}, F_U(.) * F_V(.))$:

$$E_{\xi_k}[\tilde{h}_n(\xi, \xi_k) g(\xi_k) | \xi] = \lambda g(\xi)$$

are identical to those associated with the following equation in terms of real valued square-integrable function $g(.) \in L_2(\mathcal{X}_1 \otimes [0,1], F_U(.) \otimes Uniform)$

$$\int_{\mathcal{X}_1 \otimes [0,1]} H_n(U, U_k) g(U_k, V_k) C_1(V, V_k) dU_k dV_k = \lambda g(U, V). \tag{S.17}$$

Write the limit of $h_n H_n(U, U_k)$ as $H(U, U_k)$ which equals $R(K)$ if $U = U_k$ and zero, otherwise. Since the Hilbert-Schmidt norm of the operator of (S.17) is of order $O(h_n^{-1})$, we apply results from the perturbation theory for linear operators Kato (1966) (Chapter 5, Sec 4), to conclude that (similar to Theorem 5.5.2 of Serfling (1980), pp. 194),

$$\frac{h_n}{n^3} \sum_{i,j \neq k,l} \Phi_2(\xi_i, \xi_j, \xi_k | V_l) \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j (\chi_{1j}^2 - 1), \tag{S.18}$$

where $\chi_{1j}^2$, $j = 1, \cdots$, are independent $\chi_1^2$ random variables, while $\lambda_j$, $j = 1, \cdots$, are real numbers (not necessarily distinct) corresponding to distinct solutions to the following equation

$$\int_{\mathcal{X}_1 \otimes [0,1]} H(U, U_k) C_1(V, V_k) g(U_k, V_k) C_1(V, V_k) dU_k dV_k = \lambda g(U, V); \tag{S.19}$$

as neither $H(U, U_k)$ nor $C_1(V, V_k)$ depends on $F_U(.)$ or $F_V(.)$, the same can be said about the limiting distribution in (S.18).

The same conclusion applies to the limiting distribution of the overall statistic $\widehat{CCD}(U, V)$, for exactly the same technicalities could be used to the following somewhat 'expanded' version of the U-statistic $\tilde{\Phi}_2(\xi_i, \xi_j, \xi_k | V_l)$ discussed above, constructed through the 'symmetrization' of the following:

$$\begin{aligned} &\{f_U(U_i)\}^{-2} K_{h_n}(U_j - U_i) K_{h_n}(U_k - U_i) \{F_V(V_l) - I(V_j < V_l)\} \{F_V(V_l) - I(V_k < V_l)\} \\ &+ \{f_V(V_i)\}^{-2} K_{h_n}(V_j - V_i) K_{h_n}(V_k - V_i) \{F_U(U_l) - I(U_j < U_l)\} \{F_U(U_l) - I(U_k < U_l)\}. \end{aligned}$$

Conclusions in the form of (S.18) still apply; this together with the fact that $\hat{D}_V \to 1/6$ and $\hat{D}_U \to 1/6$ in probability, completes the proof. $\square$

**Proof of Equation (2.9).**

$$B_n = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{i=1}^{n}\{\hat{F}_{V|U}(V_k|U_i) - \hat{F}_V(V_k)\}^2$$

In view of (S.2), we have

$$\begin{aligned} B_n = & \frac{1}{n^2}\sum_{i,k}\{F(V_k|U_i) - F_V(V_k)\}^2 + h_n^2\frac{1}{n^2}\sum_{i,k} a(U_i, V_k)\{F(V_k|U_i) - F_V(V_k)\} \\ & + \frac{2}{n^3}\sum_{i,j,k}\{f(U_i)\}^{-1}\{F(V_k|U_i) - F_V(V_k)\}g_n(\xi_j, U_i, V_k) \\ & + \frac{2}{n^3}\sum_{i,k}\{F(V_k|U_i) - F_V(V_k)\}T_{1n}(V_k) + O_{wp1}(h_n^4 + \delta_n^2). \end{aligned} \tag{S.20}$$

Regarding the terms on the RHS of (S.20), first note that the second term is of order $O_p(h_n^2)$, since $E[a(U_i, V_k)\{F(V_k|U_i) - F_V(V_k)\}] \neq 0$. The third term, as a result of the Law of Iterated logarithm for (zero-mean) U-statistics, is of order $O_{wp1}((nh_n/\log\log n)^{-1/2})$. Note that we do not have stronger results like (S.6) for this term, since that only holds for degenerate U-statistics (processes). $\square$

