Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Composite Coefficient of Determination and Its Application in Ultrahigh Dimensional Variable Screening

Efang Kong, Yingcun Xia & Wei Zhong

To cite this article: Efang Kong, Yingcun Xia & Wei Zhong (2018): Composite Coefficient of Determination and Its Application in Ultrahigh Dimensional Variable Screening, Journal of the American Statistical Association, DOI: 10.1080/01621459.2018.1514305

To link to this article: https://doi.org/10.1080/01621459.2018.1514305

Composite Coefficient of Determination and Its Application in Ultrahigh Dimensional Variable Screening

Efang Kong, Yingcun Xia and Wei Zhong
University of Electronic Science and Technology of China,
National University of Singapore and Xiamen University

August 17, 2018

Abstract: In this paper, we propose to measure the dependence between two random variables through a composite coefficient of determination (CCD) of a set of nonparametric regressions. These regressions take consecutive binarizations of one variable as the response and the other variable as the predictor. The resulting measure is invariant to monotonic marginal variable transformation, rendering it robust against heavy-tailed distributions and outliers, and convenient for independent testing. Estimation of CCD could be done through kernel smoothing, with a consistency rate of root-n. CCD is a natural measure of the importance of

[1]Efang Kong is Professor of School of Mathematical Sciences,University of Electronic Science and Technology of China, China. Email: singeu@live.co.uk. Her research was supported by a grant of National Natural Science Foundation of China (NNSFC) 11771066. Yingcun Xia is Professor of Department of Statistics and Applied Probability, National University of Singapore, Singapore. Email: staxyc@nus.edu.sg. His research was supported by NNSFC grant 71371095, and MOE grant of Singapore: MOE2014-T2-1-072, and NUS AcRF grant R-155-000-193-114. Wei Zhong is Professor of Wang Yanan Institute for Studies in Economics (WISE), and Department of Statistics, School of Economics, Xiamen University, China. Email: wzhong@xmu.edu.cn. His research was supported by NNSFC grants 11671334, 11401497, University Distinguished Young Researchers Program in Fujian Province and the Fundamental Research Funds for the Central Universities 20720181004. All authors equally contribute to this paper, and the authors are listed in the alphabetic order. The authors thank the Editor, the AE and three reviewers for their constructive comments, which have greatly improved the earlier version.

variables in regression and its sure screening property, when used for variable screening, is also established. Comprehensive simulation studies and real data analysis show that the newly proposed measure quite often turns out to be the most preferred compared to other existing methods both in independence testing and in variable screening.

Keywords: binarization, kernel regression, self-equitability, test of independence, ultrahigh dimensionality, variable screening.

1 Introduction

In data analysis and statistical inference, it is fundamental to quantify the strength of dependence between random variables. In numerous applications, Pearson correlation coefficient is widely used to measure linear relationships between two variables. In a univariate linear regression, the squared Pearson correlation coefficient between a response and an explanatory variable equals to the coefficient of determination R^2 which is the proportion of the variance in the response explained by linear regression on the explanatory variable. However, Pearson correlation coefficient fails to detect the nonlinear relationship between two variables. To overcome this limitation, many dependence measures have been introduced in the literature. Szekely, Rizzo and Bakirov (2007) proposed a distance correlation (DC) to measure both linear and nonlinear relationships. The appealing feature is that the distance correlation equals to zero if and only if two variables are independent. Shao and Zhang (2014) further extended the distance correlation to a martingale difference correlation (MDC) which measures the departure of conditional mean independence between two variables. Zheng, Shi and Zhang (2012) studied a pair of generalized measures of correlation (GMC) to account for asymmetries in the explained variance and nonlinearity be-

tween random variables. To measure the dependence between a continuous random variable and a categorical one, Cui, Li and Zhong (2015) developed a mean-variance (MV) index based on the Cramér-von Misses distances between conditional and unconditional distribution functions. Heller, Heller and Corfine (2013) considered a consistent test statistic for association between variables based on ranks of distances. Pfister, et al. (2016) investigated the dependence amongst a set of variables based on the Hilbert-Schmidt independence criterion (Gretton, et al., 2005).

An important application of dependence measures is to serve as marginal utilities to screen variables in ultrahigh dimensional problems. Marginal variable screening has received much attention in the recent ultrahigh dimensional analysis since the seminal work of Fan and Lv (2008) on sure independence screening (SIS). The SIS ranks predictors based on the magnitudes of Pearson correlation coefficients between the response and each predictor in Gaussian linear regression. It has been shown that under some mild conditions this procedure possesses a sure screening property, namely that all truly important predictors can be selected with probability tending to one even in ultrahigh dimensions. To deal with the limitation of Pearson correlation which only measures linear dependence, Li, Zhong and Zhu (2012) used the distance correlation (DC) of Szekely, Rizzo and Bakirov (2007) to rank the importance of the predictors, while Li et al. (2012) employed Kendall's τ rank correlation to guard against outliers. Shao and Zhang (2014) proposed a martingale difference correlation as marginal utility to identify variables which contribute to the conditional mean of the response. Cui, Li and Zhong (2015) developed a model-free feature screening method based on mean variance index for ultrahigh dimensional discriminant analysis. Zhou and Zhu (2018) proposed an independence screening method based on a modified Blum-Kiefer-Rosenblatt correlation (MBKR) (Blum, Kiefer and Rosenblatt, 1961), which considers the scaled difference between the joint cumulative distribution

function and the product of two marginal cumulative distribution functions. More recent variable screening approaches include Fan and Fan (2008), Fan and Song (2010), Hall and Miller (2009), Fan, Feng and Song (2011), Zhu, et al. (2011), Ji and Jin (2012), He, Wang and Hong (2013), Mai and Zou (2013), Liu, Li and Wu (2014) and Li, et al. (2016), among others.

Motivated by the importance of dependence measures as well as their application in ultrahigh dimensional variable screening, we propose in this paper a new metric, called the composite coefficient of determination (CCD), to measure the strength of the dependence between two random variables. CCD is calculated from the R^2 of a set of nonparametric regressions, and each regression takes consecutive binarization of one variable as response and the other variable as predictor. It enjoys several appealing properties. Firstly, CCD is zero only if the two variables are independent, so it is able to measure both linear and nonlinear relationships. Secondly, the measure is distribution-free and invariant to monotone transformations of variables. Thus, it is robust against distribution types and outliers. This is significant as most of the existing measures is not invariant to monotone transformations. Thirdly, the distribution of the estimated CCD for two independent variables is independent of their respective (marginal) distributions, and one implication of this is that permutation test could be carried out for independence testing. Although estimation of CCD is based on kernel smoothing, we could show that the estimator is consistent with a rate of at least root-n. Estimates of CCD when used for the purpose of variable screening in ultrahigh dimensional data for which no model structure is assumed, are shown to also enjoy the sure screening property and to deliver much improved performance in empirical studies than most of its competitors.

The rest of this article is organized as follows. In Section 2, we first give the definition of the composite coefficient of determination (CCD) and outline its theoretical

properties. We then discuss its estimation and establish its consistency rate together with the asymptotic distributions. Section 3 describes the sure independence screening procedure based on CCD and state the corresponding sure screening property. Some extensions of the concept of CCD, including the more general class of CCD_α and CCD for multivariate random variables are given in Section 4. Simulation studies and real data analysis are described in Section 5. All technical proofs are given in the online supplement.

