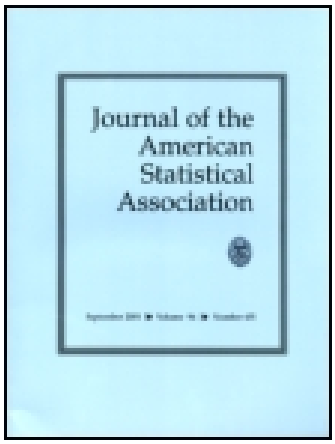

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Martingale Difference Correlation and Its Use in High-Dimensional Variable Screening

Xiaofeng Shao & Jingsi Zhang

Accepted author version posted online: 18 Feb 2014.Published online: 02 Oct 2014.

To cite this article: Xiaofeng Shao & Jingsi Zhang (2014) Martingale Difference Correlation and Its Use in High-Dimensional Variable Screening, Journal of the American Statistical Association, 109:507, 1302-1318, DOI: 10.1080/01621459.2014.887012

To link to this article: http://dx.doi.org/10.1080/01621459.2014.887012

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Martingale Difference Correlation and Its Use in High-Dimensional Variable Screening

Xiaofeng SHAO and Jingsi ZHANG

In this article, we propose a new metric, the so-called martingale difference correlation, to measure the departure of conditional mean independence between a scalar response variable V and a vector predictor variable U. Our metric is a natural extension of distance correlation proposed by Székely, Rizzo, and Bahirov, which is used to measure the dependence between V and U. The martingale difference correlation and its empirical counterpart inherit a number of desirable features of distance correlation and sample distance correlation, such as algebraic simplicity and elegant theoretical properties. We further use martingale difference correlation as a marginal utility to do high-dimensional variable screening to screen out variables that do not contribute to conditional mean of the response given the covariates. Further extension to conditional quantile screening is also described in detail and sure screening properties are rigorously justified. Both simulation results and real data illustrations demonstrate the effectiveness of martingale difference correlation-based screening procedures in comparison with the existing counterparts. Supplementary materials for this article are available online.

KEY WORDS: Conditional mean; Feature screening; High-dimensional inference; Sure screening property.

1. INTRODUCTION

Since the seminal work of Fan and Lv (2008) on sure independence screening (SIS, hereafter), there has been a recent surge of interest on (ultra)high-dimensional variable screening (or feature screening). As the existing variable selection methods (e.g., LASSO, LARS, SCAD) may not perform well when the dimension of predictor variables p is much larger than sample size n, an alternative approach that has been advocated in the literature is to first perform variable screening to reduce the dimensionality p to some moderate scale, and then apply variable selection methods in the second stage. The main goal of this article is to propose a new model-free screening procedure based on a new marginal utility, namely the martingale difference correlation (MDC, hereafter), which measures the conditional mean (in)dependence between two variables. Our procedure can be easily extended to perform conditional quantile screening, that is, screening out variables that do not contribute to conditional quantiles of the response given the covariates, which is useful for analyzing high-dimensional heterogeneous data.

The literature on high-dimensional variable screening has been growing rapidly in recent years. In one direction, great efforts have been made to relax the linear model assumption imposed in Fan and Lv (2008). Fan and Song (2010) proposed a new feature-screening method based on ranking the maximum marginal likelihood estimates in generalized linear models. Within the scope of the ultrahigh-dimensional additive models, Fan, Feng, and Song (2011) fitted a marginal nonparametric regression model for each covariate using B-spline approximation and ranked their importance according to several goodness-of-fit criteria. In another direction, more sophisticated dependence metrics have been applied to detect nonlinear relationship and make the screening procedure less model-dependent; see Hall and Miller (2009), Zhu et al. (2011), and Li, Zhong, and Zhu (2012), among others. In particular, Li, Zhong, and Zhu (2012) introduced a distance correlation (Székely, Rizzo, and Bakirov 2007) based screening procedure (DC-SIS, hereafter), which can handle grouped predictors and/or multivariate responses and demonstrate superior finite sample performance over SIS, its correlation-based counterpart proposed by Fan and Lv (2008). By employing Kendall's τ, Li et al. (2012) constructed a robust rank correlation screening method, which can guard against outliers or influence points. Also see Delaigle and Hall (2012) for a transformation-based variable ranking method to deal with heavy-tailed data. For censored data, various variable screening methods have also been developed; see Fan, Feng, and Wu (2010), Zhao and Li (2012), Gorst-Rasmussen and Scheike (2013), and He, Wang, and Hong (2013) for recent contributions.