To prepare the ground for the proof of Theorem 2, we need to introduce the following class of functions and make use of results concerning bounds on its entropy (covering) numbers. For $s > 0$, the class of all functions on a bounded set $\mathcal{D}$ in $R^d$ that possess uniformly bounded partial derivatives up to order $[s]$ (the greatest integer smaller than $s$) and whose highest partial derivatives are Lipschitz of order $s - [s]$. Specifically, define for any vector $k = (k_1, \cdots, k_d)$ of $d$ integers, the differential operator

$$D^k = \frac{\partial^{|k|}}{\partial_{x_1}^{k_1}\cdots\partial_{x_d}^{k_d}}, \quad |k| = \sum k_i.$$

For a function $g : \mathcal{D} \mapsto R$, let

$$\|g\|_s = \max_{|k|\le[s]} \sup_{\xi} |D^k g(\xi)| + \max_{|k|=[s]} \sup_{\xi,\xi'} \frac{|D^k g(\xi) - D^k g(\xi')|}{\|\xi - \xi'\|^{s-[s]}}$$

with the suprema taken over all $\xi, \xi'$ in the interior of $\mathcal{D}$ with $\xi \neq \xi'$. For $s$, $M > 0$, let $C_M^s(\mathcal{D})$ be the set of all continuous functions $g : \mathcal{D} \mapsto R$ with $\|g\|_s \le M$.

**Proof of Theorem 2**. It suffices to show that there exist $c_1$ and $k_1 > 0$, such that for $k = 1, \cdots, p$, and $U = X_k$, $V = Y$,

$$\Pr(|\widehat{CCD}(U,V) - CCD(U,V)| > cn^{-\tau}) \le c_1 \exp(-k_1 n^{1-2\tau}), \tag{S.21}$$

To this aim, we make use of the following identity

$$\begin{aligned} g(\xi_i|v,\hat{F}) \triangleq |\hat{F}(v|U_i) - I(V_i < v)|^2 = I(V_i < v) \\ \{1 - \hat{F}(v|U_i)\}^2 + [1 - I(V_i < v)]\{\hat{F}(v|U_i)\}^2. \end{aligned}$$

Under (A2), we conclude that there exists some $M > 0$, such that with probability one, $\{\hat{F}(v|.) : v \in \mathcal{X}_2\} \subset C_M^1(\mathcal{X}_1)$. So are $\{1 - \hat{F}(v|.)\}^2$ and $\{\hat{F}(v|.)\}^2$. On the other hand, $\{I(V \le v) : v \in \mathcal{X}_2\}$ is a polynomial class. Thus, according to Theorem 2.7.1 and Lemma 2.6.18 of van der Vaart and Wellner (1996) (pp. 147; pp. 155), the entropy numbers of $\mathcal{F}_2 \triangleq \{|\widetilde{F}(v|U_i) - I(V_i < v)|^2 : v \in \mathcal{X}_2, \widetilde{F}(.|v) \in C_M^1(\mathcal{X}_1)\}$ satisfy condition (2.14.8) of van der Vaart and Wellner (1996) (pp. 246). Consequently we could apply their Theorem 2.14.10 with the constants therein $\delta = 0$, $W = 1$, to conclude that there exist constants $c_2, k_2 > 0$, such that such that for every $t > 0$,

$$\Pr\Bigg( \sup_{\substack{v \,\in\, [0,1] \\ g(.|.,.) \,\in\, \mathcal{F}_2}} \Bigg| \frac{1}{n} \sum_{i=1}^{n} g(\xi_i|v,\widetilde{F}) - E\{g(\xi_i|v,\widetilde{F})\} \Bigg| > t \Bigg) \le c_2 \exp\Big(k_2 n^{5/6} t^{5/3} - 2nt^2\Big). \tag{S.22}$$

Next, note that for any given $v \in \mathcal{X}_2$, $\widetilde{F}(.|v) \in C_M^1(\mathcal{X}_1)$,

$$E\{g(\xi_i|v,\widetilde{F})\} = E_{U_i}[F(v|U_i)\{1 - F(v|U_i)\}] + E_{U_i}[\{\widetilde{F}(v|U_i) - F(v|U_i)\}^2]. \tag{S.23}$$

This should be used in conjunction with the identity that under (A3)

$$\hat{F}(v|U) - F(v|U) = O_{wp1}(h_n^2 + (nh_n/\log n)^{-1/2}) = o_{wp1}(n^{-\tau}) \tag{S.24}$$

uniformly in $v$ and $U$. On the other hand, for any $t > 0$,

$$\Pr\Big(\Big|\frac{1}{n}\sum_{k=1}^{n} E_U[F(V_k|U)\{1-F(V_k|U)\}] - E\{C_{U|V}(U)\}\Big| \ge t\Big) \le \exp\Big(-\frac{16nt^2}{1+8t/3}\Big). \tag{S.25}$$