2 Composite Coefficient of Determination

2.1 Definition of CCD

For any two random variables U and V, we consider a regression of the indicator function, $I_V(v) = I(V \leq v)$ for any given v, on U. The expectation conditional $I_V(v)$ on U is

$$E(I_V(v)|U) = F_{V|U}(v|U),$$

which is the conditional distribution function of V given U. Then, the expectation of the squared "regression error" of $I_V(v)$ on U is

$$C_{V|U}(v) \triangleq E\left[I_V(v) - E(I_V(v)|U)\right]^2 = F_V(v) - E\left[F_{V|U}(v|U)\right]^2, \tag{2.1}$$

where $F_V(v) = E(I_V(v)) = P(V \leq v)$ is the marginal distribution function of V. In cases where V is a deterministic function of U, $C_{V|U}(v) = 0$ for all v; on the other hand, if V and U are independent, then $E(I_V(v)|U) = E(I_V(v))$ and thus $C_{V|U}(v) = \text{Var}(I_V(v)) = F_V(v)(1 - F_V(v))$. Consequently, $E\{C_{V|U}(V)\}$ calibrates the difference between the marginal distribution function of V and its conditional counterpart given the value of U. Furthermore, $E\{C_{V|U}(V)\}$ is invariant to monotone transformation (Lemma 1), and it possesses the so-called self-equitability property proposed in Kinney and Atwal (2014): namely if $g(\cdot)$ is a deterministic function such

that $U \leftrightarrow g(U) \leftrightarrow V$ forms a Markov chain (to be understood as given $g(U)$, V is independent of U), then

$$E\{C_{V|U}(V)\} = E\{C_{V|g(U)}(V)\}. \tag{2.2}$$

In other words, different types of dependence are made comparable in terms of $E\{C_{V|U}(V)\}$. For example, if $V = U^2 + \epsilon$ and ϵ is independent of U, we would have $E\{C_{V|U}(V)\} = E\{C_{V|U^2}(V)\}$; this is consistent with the heuristic reasoning that in this case V depends on U to the same extent as it does on U^2. Next, we define

$$D_V \triangleq E\{\mathrm{Var}(I_V(V))\} = E\{F_V(V)(1 - F_V(V))\}. \tag{2.3}$$

Thus, $1 - E(C_{V|U}(V))/D_V$ could be used to measure the strength of dependence of V on U because it equals to zero when two variables are independent and becomes the unity when V is perfectly determined by U. From the perspective of regression, $C_{V|U}(v)$ is the residual sum of squares, and $1 - E(C_{V|U}(V))/D_V$ is a composite coefficient of determination of all the regressions with v going through all possible values of V. To obtain a measure which is symmetric about U and V, we exchange the roles of V and U, and obtain an analogue $1 - E(C_{U|V}(U))/D_U$ to assess the dependency of U on V. The average of the two is then used as an overall criterion to evaluate the strength of association between U and V:

$$CCD(U, V) \triangleq 1 - \frac{1}{2}\left\{\frac{E\{C_{V|U}(V)\}}{D_V} + \frac{E\{C_{U|V}(U)\}}{D_U}\right\}. \tag{2.4}$$

We call this measure the composite coefficient of determination (CCD) between U and V. As noted earlier, standing alone $E\{C_{V|U}(V)\}$ and $E\{C_{U|V}(U)\}$ are both self-equitable (see (2.2)), so as a combination of these two, $CCD(U, V)$ necessarily inherits this property, at least to some extent. Moreover, the averaging in (2.4) serves not only the purpose of producing a symmetric measure, but also has its

practical implications: when used for variable screening in regression models, we find that tests of independence based on CCD performs much better that those based on $1 - E\{C_{V|U}(V)\}/D_V$ or $1 - E\{C_{U|V}(U)\}/D_U$ alone.

REMARK 1. As implied by (2.3), $D_V \equiv 1/6$ not matter what distribution the random variable V follows, and $D_V (= \mathrm{Var}(I_V(\cdot)))$ is engaged in definition (2.4) so that CCD is 'standardized' to have a range of $[0, 1]$. One might also consider an alternative definition:

$$CCD_1(U, V) \triangleq 1 - \frac{1}{2}\left\{E\Big[\frac{C_{V|U}(V)}{\mathrm{Var}(I_V(V))}\Big] + E\Big[\frac{C_{U|V}(U)}{\mathrm{Var}(I_U(U))}\Big]\right\}; \tag{2.5}$$

compared with (2.4), (2.5) differs in the order in which ratio and expectation are taken. The range of this alternative definition is also $[0, 1]$, and it shares all the characteristics of CCD as summarized in Lemma 1. While CCD can be estimated with a consistency rate of at least root-n (Theorem 1), and enjoys the sure screening property (Theorem 2) when used for the purpose of variable screening, the same assertions can be made about CCD_1. The proof follows exactly the same line of reasoning as in the case of CCD, only with slightly more cumbersome notations.

REMARK 2. CCD is related to some existing measures in the literature. For example, since $C_{V|U}(v) = E\{\mathrm{Var}(I_V(v)|U)\}$, the quantity $C_{V|U}(v)/\{F_V(v)(1 - F_V(v))\}$ is in fact the coefficient of determination (CD) associated with regression $I_V(v)$ on U. Zheng, Shi and Zhang (2012) suggested the use of the CD associated with regression of V on U to measure the strength of dependency, which they call the generalized measures of correlation (GMC). As CCD is formed as an integration of these CD's associated with different v, it is more fine-tuned to detect the presence of dependency between U and V, than the relatively crude GMC measurement. Another example is the mean-variance index of Cui, Li and Zhong (2015), defined as $E[\mathrm{Var}\{F_V(V|U)\}]$ which equals $D_V - E\{C_{V|U}(V)\}$, and is used to measure the dependence between a

continuous variable V and a categorical variable U.

REMARK 3. Making use of (2.1), i.e.

$$E(C_{V|U}(V)) = \frac{1}{2} - \int F^2_{V|U}(v|u)dF_V(v)dF_U(u),$$

one might suspect some sort of equivalence between the CCD and a Kolmogorov-Smirnov type of statistic which, befitting the current context, is given by

$$\int [F_{V|U}(v|u) - F_V(v)]^2 dF_V(v)dF_U(u) = \frac{1}{3} - \int F^2_{V|U}(v|u)dF_V(v)dF_U(u). \quad (2.6)$$

We use $E\{C_{V|U}(V)\}$, rather than (2.6) in the formation of CCD, because it can be estimated with a consistency rate of at least $n^{-1/2}$ (to be described in Section 2.1), while its Kolmogorov-Smirnov-type alternative (2.6) could only be estimated at a standard nonparametric rate; see (2.9).

We now present some basic properties of CCD in the lemma below.

Lemma 1. *For any two random variables U and V,*

1. *$CCD(U,V) \geq 0$, where the equality holds only when U and V are independent.*
2. *$CCD(U,V) \leq 1$, where the equality holds only when there is an one-to-one correspondence between U and V.*
3. *The CCD is symmetric, that is, $CCD(U,V) = CCD(V,U)$.*
4. *For any strictly monotone transformations $M(\cdot)$ and $N(\cdot)$, $CCD(M(U), N(V)) = CCD(U,V)$.*
5. *If (U,V) is bivariate Gaussian with correlation coefficient ρ, then $CCD(U,V)$ is strictly increasing in $|\rho|$.*

Properties 1 and 2 together dictate that the proposed metric takes values between 0 (independent) and 1, and the strongest association is between two variables which are deterministic functions of each other, regardless of the specific form of their quantitative relationship. For example, in the eyes of CCD, the strength of the association between U and V in the following two distinct examples are identical (also maximal):

$$(a)\ V = \sin(\pi U),\ U \sim unif(0,1); \quad (b)\ V = U, \quad U \sim unif(0,1);$$

where $unif(0,1)$ stands for uniform random variables on $[0,1]$. In this sense we could say that CCD is fair, while the same thing couldn't always be said about other more sophisticated measures; for example, the distance correlations of the two pairs above are indeed different. Property 3 indicates that CCD can be considered as a symmetric distance between two variables. Property 4 implies that CCD is distribution-free; as a result, if U and V are both continuous, with marginal distribution functions $F_U(.)$ and $F_V(.)$, then $CCD(U,V)$=$CCD(F_U(U), F_V(V))$. The last property, also known as Rényi's Axiom 7 (Rényi , 1959), relates CCD to the Pearson correlation coefficient for normal random variables. This property guarantees that in Gaussian linear regression models, the CCD-based variable screening (see Section 3) should behave very much like the SIS based on R^2.

2.2 Estimation of CCD

We focus on the estimation of $CCD(U,V)$ for continuous random variables U and V, with bounded compact support $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Let $(U_i, V_i), i = 1, \cdots, n$, be independent and identically distributed (IID) observations of (U,V). Throughout this paper, $K(\cdot)$ denotes some kernel (density) function, h_n is a smoothing parameter, and $h_n \to 0$ as $n \to \infty$.