Together, assumption (A3), (S.22), (S.23), (S.24) and (S.25) imply that there exist constants $c_2, k_2 > 0$ (depending on $c$) , so that for every $\tau \in [0, 1/2)$,

$$\Pr\Big(\Big|\frac{1}{n}\sum_{k=1}^{n} \hat{C}_{V|U}(V_k) - E\{C_{U|V}(U)\}\Big| \ge cn^{-\tau}\Big) \le c_2\exp(-k_2 n^{1-2\tau}). \tag{S.26}$$

As for the denominator $\hat{D}_V$, we make use of (S.4), and the following Bernstein type of inequalities for independent sums as well as U-statistics: that for any $t > 0$,

$$\Pr\Big(\Big|\frac{1}{n}\sum_{k=1}^{n} P_V(V_k)\{1-P_V(V_k)\} - D_V]\Big| \ge t\Big) \le \exp\Big[-\frac{nt^2}{1/16+t/6}\Big]; \tag{S.27}$$

$$\Pr(|D_n| \ge t) \le \exp\Big[-\frac{nt^2}{1/16+t/6}\Big]. \tag{S.28}$$

The Bernstein type of inequality in the form of

$$\Pr\Big(\Big|\frac{\sum_{k=1}^{n}\hat{C}_{V|U}(V_k)}{\hat{D}_V} - \frac{E\{C_{U|V}(U)\}}{D_V}\Big| \ge cn^{-\tau}\Big) \le c_2\exp(-k_2 n^{1-2\tau})$$

for some $c_2$, $k_2 > 0$ just follows from (S.26), (S.27) and (S.28). □

**Proof of Theorem 3.** The proof of Theorem 3 shares the similar spirit of Theorem 2 in Fan, Feng and Song (2011) and Theorem 3 in Zhou and Zhu (2018). For any constant $c > 0$, the cardinality of the set $\{k : |\omega_k > cn^{-\tau}|, 1 \le k \le p\}$ is bounded by $O\left(n^{\tau}\sum_{k}^{p}|\omega_k|\right)$. In view of Theorem 2, i.e.

$$\Pr\left(\max_{1\le k\le p}|\hat{\omega}_k - \omega_k| > cn^{-\tau}\right) \le O\Big(p\exp(-an^{1-2\tau})\Big),$$

we know the event $\max_{1\le k\le p}|\hat{\omega}_k - \omega_k| \le cn^{-\tau}$ occurs with probability $1-O\Big(p\exp(-an^{1-2\tau})\Big)$. Consequently

$$\begin{aligned}\hat{\mathcal{D}} \overset{def}{=} \{k : \hat{\omega}_k > cn^{-\tau}\} &= \{k : \hat{\omega}_k > 2cn^{-\tau},\, |\hat{\omega}_k - \omega_k| > cn^{-\tau}\} \\ &\quad \cup \{k : \hat{\omega}_k > 2cn^{-\tau},\, |\hat{\omega}_k - \omega_k| < cn^{-\tau}\}\end{aligned}$$

would be identical to $\hat{\mathcal{D}}_1 = \{k : \hat{\omega}_k > 2cn^{-\tau}, |\hat{\omega}_k - \omega_k| < cn^{-\tau}\}$ with probability $1 - O\Big(p\exp(-an^{1-2\tau})\Big)$. Since $\hat{\mathcal{D}}_1 \subseteq \{k : \omega_k > cn^{-\tau}\}$ and the cardinality of the latter is of order $O\left(n^{\tau}\sum_k^p |\omega_k|\right)$, the proof is thus complete. $\square$

**Proof of Theorem 4**. To illustrate, we study the convergence rate concerning definition (4.1) with $\alpha > 0 (\neq 2)$; with (4.2), the notations will be more complicated, but the proof follows the same line of reasoning.

Let $\tau_n = h_n^2 + \delta_n$. We start with the Taylor expansion of $|\hat{F}(v|U_i) - I(V_i < v)|^{\alpha}$ at $|F(v|U_i) - I(V_i < v)|^{\alpha}$, and since $\hat{F}(v|U_i) - F(v|U_i) = O(\tau_n)$, our discussion is necessarily restricted to $(u, v) \in \mathcal{X}$, such that