For any given point $(u, v) \in \mathcal{X} \triangleq \mathcal{X}_1 \otimes \mathcal{X}_2$, define

$$\hat{f}_U(u) = \frac{1}{n}\sum_{i=1}^n K_{h_n}(U_i - u), \quad \hat{F}(v|u) = \{\hat{f}_U(u)\}^{-1}\frac{1}{n}\sum_{i=1}^n K_{h_n}(U_i - u)I(V_i < v),$$

$$\hat{C}_{V|U}(v) = \frac{1}{n}\sum_{j=1}^n \{\hat{F}_{V|U}(v|U_j) - I(V_j < v)\}^2,$$

where $I(\cdot)$ is the indicator function and $K_{h_n}(\cdot) = K(\cdot/h_n)/h_n$. An estimate of $E\{C_{V|U}(V)\}$ is thus given by

$$\frac{1}{n}\sum_{k=1}^n \hat{C}_{V|U}(V_k) = \frac{1}{n^2}\sum_{j,k=1}^n \{\hat{F}_{V|U}(V_k|U_j) - I(V_j < V_k)\}^2. \tag{2.7}$$

As for the estimation of the denominator D_V, write

$$\hat{F}_V(v) = \frac{1}{n}\sum_{i=1}^n I(V_i < v), \quad \hat{D}_V = \frac{1}{n}\sum_{k=1}^n \hat{F}_V(V_k)\{1 - \hat{F}_V(V_k)\}.$$

Estimates of $E\{C_{U|V}(u)\}$ and D_U, denoted by $\hat{C}_{U|V}(u)$ and $\hat{D}_U$, are defined in a similar manner as their analogues. Finally, an estimate of $CCD(U, V)$ is given by

$$\widehat{CCD}(U, V) = 1 - \frac{1}{2}\left\{\frac{\sum_{k=1}^n \hat{C}_{V|U}(V_k)}{n\hat{D}_V} + \frac{\sum_{k=1}^n \hat{C}_{U|V}(U_k)}{n\hat{D}_U}\right\}. \tag{2.8}$$

To study the convergence rate of estimator (2.8), we need to impose the following assumptions and notations. Write $R(K) = \int K^2(v)dv$.

(A1) The marginal probability density functions of U and V are all bounded away from zero on their supports.

(A2) Both the marginal and the joint density functions have bounded third-order (partial) derivatives.

(A3) The kernel function $K(\cdot)$ is bounded, symmetric with bounded support.

(A4) $nh_n^8 \to 0$, $nh_n^2/\log^2 n \to \infty$, as $n \to \infty$.

Theorem 1. *Suppose conditions (A1)-(A4) hold. When $CCD(U, V) > 0$,*

$$n^{1/2}\{\widehat{CCD}(U, V) - CCD(U, V)\} \xrightarrow{d} N(0, \Phi),$$

for some constant Φ, *which depends on both the marginal and the conditional distribution functions. Yet when* U *and* V *are independent, i.e.* $CCD(U,V) = 0$, *then* $\widehat{CCD}(U,V)$ *converges to zero at a faster rate of* $(nh_n)^{-1}$, *and*

$$nh_n\widehat{CCD}(U,V) + R(K) \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j(\chi^2_{1j} - 1),$$

where χ^2_{1j}, $j = 1, \cdots$, *are independent* χ^2_1 *random variables, while* λ_j, $j = 1, \cdots$, *are the eigen values of a Hilbert-Schimidt integral operator which is independent of both marginal distributions* $F_U(.)$, $F_V(.)$, *and only depends on the kernel function* $K(.)$.

REMARK 4. Theorem 1 indicates that the estimator of CCD(U, V) is root-n consistent without undersmoothing. This is a desirable property of CCD, while it is not enjoyed by other estimators which are seemingly closely related to CCD. For example, consider the estimate of the Kolmogorov-Smirnov type of statistic (2.6)

$$K_n = \frac{1}{n^2} \sum_{j,k=1}^{n} \{\hat{F}_{V|U}(V_k|U_j) - \hat{F}_V(V_k)\}^2.$$

Under the same set of conditions (A1)-(A4), we show in the supplement that

$$K_n = \int \{F_{V|U}(v|u) - F_V(v)\}^2 dF_V(v) dF_U(u) + O_p(h_n^2 + (nh_n)^{-1/2}), \tag{2.9}$$

where the coefficient of the term $O_p(h_n^2)$ on the right hand side is not diminishing, which means undersmoothing is required for K_n to be root-n consistent.

REMARK 5. From Theorem 1, we know that when two random variables are independent, the (asymptotic) distribution of their estimated CCD is independent of their respective distributions; we could utilize this fact to simulate the null distribution for the purpose of hypothesis testing.

3 CCD-based Variable Screening

An important application of dependence measures is to serve as marginal utilities to screen out irrelevant variables in ultrahigh dimensional problems. In this section, we study a sure independence screening procedure based on CCD.

Let Y be the response variable, and $\mathbf{x} = (X_1, \cdots, X_p)^\top$ be the predictor vector, where $p \gg n$ and $\top$ stands for the transpose of a vector. Without specifying any structure on the regression model which relates Y with $\mathbf{x}$, we define

$$\mathcal{D} = \{k : F_{Y|X_k}(y|\cdot) \text{ functionally depends on } X_k \text{ for some } y\},$$

as the active predictor subset, and denote by $\mathcal{I} = \{1, 2, \cdots, p\} \setminus \mathcal{D}$ the inactive predictor subset. The goal of variable screening is to select a subset of predictors of a moderate size which contains the active set $\mathcal{D}$.

For $k = 1, \cdots, p$, let $\omega_k = CCD(X_k, Y)$, the CCD index associated with Y and the kth explanatory variable. With IID observations $\{(\mathbf{x}_i, Y_i) : 1 \le i \le n\}$, we estimate ω_k as $\hat{\omega}_k = \widehat{CCD}(X_k, Y)$; the chosen subset is then defined as

$$\hat{\mathcal{D}} = \{k : 1 \le k \le p, \hat{\omega}_k \ge cn^{-\tau}\}, \tag{3.1}$$

where constants c and τ are as given in the following assumptions.

(A5) There exist constants $c > 0$ and $\tau \in [0, 1/2)$, such that $\min_{k\in\mathcal{D}} \omega_k \ge 2cn^{-\tau}$.

(A6) The bandwidth h_n satisfies $n^\tau h_n^2 \to 0$, $n^{1-2\tau} h_n / \log n \to \infty$ as $n \to \infty$.

Condition (A5) is commonly seen in the independence screening literature, for example, (B2) in Li, Zhong and Zhu (2012), (A2) in Shao and Zhang (2014), and Condition 3 of Fan and Lv (2008). It is assumed that the active predictors all meet the requirement of displaying 'significant' correlation with Y in terms of CCD. Meanwhile the signals of active predictors are allowed to diminish at a certain rate of n as n goes to infinity, which is only reasonable, for the total number of predictors p is also allowed to increase with n. In a sense this condition also implicitly put restrictions on dependence among predictors, such as important predictors are not rendered marginally 'insignificant' due to dependence among predictors, though it is hard to give a mathematical form or to quantify.

Theorem 2. Under conditions (A1)-(A6), there exists constant $a > 0$ such that

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

If further $\log(p) = o(n^{1-2\tau})$ as $n \to \infty$, then

$$\Pr(\mathcal{D} \subseteq \hat{\mathcal{D}}) \ge 1 - O\Big(s_n \exp(-an^{1-2\tau})\Big).$$

where s_n is the cardinality of $\mathcal{D}$.

The second part of Theorem 2 implies that even when p increases with n at an exponential rate, the chosen subset $\hat{\mathcal{D}}$ of (3.1) still contains the true active set with probability approaching one as the sample size tends to infinity.

Next, we study the size of the chosen model after the CCD-based variable screening in the asymptotical sense.

Theorem 3. Under conditions (A1)-(A6), there exists a constant $a > 0$, such that,

$$\Pr\left(|\hat{\mathcal{D}}| \le O\left(n^{\tau}\sum_{k}^{p}|\omega_k|\right)\right) \ge 1 - O\Big(p\exp(-an^{1-2\tau})\Big). \tag{3.2}$$

REMARK 6.Marginal variable screening methods may fail to identify important predictors which are nevertheless marginally independent of the response due to dependence among the covariates. Therefore iterative procedures of variable screening are often carried out to mitigate the risk of missing the truly important predictors. Following ideas similar to those in Zhu, et al. (2011) and Zhong and Zhu (2015), an iterative procedure of CCD-SIS could also be proposed. Details are given in the supplement.

4 Extensions

4.1 The class of $\alpha-$norm CCD

It is important to note that the CCD as defined in (2.4) is based on the squared norm, since $C_{V|U}(v) = E\{I_V(v) - E(I_V(v)|U)\}^2$, where $E(I_V(v)|U)$ is the minimizer

of $E\{(I_V(v)-a)^2|U\}$, while $\mathrm{Var}(I_V(v)) = E\{I_V(v)-E(I_V(v))\}^2$, where $E(I_V(v))$ is the minimizer of $E\{(I_V(v)-a)^2\}$. One might then wonder whether any other norm could be used instead which, when used for the purpose of independence test and variable screening, leads to better performance. A naive approach is to replace the squared norm with the $\alpha-$norm:

$$C_{V|U}(v;\alpha) \triangleq E\{|I_V(v)-E(I_V(v)|U)|^\alpha\} = E\{|I_V(v)-F_{V|U}(v|U)|^\alpha\}. \tag{4.1}$$

An estimate of $E\{C_{V|U}(V;\alpha)\}$ could be formed in exactly the same manner as (2.7), only with $|.|^2$ replaced with $|.|^\alpha$. Yet unlike in the squared norm case, it can be shown that this estimator has a bias of order $O(h_n^2)$.

Next, we need to devise an $\alpha-$norm alternative to the 'standardisation' factor. An obvious candidate would be $E\{|I_V(v)-F_V(v)|^\alpha\}$, yet the problem is we do not always have $C_{V|U}(v;\alpha)/E\{|I(V\le v)-F_V(v)|^\alpha\} \le 1$; and a proper 'standardisation' factor is expected to keep the ratio between 0 (deterministic) and 1 (independent). With this in mind, for $\alpha > 0\,(\alpha \neq 1)$, we define

$$Q_V^\alpha(v) = \frac{1}{1+\left\{\frac{F_V(v)}{1-F_V(v)}\right\}^{1/(1-\alpha)}}, \qquad Q_{V|U}^\alpha(v|u) = \frac{1}{1+\left\{\frac{F_{V|U}(v|u)}{1-F_{V|U}(v|u)}\right\}^{1/(1-\alpha)}},$$

which are respectively the minimizer of $E[|I(V_i\le v)-a|^\alpha]$ and $E[|I(V_i\le v)-a|^\alpha|U=u]$, for given values of α, v and u. With $\alpha=1$, these are either 0 or 1, depending on whether $F_V(v)$ (or $F_{V|U}(v|u)$) is less than $1/2$. Consequently, the analogue of (2.1) in this case,

$$C_{V|U}^\alpha(v) \triangleq E[|I(V\le v)-Q_{V|U}^\alpha(v|u)|^\alpha], \tag{4.2}$$

is always no greater than the corresponding standardisation factor

$$D_V^\alpha \triangleq E[|I(V\le v)-Q_V^\alpha(v)|^\alpha],$$

and equality holds only when $F_V(v) = F_{V|U}(v|u)\,(\alpha\neq 1)$. Then, we can follow the definition of $CCD(U,V)$ to define

$$CCD_\alpha(U,V) = 1-\{E\{C_{V|U}^\alpha(V)\}/D_V^\alpha + E\{C_{U|V}^\alpha(U)\}/D_U^\alpha\}. \tag{4.3}$$

Its estimate, $\widehat{CCD}_\alpha(U,V)$, can be formulated along the lines described in Section 2.1. Similar to (4.1), estimation of $E\{C^\alpha_{V|U}(V)\}$ is also overshadowed by an asymptotic bias of order $O(h_n^2)$. However, root-n consistency is still possible if a smaller bandwidth is used.

Theorem 4. *Under some conditions (A1)-(A4), if $\alpha \neq 2$ and $nh_n^4 \to 0$, we have*

$$\widehat{CCD}_\alpha(U,V) - CCD_\alpha(U,V) = O_p(n^{-1/2}). \tag{4.4}$$

Compared with $\widehat{CCD}(U,V)$ where $\alpha = 2$, root-n consistency of $\widehat{CCD}_\alpha(U,V)$ calls for smaller bandwidth ($nh_n^4 \to 0$). Also, for independent pair U and V, $\widehat{CCD}_\alpha(U,V)$ is not able to achieve a faster rate, if $\alpha \neq 2$. Though in theory it is impossible to obtain results regarding the optimal choice for α, our experience working with simulated data suggests that $\widehat{CCD}_\alpha(U,V)$ with $\alpha < 1$ in general performs better than with $\alpha = 2$ when used for variable screening. Heuristically this might be explained by a simple fact that, with $0 < a, b < 1$ and $0 < \alpha < 2$, $(a-b)^2 < |a-b|^\alpha$; in other words, the difference between a and b will be magnified with smaller α.

4.2 Multivariate CCD

It is also important to test the independence between two random vectors or mutual independence among multivariate random variables. Let $\mathbf{u} = (U_1, \ldots, U_p)^\top$ and $\mathbf{v} = (V_1, \ldots, V_q)^\top$ be two random vectors of dimensions p and q, respectively. There are a number of possible ways to extend the definition of CCD to multivariate case. For example, to characterize joint independence, we might consider

$$MCCD_\alpha(\mathbf{u},\mathbf{v}) \triangleq \max_{\mathbf{a}\in R^p, \mathbf{b}\in R^q} CCD_\alpha(\mathbf{a}^\top\mathbf{u}, \mathbf{b}^\top\mathbf{v}), \tag{4.5}$$

Some properties of this metric are as follows.

Lemma 2. For any two random vectors $\mathbf{u}$ and $\mathbf{v}$,

1. $MCCD_\alpha(\mathbf{u},\mathbf{v}) = 0$ if and only if $\mathbf{u}$ is independent of $\mathbf{v}$.

2. $MCCD_\alpha(\mathbf{u}, \mathbf{v}) = 1$ if and only if some (nontrivial) linear combination of $\mathbf{u}$ is a monotone function of some linear combination of $\mathbf{v}$.

The projection directions, referred to the $\mathbf{a}$ and $\mathbf{b}$ at which the maximization in (4.6) is realized, can be estimated using method in Xia (2008). Random projections in computer science Bingham and Mannila (2001) could also be used for this purpose. To ease the computational burden, we can also consider the below measure of pairwise dependence between $\mathbf{u}$ and $\mathbf{v}$,

$$MCCD^*_\alpha(\mathbf{u}, \mathbf{v}) \stackrel{def}{=} \frac{1}{pq} \sum_{k=1}^{p} \sum_{\ell=1}^{q} CCD_\alpha(U_k, V_\ell). \tag{4.6}$$

Obviously, $MCCD^*_\alpha(\mathbf{u}, \mathbf{v}) \leq MCCD_\alpha(\mathbf{u}, \mathbf{v})$. However, there are interesting discussion about the "equivalence" between them; see Sun (1998).

Pfister, et al. (2016) investigated the problem of testing whether a set of random variables, $\mathbf{u} = (U_1, \ldots, U_p)^\top$, are jointly (mutually) independent. Similarly, we can define the mutual dependence and pairwise dependence as follows

$$MCCD_\alpha(\mathbf{u}) = \frac{1}{p} \sum_{k=1}^{p} \max_{\theta_k} CCD_\alpha(U_k, \theta_k^\top U_{[-k]}), \tag{4.7}$$

where $U_{[-k]}$ the remaining elements of $\mathbf{u}$ after U_k is removed, and

$$MCCD^*_\alpha(\mathbf{u}) = \frac{1}{p^2 - p} \sum_{k \neq \ell} CCD_\alpha(U_k, U_\ell).$$

Alternatively, one might opt for the largest among all $CCD_\alpha(U_k, U_\ell)$, $k \neq l$. While their average is more conservative, the largest among them would have a higher rejection rate but less stable. Somewhere in the middle, for example, we could consider the summation of the largest p pairs, to achieve some sort of a balance

$$MCCD^+_\alpha(\mathbf{u}) = \frac{1}{p} \sum_{i=1}^{p} CCD^{(i)}_\alpha(U_k, U_\ell), \tag{4.8}$$

where $CCD^{(i)}_\alpha(U_k, U_\ell)$ is the ith largest values in $\{CCD(U_k, U_\ell) : k, \ell = 1, ..., p, \text{ and } k \neq \ell\}$. Our simulations suggest that $MCCD^+_\alpha$ has better performance than $MCCD_\alpha(\mathbf{u})$ and $MCCD^*_\alpha(\mathbf{u})$.