$$F(v|u)/\tau_n \to \infty, \quad \text{and } 1 - F(v|u)/\tau_n \to \infty.$$

Noting (S.2), we have

$$\begin{aligned}|\hat{F}(v|U_i) - I(V_i < v)|^{\alpha} &= I(V_i < v)\{1 - \hat{F}(v|U_i)\}^{\alpha} + \{1 - I(V_i < v)\}\{\hat{F}(v|U_i)\}^{\alpha} \\ &= g_1(\xi_i|v,\alpha) + \frac{\alpha}{2}h_n^2 a(U_i, v) g_2(\xi_i|v,\alpha) + \alpha\frac{1}{n}\sum_{i=j}^{n} g_n(\xi_j, U_i, v) g_2(\xi_i|v,\alpha) + O(\tau_n^2),\end{aligned}$$

where $a(u, v)$ and $g_n(\xi_i, u, v)$ are as defined in (S.2),

$$\begin{aligned}&g_1(\xi_i|v,\alpha) \triangleq I(V_i < v)\{1 - F(v|U_i)\}^{\alpha} + [1 - I(V_i < v)]\{F(v|U_i)\}^{\alpha}, \\ &g_2(\xi_i|v,\alpha) \triangleq \{1 - I(V_i < v)\}\{F(v|U_i)\}^{\alpha-1} - I(V_i < v)\{1 - F(v|U_i)\}^{\alpha-1} \\ &a(u, v) = \{f(u)\}^{-1}\{F_U^{(2)}(u, v) - F_{V|U}(v|u) f^{(2)}(u)\}, \\ &g_n(\xi_i, u, v) = K_{h_n}(U_i - u)\{I(V_i < v) - F_{V|U}(v|u)\} - \{c_n(u, v) - a_n(u) F_{V|U}(v|u)\}.\end{aligned}$$

Therefore, with $\Phi(\xi_i, \xi_j|v,\alpha) = g_n(\xi_j, U_i, v) g_2(\xi_i|v,\alpha) + g_n(\xi_i, U_j, v) g_2(\xi_j|v,\alpha)$, we have

$$\begin{aligned}\hat{C}_{V|U}(v;\alpha) = \frac{1}{n}\sum_{i=1}^{n} |\hat{F}(v|U_i) - I(V_i < v)|^{\alpha} &= \frac{\alpha}{2n} h_n^2 \sum_{i=1}^{n} a(U_i, v) g_2(\xi_i|v,\alpha) \\ &+ \frac{\alpha}{n^2}\sum_{i\le j} \Phi(\xi_i, \xi_j|v,,\alpha) + \frac{1}{n}\sum_{i=1}^{n} g_1(\xi_i|v,\alpha) + O_{wp1}(h_n^4 + \delta_n^2). \quad \text{(S.29)}\end{aligned}$$

For the term of order $h_n^2$, standard results on U-statistics indicate that

$$\frac{1}{n^2}\sum_{i,k=1}^{n} a(U_i,V_k)g_2(\xi_i|V_k,\alpha) - E[a(U_i,V_k)g_2(\xi_i|V_k,\alpha)] = O_p(n^{-1/2}), \quad \text{(S.30)}$$

where the expectation term,

$$\begin{aligned} E[a(U_i,V_k)g_2(\xi_i|V_k,\alpha)] = E[a(U_i,V_k)\{(1-F(V_k|U_i))\{F(V_k|U_i)\}^{\alpha-1} \quad \text{(S.31)} \\ -F(V_k|U_i)\{1-F(V_k|U_i)\}^{\alpha-1}\}], \end{aligned}$$

however, is not zero unless $\alpha = 2$. This implies that $\hat{C}_{V|U}(v;\alpha)$ when taken summation over $v = V_k$, $k = 1, \cdots, n$, it would have a bias of order $O(h_n^2)$, with the coefficient non-diminishing. The rest of the terms in (S.29) could be dealt with in a similar manner as we have seen in the proof of Theorem 1. The situation stays the same Next, through arguments similar to those used in the proof of (S.6), we claim that $\mathcal{F}_1 = \{\Phi(\xi_i, \xi_j|v,,\alpha) : v \in \mathcal{X}_2\}$ forms a polynomial class, and based on Theorem 9 of Nolan and Pollard (1987) we have

$$\begin{aligned} &\sup_v \left|\frac{1}{n^2}\sum_{i\le j}\Phi(\xi_i,\xi_j|v,,\alpha) - \frac{1}{n}E_{\xi_i}[g_n(\xi_j,U_i,v)g_2(\xi_i|v,\alpha)]\right| = o_{wp1}(n^{-1}), \\ &\frac{1}{n^2}\sum_{j,k} E_{\xi_i}[g_n(\xi_j,U_i,V_k)g_2(\xi_i|V_k,\alpha)] = E[g_n(\xi_j,U_i,V_k)g_2(\xi_i|V_k,\alpha)] + O_p(n^{-1/2}), \\ &E[g_n(\xi_j,U_i,V_k)g_2(\xi_i|V_k,\alpha)] = 0. \end{aligned}$$