5 Numerical Studies

Note that binarization of a variable is invariant to the monotonic transformation. In the following calculation, we first transform the data to uniformly distributed on [0, 1] marginally before the estimation of CCD. In our simulations, we take $\alpha = 0.5$, and use the Gaussian kernel and the rule-of-thumb bandwidth $h = 1.06n^{-2/7}$ for undersmoothing as required for $\alpha \neq 2$ (Theorem 4).

5.1 Simulations

In the following simulations, we first compare the testing power of our method with other existing methods, including distance correlation (DC) test in Szekely, Rizzo and Bakirov (2007), the mutual information (MI) test using the KNN method with $k = 6$ as Kinney and Atwal (2014), the rank-based test (HHG) in Heller, Heller and Corfine (2013) and the dependence of multi-variables (dHSIC) as in Pfister, et al. (2016). Certainly there are many other competing methods; yet as demonstrated in Ding and Li (2015), the HHG test of Heller, Heller and Corfine (2013) is quite often the best performer. The comparison is carried out on all the models used in HHG. Due to the complicated distribution of the statistics, we use the exact (permutation) test, which usually have a precise control over the type-I error.

Example 1. Consider the same models as in Heller, Heller and Corfine (2013) but with more choices of signal to noise ratio (SNR) in order to highlight the difference in performance of different methods.

Four independent clouds: $(X, Y) = (x_0 + U, y_0 + V)$ where $U, V \overset{IID}{\sim} N(0, 1)$ and x_0, y_0 are IID taking values -1, 1 with probability 0.5 each. For this model, X and Y are independent.

W: $X \sim unif(-1, 1), Y = C(X^2 - 0.5)^2 + \varepsilon,$ where $\varepsilon \sim unif(0, 1)$.

Diamond: U and V are IID $unif(-1,1)$, $U' = U\cos(\frac{1}{4}\pi)+V\sin(\frac{1}{4}\pi)$, $V' = -U\sin(\frac{1}{4}\pi)+V\cos(\frac{1}{4}\pi)$; U'' and V'' are IID and follow $unif(-1,1)$; and $(X,Y) = (R < C)\times(U',V') + (R > C)\times(U'',V'')$.

Parabola: $X \sim unif(-1,1), Y = CX^2 + \varepsilon$, where $\varepsilon \sim unif(0,1)$.

Two parabola: $X \sim unif(-1,1), Y = (CX^2+\varepsilon)V$, where $V = \pm 1$ with equal probabilities.

Circle: $U \sim unif(0,1), X = C\sin(2\pi U) + \epsilon_1, Y = 4.2\cos(2\pi U) + \epsilon_2$, where $\epsilon_1, \epsilon_2 \overset{IID}{\sim} N(0,1)$.

Variance: $Y = \epsilon\sqrt{CX^2+1}$, where $X, \epsilon \overset{IID}{\sim} N(0,1)$.

Log: $Y = C\log(X^2) + \epsilon$, where $X,\ \epsilon \overset{IID}{\sim} N(0,1)$.

For each model except the first one, the coefficient C controls the SNR: the greater C is, the greater the SNR. For any given model and fixed sample size, the value of C is chosen such that the highest rejection rate among these five contesting methods is roughly 95%. Based on 2000 replications, the rejection rate (testing power) of all five methods are gathered in Table 1 with the highest values highlighted (boldface). Except for model **Circle**, in which case CCD is marginally outperformed by dHSIC, the testing power of CCD is sometimes comparable while other times significantly higher than those of all its competitors. It is interesting to see that in most settings, and as sample size n increases from 200, 500, and to 1000, the advantage of CCD over the other methods becomes more dominant.

Next, we compare the efficiencies of different methods in variable screening for ultrahigh dimensional data. The methods compared below include nonparametric independence screening (NIS) in Fan, Feng and Song (2011), sure independence screening via distance correlation (DC) in Li, Zhong and Zhu (2012), sure independence

Table 1: Testing Power of different methods for all the models in Example 1.

n	model	C	DC	MI	dHSIC	HHG	CCD
200	Independent	—	0.045	0.050	0.045	0.046	0.052
	W	1.20	0.311	0.961	0.637	0.674	**0.972**
	Diamond	0.70	0.113	0.862	0.930	**0.980**	0.967
	Parabola	0.25	0.233	0.898	0.509	0.670	**0.958**
	Two parabola	0.35	0.074	**0.983**	0.672	0.780	0.967
	Circle	2.75	0.091	0.775	**0.939**	0.895	0.812
	Variance	1.20	0.466	0.962	0.795	**0.968**	0.944
	Log	0.18	0.829	0.811	0.853	0.909	**0.943**
500	Independent	—	0.061	0.054	0.049	0.052	0.053
	W	0.50	0.104	0.901	0.285	0.368	**0.942**
	Diamond	0.35	0.063	0.742	0.663	0.843	**0.967**
	Parabola	0.12	0.115	0.885	0.299	0.494	**0.964**
	Two parabola	0.16	0.058	0.950	0.382	0.533	**0.944**
	Circle	2.10	0.143	0.702	0.955	**0.956**	0.823
	Variance	0.45	0.340	0.942	0.675	**0.949**	0.948
	Log	0.12	0.914	0.823	0.908	0.954	**0.970**
1000	Independent	—	0.053	0.051	0.042	0.048	0.051
	W	0.28	0.076	0.903	0.219	0.255	**0.959**
	Diamond	0.20	0.049	0.662	0.450	0.596	**0.970**
	Parabola	0.07	0.081	0.876	0.241	0.364	**0.967**
	Two parabola	0.10	0.052	0.954	0.318	0.448	**0.963**
	Circle	1.80	0.148	0.565	0.945	**0.965**	0.777
	Variance	0.25	0.223	0.889	0.626	0.919	**0.937**
	Log	0.08	0.862	0.703	0.875	0.926	**0.931**

ranking and screening (SIRS) in Zhu, et al. (2011), Quantile-adaptive variable screening (QaSIS$_\tau$) for the τth conditional quantile of the response in He, Wang and Hong (2013), HHG (Heller, Heller and Corfine, 2013) and MBKR in (Zhou and Zhu, 2018). Each method is evaluated based on their averaged smallest model sizes for the important variables. For example, if a model has three truly important variables: X_1, X_2 and X_3, and if according to a certain criterion, the rankings of these three variables, are k_1, k_2, k_3, respectively, then the averaged model size in this case is given by $(k_1 + k_2 + k_3)/3$. The smaller the averaged model size is, the better the variable screening method is considered to be.

Example 2. We consider the regression models used in both Li, Zhong and Zhu (2012) and He, Wang and Hong (2013):

$$(M1) \quad Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 I(X_{12} < 0) + c_4\beta_4 X_{22} + \varepsilon;$$

$$(M2) \quad Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 I(X_{12} < 0) + c_4\beta_3 X_{22} + \varepsilon;$$

$$(M3) \quad Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 I(X_{12} < 0) X_{22} + \varepsilon;$$

$$(M4) \quad Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 I(X_{12} < 0) + \exp(c_4|X_{22}|)\varepsilon;$$

$$(M5) \quad Y = 5g_1(X_1) + 3g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sqrt{1.74}\varepsilon;$$

$$(M6) \quad Y = 2X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5 + \exp(X_{20} + X_{21} + X_{22})\varepsilon;$$

$$(M7) \quad Y = 2(X_1^2 + X_2^2) + \exp((X_{18} + ... + X_{30})/5)\varepsilon;$$

$$(M8) \quad Y = min(Y, C), \text{ where } C \sim 0.4N(-5, 4) + 0.1N(5, 1) + 0.5N(55, 1).$$

The coefficients in (M1)-(M4) are exactly the same as in Li, Zhong and Zhu (2012); in (M5), $g_1(x) = x$, $g_2(x) = (2x-1)^2$, $g_3(x) = \sin(2\pi x)/(2 - \sin(2\pi x))$, and $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3$; in (M8), Y is firstly generated in the same way as in (M6). As for the covariates, we consider three designs: (A) $X \sim N(0, \Sigma)$ where $\Sigma = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}$, $i, j = 1, \ldots, p$; (B) the same as (A), but with even-index covariates (i.e. X_{2k}, $k \geq 1$) replaced with Z_k: $Z_k = -1$, if $X_{2k} < 0$ and $Z_k = 0.5$, otherwise; (C) is derived from (A) but with $\sigma_{ij} \equiv 0.85$ when $i \neq j$, and 1 otherwise.

For any given model and covariate design, we consider two settings: $\rho = 0$ or 0.8 as in Li, Zhong and Zhu (2012). With $n = 200$, $p = 2000$, and 1000 repetitions, the averaged model size (AMS) of all the competing methods are reported in Table 2; a smaller AMS indicates a more effective method. For (M1), i.e. a linear dependence between Y and its covariates, all methods perform comparably well and the DC method is the best. For all the other models, HHG and CCD are clearly in a league of their own. While the performance of these methods are relatively comparable for design A, with design B, CCD does stand out as the clear winner.

Table 2: Mean of the smallest averaged model sizes for models in Example 2

ρ	model	DC	QaSIS$_{0.75}$	NIS	SIRS	HHG	MBKR	CCD
				[Design A]				
0.0	1	81.17	151.39	98.05	77.21	140.13	80.84	107.26
	2	233.87	272.50	371.95	463.75	126.38	211.67	98.82
	3	134.01	190.01	613.79	499.36	29.17	141.01	65.13
	4	200.40	255.81	653.83	347.63	188.47	149.38	186.09
	5	470.93	589.36	522.80	503.71	435.62	456.88	404.39
	6	133.04	348.38	640.50	178.85	165.37	141.05	149.48
	7	816.11	860.36	854.98	952.80	816.67	690.79	764.90
	8	519.70	571.49	572.65	556.34	519.47	492.58	489.76
0.8	1	12.01	26.67	12.38	15.27	21.12	11.93	14.58
	2	8.56	11.70	5.91	470.50	5.99	7.73	6.04
	3	19.17	71.03	231.06	497.17	7.18	15.24	8.44
	4	64.43	112.92	597.89	234.96	47.52	23.88	40.76
	5	2.97	4.08	2.71	56.64	3.19	3.67	2.77
	6	26.43	19.23	685.68	9.20	4.63	5.04	4.61
	7	35.06	379.30	558.93	246.17	34.41	51.44	24.70
	8	54.75	50.67	54.72	485.18	46.84	93.32	46.00
				[Design B]				
0	1	158.63	222.76	158.64	112.92	236.30	108.16	176.05
	2	270.82	358.14	379.47	271.98	236.21	194.02	142.09
	3	185.94	380.09	469.80	197.92	141.31	110.15	98.95
	4	249.58	411.28	569.48	190.09	311.73	183.08	260.62
	5	177.01	195.42	198.65	257.82	139.74	218.65	155.40
	6	259.86	463.87	530.41	286.30	268.25	231.35	206.31
	7	861.69	906.12	894.05	900.43	860.01	824.68	809.83
	8	257.53	542.91	319.90	299.57	214.66	245.76	185.68
				[Design C]				
0.85	1	584.56	543.69	554.59	587.54	554.09	583.28	561.26
	2	278.62	282.16	164.02	780.79	140.12	291.61	162.06
	3	264.86	370.33	292.84	805.35	123.11	245.59	159.83
	4	444.61	617.77	807.50	766.71	390.10	446.94	374.52
	5	812.77	772.75	816.40	861.76	536.03	725.86	522.45
	6	506.20	745.11	909.90	721.18	275.02	501.46	323.23
	7	709.53	868.38	928.59	868.24	641.24	693.86	635.20
	8	770.68	765.54	799.83	834.88	661.83	741.71	690.01

Example 3. This is Example 5 of Pfister, et al. (2016), which is proposed for the testing of mutual independence of multiple random variables. The two designs are

considered as follows:

$$\begin{aligned} \text{Dense:} \quad & X_i = CH^2 + \varepsilon_i, i = 1, ..., p; \\ \text{Sparse:} \quad & X_i = \begin{cases} CH^2 + \varepsilon_i, & i = 1, 2; \\ \varepsilon_i, & i > 2 \end{cases} \end{aligned}$$

where $H, \varepsilon_1, \ldots, \varepsilon_p$ are IID N(0, 1). For different values of C, which controls the SNR, we compare our method with dHSIC of Pfister, et al. (2016) in terms of their (empirical) testing power.

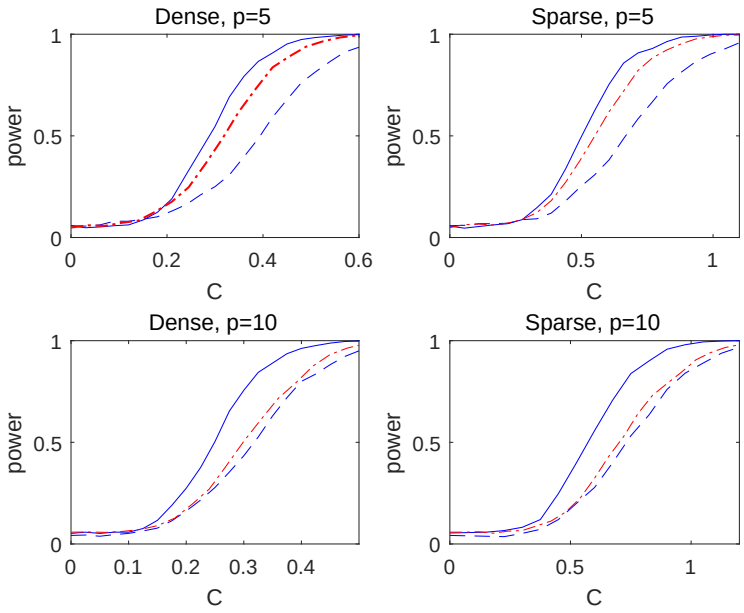

Figure 1: Powers of testing of dHSIC (dashed line), MCCD* (dash-dot line) and MCCD$^+$ (solid line) for different signal-to-noise ratios (C) in Example 3.

With sample size $n = 100$, and based on 1000 repetitions, the testing power of these two competing methods for the two different designs above with $p = 5$ and $p = 10$ are plotted in Figure 1. As we can see that in all cases, our methods are more powerful than the dHSIC of Pfister, et al. (2016) in detecting the presence of dependence among multivariate random variables.

5.2 Real data examples

Example 4 (Boston housing data)**.** The Boston housing data set from the 1970 US census was collected to investigate property values in the suburbs of Boston (Harrison and Rubinfeld, 1978). The data contain 506 observations on 13 covariates and on the median value of owner-occupied homes ('*medv*'), which serves as the response variable. The original data and the covariates description can be found in R package *mlbench*. The main goal is to predict house values using the information of the 13 covariates. Harrison and Rubinfeld (1978) established a linear regression model of the house values using the transformed forms of 13 covariates. The data have also been studied by many other researchers (Lin and Zhang, 2006; Ravikumar, et al., 2009; Fan, Ma and Dai, 2014). These empirical results show that the relationships between the response and the covariates are definitively not linear. Figure 2 contains scatter plots of the response '*medv*' versus every predictor superimposed with fitted quartic spline curves (expect a binary variable '*chas*', a dummy variable indicates whether tract bounds Charles river).

To evaluate the finite-sample performance of the various screening methods in an ultrahigh dimensional setting, we augment the data set by some 'artificial' predictors generated as follows. First, for each covariate X_j, we generate 30 extra variables using combination $X_j/4+Z_k$, where $(Z_1,\ldots,Z_{30})$ are independent standard normal random variables. These 'artificial' predictors would display spurious associations with the response, thus bringing more challenges to variable screening. Secondly, for each covariate X_j, we further generate 69 independent auxiliary variables by bootstrapping X_j. Bootstrapping instead of generating samples from some specific distribution can ensure that the resulting auxiliary variables have the same data structure as the original predictors. Note that the binary variable '*chas*' is not considered for comparison study and is directly included in the regression model in the later analysis. Therefore, after combining the original 12 covariates with the artificial predictors, the