Together these lead to the conclusion that

$$\frac{1}{n^3}\sum_{i,j,k}\Phi(\xi_i,\xi_j|V_k,,\alpha) = O_p(n^{-1/2}). \quad \text{(S.32)}$$

As for the remaining term in (S.29), standard results on U-statistics

$$\frac{1}{n^2}\sum_{i,k}^{n} g_1(\xi_i|V_k,\alpha) = E\{C_{U|V}(V)\} + O_p(n^{-1/2}). \quad \text{(S.33)}$$

Together (S.29), (S.30), (S.32), (S.32) and (S.33) yield

$$n^{-1}\sum_k \hat{C}_{V|U}(V_k;\alpha) = E\{C_{V|U}(;\alpha)\} + O(h_n^2) + O_p(n^{-1/2});$$

which, with $nh_n^4 \to 0$, results in the root-$n$ consistency of the term on the LHS. $\square$

# 2 An Iterative CCD-Based Variable Screening

Nearly all marginal variable screening methods share a common inherent weakness, namely their failure to identify the important covariates which are marginally independent of the response variable due to strong correlations among the covariates. Some iterative procedure of variable screening could then be adopted to reduce the risk of missing the truly important covariates. As the CCD-based variable screening method is model-free, following Zhu, et al. (2011) and Zhong and Zhu (2015), we propose the following iterative procedure to improve screening performances. The essence of this procedure is the projection of the unselected covariates to the orthogonal space of the already selected variables.

Denote the original data by $(Y, \mathbf{X})$, where $Y = (Y_1, \cdots, Y_n)$ and $\mathbf{X}$ is the $n \times d$ design matrix with columns given by $X_{(k)}$, $k = 1, \cdots, d$. Write $\mathcal{S}_0 = \{1, 2, \cdots, p\}$. Let $d$ be a preferred model size and $\mathcal{S}_1$ denote the index set of the $p_1 (< d)$ covariates selected using CCD-SIS based on $(Y, \mathbf{X})$.

**Step 1.** Denote by $\mathbf{X}_1$, the $n \times p_1$ sub-matrix of $\mathbf{X}$ consisting of columns indexed by $\mathcal{S}_1$. Compute $\mathbf{X}_{new} = \left\{ \mathbf{I}_n - \mathbf{X}_1 \left( \mathbf{X}_1^\top \mathbf{X}_1 \right)^{-1} \mathbf{X}_1^\top \right\} \mathbf{X}_1^c$, where $\mathbf{X}_1^c$ is a $n \times (d - p_1)$ sub-matrix of $\mathbf{X}$ consisting of columns indexed by $\mathcal{S}_0 \backslash \mathcal{S}_1$. Note that the columns of $\mathbf{X}_{new}$ should keep their labels (indices) as with the original $\mathbf{X}$.

**Step 2.** Apply CCD-SIS on the new data $(Y, \mathbf{X}_{new})$ and denote the index set of the newly selected $p_2$ covariates as $\mathcal{S}_2$.

**Step 3.** Update $\mathcal{S}_1 = \mathcal{S}_1 \cup \mathcal{S}_2$ and repeat **Steps 1 - 3** until the size of the selected model equals to $d$.

Note that for the design matrix $X_1$, the splines base of each selected variable can also be used to remove the nonlinear correlation between variables.

# References


Fan, J., Feng, Y. and Song, R. (2011), "Nonparametric independence screening in sparse ultrahigh dimensional additive models," *Journal of the American Statistical Association*, **106**, 544–557.

Kato, T. (1995), *Perturbation theory for linear operators.* Springer-Verlag: New York.

Nolan, D. and Pollard, D. (1987), *U-processes: rates of convergence, The Annals of Statistics*, **15**, 780–799.

Serfling, R. (1980), *Approximation theorems of Mathematical Statistics.* John Wiley and Sons.

van der Vaart, A. and Wellner, J. (1996), *Weak convergence and empirical processes: with applications to statistics.* Springer-Verlag: New York.

Zhong, W. and Zhu, L. (2015), "An Iterative Approach to Distance Correlation Based Sure Independence Screening," *Journal of Statistical Computation and Simulation*, **85**, 2331-2345.

Zhou, Y. and Zhu, L. (2018), "Model-Free Feature Screening for Ultrahigh Dimensional Data through a Modified Blum-Kiefer-Rosenblatt Correlation," *Statistica Sinica,* to appear, doi:10.5705/ss.202016.0264.

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**, 1464–1475.