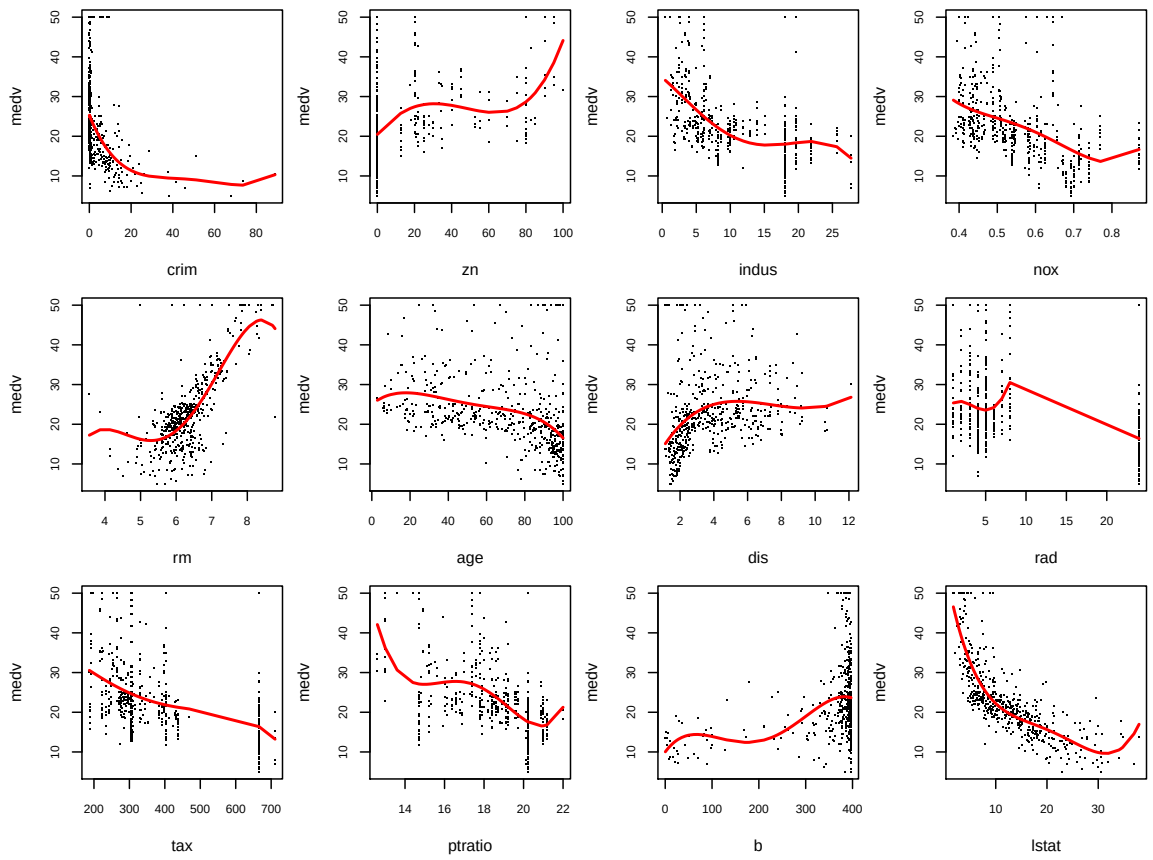

Figure 2: Scatter plots of the response '*medv*' versus each predictor with the fitted quartic spline curves

dimensionality of the expanded data is now 1200.

Next we apply different independence screening methods to the augmented data set and record the rank of each covariate. Table 3 reports the averaged rank of each original covariate based on 50 augmented data sets obtained by repeating the above procedure 50 times. We can see that covariates '*rm*' (average number of rooms per dwelling), '*lstat*' (percentage of lower status of the population) are identified as the most important factors for housing values by most methods. CCD-based method in general assigns higher ranks to the original covariates than other screening methods do. For example, '*b*' $(1000(B-0.63)^2$ where B is the proportion of blacks by town), '*age*' (proportion of homes built before 1940), '*dis*' (weighted distances to five employment centers) are ranked in the top by our method. In view of Figure 2, it

is interesting to see that these covariates should be picked up by our CCD-based method. For instance, when '*dis*' is in a lower level, there exists a clear increasing trend between the response and '*dis*'. In addition, we fit an additive model based on the selected $2[n/\log(n)] = 162$ variables by the screening methods and then apply the penalized grouped Lasso (Huang, Horowitz and Wei, 2010) to further select important variables and estimate unknown additive functions. The average of the adjusted R^2 values based on 10 simulations of artificial predictors are also reported in Table 3. Since the CCD-based procedure identifies more truly relevant covariates, the corresponding regression model has higher adjusted R^2 value than models selected using other methods. This leads us to conclude that CCD-SIS is more powerful to detect nonlinear dependence between two variables.

As a final touch, we examine the prediction performances of the models with variables selected by these methods. We split the data randomly to a training set of 406 observations and a testing set containing 100 observations. We use the training set to estimate the additive model, and the testing set to check the prediction performance. The mean squared prediction error (MSPE) is defined as $MSPE = n.test^{-1} \sum_{i=1}^{n.test} (\hat{Y}_i^* - Y_i^*)^2$, where $n.test$ is the size of testing data, $\hat{Y}_i^*$ and Y_i^* denote the predicted and true (observed) values of the ith response in the testing data, respectively. We run this experiment 50 times and the median MSPE values are also included in Table 3.

Example 5 (Gene expression data)**.** Various genetic diseases are caused by abnormality in gene expressions which are partly regulated by DNA copy number alternations (CNAs). Therefore, it is important to study the dependence between CNAs and certain gene expression in cell development pathways. The Cancer Genome Atlas (TCGA)[1] project applies high-throughput genome analysis techniques to catalogue genetic mutations responsible for cancer. In this example, we study a TCGA data

[1]http://cancergenome.nih.gov

Table 3: The averaged ranks for each covariate, adjusted R^2 values of the additive models, and the MPSE for Boston housing data.

Covariates	SIS	DC	SIRS	NIS	$\text{QaSIS}_{0.5}$	CCD
crim	95.8	34.0	130.5	35.1	65.6	36.2
zn	152.9	182.3	219.3	119.6	213.7	180.6
indus	34.0	33.0	33.0	34.0	32.5	33.0
nox	75.0	36.0	64.8	75.7	67.2	34.0
rm	2.1	2.0	32.0	2.0	1.0	3.6
age	109.2	77.6	69.5	140.6	94.2	35.0
dis	264.7	246.3	238.9	212.5	177.7	101.0
rad	100.9	128.4	124.7	87.4	71.6	64.9
tax	48.2	52.8	47.0	47.2	48.4	39.0
ptratio	33.0	35.0	104.5	33.0	68.6	38.0
b	223.3	228.5	190.000	109.2	242.2	37.0
lstat	1.0	1.0	1.0	1.0	2.0	1.0
Adj R^2	0.846	0.847	0.847	0.854	0.849	0.879
MSPE	34.24	34.43	41.10	38.16	37.33	33.91

set which contains observations of 275 patients on 563 gene expression levels and 341 CNAs. Nonlinear dependence relationships between CNAs and gene expression always exhibit in the pathway studies. Thus, the commonly used Pearson correlation is not sufficient to characterize their relationships. As an illustrative example, we investigate significant CNAs for one important gene expression level, FAS gene, which encodes the FAS receptor. It can trigger apoptosis and play a critical role in the physiological regulation of programmed cell death. It has also been implicated in the pathogenesis of various malignancies and diseases of the immune system (Randhawa, et al., 2010; Cao, et al., 2010).

We apply our proposed CCD to measure the dependence between each CNA and FAS gene expression level and select three important CNAs: TP53, ALOX12, UBB. Figure 3 displays the scatter plots of the response FAS gene expression level versus three selected CNAs TP53, ALOX12, UBB. It is clearly seen that the relationship between the response and each top CNA is nonlinear. These results coincide with the existing genetic and medical findings in the literature. It has been shown that TP53 is

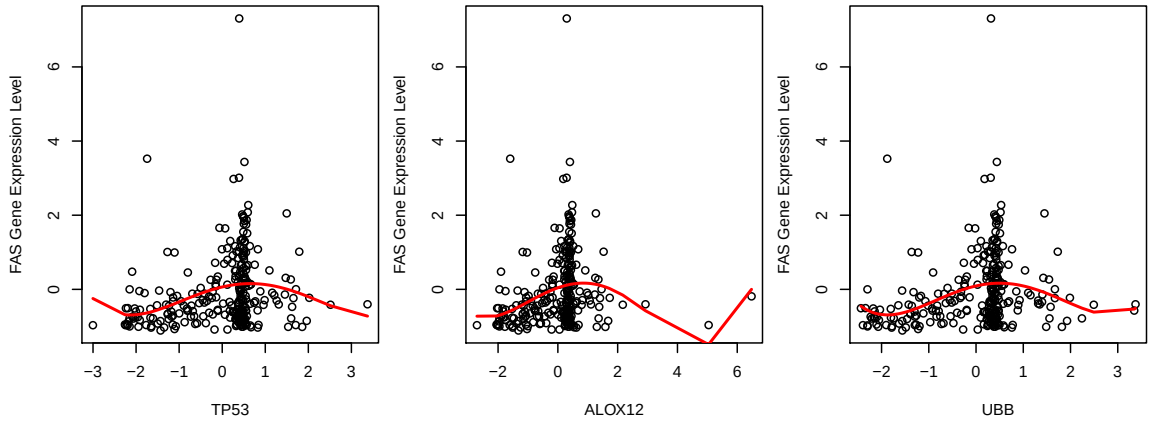

Figure 3: The scatter plots of the response FAS gene expression level versus three important CNAs TP53, ALOX12, UBB selected by CCD with quadratic spline fit curves.

the most commonly mutated tumor suppressor in human cancers and a large variety of TP53 mutations including deletions, missense and splicing alternations contribute to tumour progression (Kandoth, et al., 2013; Ciriello, et al., 2013; Liu, et al., 2016). ALOX12 has also been recognised as a major source of oxidative stress which may induce apoptosis (Miller, et al., 2015) and allelic variants of ALOX12 are associated with diseases including schizophrenia, atherosclerosis, cancers and menopause (Witola, et al. , 2014; Liu, et al., 2010). UBB encodes the ubiquitin which is closely related to the degradation of abnormal proteins and consequently involves in the regulation of gene expression (Conaway, Brower and Conaway, 2002). Many types of cancers also display an elevated level of ubiquitin (Oh, et al., 2013).

Supplementary Material. All technical proofs of results in this paper and an iterative procedure of CCD-SIS are included in a separate online supplemental file.

References

Bingham, E. and Mannila, H. (2001), "Random projection in dimensionality reduction: applications to image and text data". Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. 245-250.

Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961), "Distribution free tests of independence based on the sample distribution function," *The Annals of Mathematical Statistics*, **32**, 485–498.

Cao, Y., Miao, X., Huang, M., Deng, L., Lin, D., Zeng, Y. and Shao, J. (2010), "Polymorphisms of death pathway genes FAS and FASL and risk of nasopharyngeal carcinoma," *Molecular Carcinogenesis*, **49**, 944–950.

Ciriello, G., Miller, M., Aksoy, B., Senbabaoglu, Y., Schultz, N. and Sander, C. (2013), "Emerging landscape of oncogenic signatures across human cancers", *Nature Genetics*, **45**, 1127C1133.

Conaway, R. C., Brower, C. S. and Conaway, J. W. (2002), "Emerging roles of ubiquitin in transcription regulation", *Science*, **296**, 1254–1258.

Cui, H., Li, R. and Zhong, W. (2015), "Model-free feature screening for ultrahigh dimensional discriminant analysis," *Journal of the American Statistical Association,* **110,** 630-641.

Ding, A.A. and Li, Y. (2015), "Copula correlation: an equitable dependence measure and extension of Pearson's correlation," *arXiv:1312.7214.*

Fan, J. and Fan, Y. (2008), "High-dimensional classification using features annealed independence rules," *The Annals of Statistics*, **36**, 2605–2637.

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.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space (with discussion)," *Journal of the Royal Statistical Society, Series B*, **70**, 849–911.

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

Fan, J. and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with NP-Dimensionality," *The Annals of Statistics*, **38**, 3567–3604.

Gretton, A., O. Bousquet, A. Smola, and B. Schölkopf (2005), "Measuring statistical dependence with Hilbert-Schmidt norms." In *Algorithmic Learning Theory*, 63–67. Springer-Verlag.

Hall, P. and Miller, H. (2009), "Using generalized correlation to effect variable selection in very high dimensional problems," *Journal of Computational and Graphical Statistics*, **18**, 533–550.

Harrison, D. and Rubinfeld, D.L. (1978), "Hedonic prices and the demand for clean air," *Journal of Environmental Economics and Management*, **5**, 81–102.

He, X., Wang, L. and Hong, H. (2013), "Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data", *The Annals of Statistics*, **41**, 342–369.

Heller, R., Heller, Y. and Corfine, M. (2013), "A consistent multivariate test of association based on ranks of distances," *Biometrika*, **100**, 503–510.

Huang, J., Horowitz, J. and Wei, F. (2010), "Variable selection in nonparametric additive models," *The Annals of Statistics,* **38**, 2282–2313.

Ji, P. and Jin, J. (2012), "UPS delivers optimal phase diagram in high dimensional variable selection," *The Annals of Statistics*, **40**, 73–103.

Kandoth, C., McLellan, M., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J., Wyczalkowski, M. et al. (2013), "Mutational landscape and significance across 12 major cancer types," *Nature*, **502**, 333–339.

Kinney, J. and Atwal, G. (2014). "Equitability, mutual information, and the maximal information coefficient," *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 3354–3359.

Li, G., Peng. H., Zhang J. and Zhu, L. (2012), "Robust rank correlation based screening," *The Annals of Statistics*, **40**, 1846–1877.

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

Lin, Y., and Zhang, H. H. (2006), "Component selection and smoothing in multivariate nonparametric regression," *The Annals of Statistics,* **34**, 2272–2297.

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

Liu, P., Lu, Y., Recker, R., Deng, H. and Dvornyk, V. (2010) "ALOX12 gene is associated with the onset of natural menopause in white women," *Menopause,* **17**, 152–156.

Liu, Y., Chen, C., Xu, Z., Scuoppo, C., Rillahan, C., Gao, J., Spitzer, B., Bosbach, B., Kastenhuber, E. et al. (2016), "Deletions linked to TP53 loss drive cancer through p53-independent mechanisms," *Nature*, **531**, 471–475.

Mai, Q. and Zou, H. (2013), "The Kolmogorov filter for variable screening in high-dimensional binary classification," *Biometrika*, **100**, 229–234.

Miller, M., Wolf, E., Sadeh, N., Logue, M., Spielberg, J., Hayes, J., Sperbeck, E., and Schichman, S. (2015), "A novel locus in the oxidative stress-related gene ALOX12 moderates the association between PTSD and thickness of the prefrontal cortex," *Psychoneuroendocrinology,* **62,** 359–365.

Oh, C., Park, S., Lee, E. and Yooa, Y. (2013), "Downregulation of ubiquitin level via knockdown of polyubiquitin gene Ubb as potential cancer therapeutic intervention", *Scientific Reports,* **3**, 2623.

Pfister, N., Bühlmann, P., Schölkopf, B. and Peters, J. (2016), "Kernel-based tests for joint independence." (To appear in the *Journal of the Royal Statistical Society, Series B*)

Randhawa, S.R., Chahine, B.G., Lowery-Nordberg, M., Cotelingam, J.D. and Casillas, A.M. (2010), "Underexpression and overexpression of Fas and Fas ligand: a double-edged sword," *Annals of Allergy, Asthma & Immunology,* **104**, 286–292.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2009), "SpAM: Sparse additive models", *In Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference.*

Rényi, A. (1959), "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, **10**, 441–451.

Rosasco, L., Belkin, M., and De Vito, E.(2010), "On Learning with Integral Operators." *Journal of Machine Learning Research*, **11**, 905-934.

Shannon, C.E. and Weaver, W. (1949), *The mathematical theory of communication.* University of Illinois Press, Urbana, Illinois.

Shao, X. and Zhang, J. (2014), "Martingale difference correlation and its use in high-dimensional variable selection," *Journal of the American Statistical Association,* **109,** 1302-1318.

Sun, Y. (1998), "The almost equivalence of pairwise and mutual independence and the duality with exchangeability", *Probability Theory and Related Fields*, **112**, 425-456.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007), "Measuring and testing dependence by correlation of distances," *Annals of Statistics*, **35**, 2769-2794.

Witola, W., Liu, S., Montpetit, A., Welti, R., Hypolite, M., Roth, M., Zhou, Y., Mui, E., and Cesbron-Delauw, M. (2014), "ALOX12 in Human Toxoplasmosis", *Infection and Immunity* , **82,** 2670–2679.

Xia, Y. (2008), "A semiparametric approach to canonical analysis". *Journal of the Royal Statistical Society Series B*, **70**, 519-543.

Li, D., Ke, Y. and Zhang, W. (2016). Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models. *The Annals of Statistics*, **43**, 2676-2705.

Zheng, S., Shi, N. and Zhang, Z. (2012), "Generalized measures of correlation for asymmetry, nonlinearity and beyond," *Journal of the American Statistical Association,* **107,** 1239-1252.

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.