In this article, we introduce a new notion called martingale difference correlation to measure the departure of (U, V) from the relationship that

$$\mathbb{E}(V|U) = \mathbb{E}(V) \text{ almost surely}, \tag{1}$$

that is, the conditional mean of V given U is independent of U. Due to the similarity of (1) to the martingale difference concept in probability theory, we call $V - \mathbb{E}(V)$ a martingale difference with respect to U if (1) holds. Our definition of MDC borrows the idea from the recently proposed distance correlation (DC, hereafter), which is used to measure the dependence between two random vectors. Our MDC-based screening procedure is parallel to the DC-based one developed recently by Li, Zhong, and Zhu (2012). In particular, the MDC is defined in a way so that (1) holds if and only if $\text{MDC}(V|U) = 0$ and otherwise $\text{MDC}(V|U) > 0$. Furthermore, the MDC of two variables in a

Xiaofeng Shao is Associate Professor, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL (E-mail: *xshao@illinois.edu*). Jingsi Zhang is Ph.D. student in the Department of Statistics, Northwestern University, Evanston, IL (E-mail: *jingsizhang2016@u.northwestern.edu*). Both authors equally contribute to this article, and the authors are listed in the alphabetical order. Shao's research was partially supported by NSF grants DMS08-04937 and DMS11-04545. The authors thank all the contributors of He, Wang and Hong (2013) for providing the R codes used in their article and for a clarification of their example 3.a in numerical simulations. Thanks also go to Professor Runze Li and Dr. Lukas Meier for providing the R codes used in the papers by Li, Zhong, and Zhu (2012), and Meier, van de Geer, and Bühlmann (2009), respectively. The authors are also grateful to the associate editor and three referees for their constructive comments that led to a substantial improvement of the article.

DOI: 10.1080/01621459.2014.887012

bivariate normal vector is a strictly increasing function of the absolute value of the Pearson correlation of these two normal random variables. Like DC-SIS, these are two key properties that we need to extend the Pearson correlation-based screening (Fan and Lv 2008). However, unlike DC-SIS, the active set of variables the MDC-based screening procedure aims to keep is generally different from the set DC-SIS targets, and this difference will be elaborated later.

One attractive feature of the MDC-based screening procedure is that it can be easily modified to screen out variables that do not contribute to a particular aspect of the conditional distribution of the response given the covariates, such as conditional quantiles. For high-dimensional heterogenous data, He, Wang, and Hong (2013) proposed a new framework called quantile-adaptive model-free screening and advocated a quantile-adaptive approach which allows the set of active variables to be different when modeling different conditional quantiles. The screening procedure in He, Wang, and Hong (2013) involves the estimation of marginal quantile regression model nonparametrically using B-spline approximation and its practical implementation requires the choice of spline basis functions and knots, etc., whereas no tuning parameter is involved in the calculation of sample marginal utilities for our MDC-based conditional quantile screening procedure. On the other hand, it was mentioned in He, Wang, and Hong (2013) that adaptive choice of number of knots and spline basis functions is often not needed as optimal estimation of marginal quantile regression is not the goal in the high-dimensional screening and a standard procedure using cubic spline and 3 or 4 internal knots are flexible enough to approximate many smooth functions encountered in practice, and is seen to work well in many settings. See Section 4.2 for a detailed finite sample comparison between QaSIS of He, Wang, and Hong (2013) and our MDC-SISQ.

In this article, we obtain a number of useful results related to the properties of the MDC and its sample counterpart as well as MDC-based screening procedures. For the latter, we provide a rigorous justification of the sure screening property under a general model-free setup. In particular, our theory allows for the ultrahigh-dimensional case when $p = o(\exp(n^{\alpha}))$ for some $\alpha > 0$ under suitable moment assumptions on the covariates and the response variable. Monte Carlo simulations and real-data illustrations are conducted to evaluate the performance of MDC-based screening procedures in comparison with the counterparts based on other dependence metrics. Our results indicate that the MDC-based screening procedures are comparable to the relatively superior one among alternative methods under examination in most examples and can be superior to SIS and DC-SIS in many model settings.

The rest of this article is organized as follows. Section 2 introduces martingale difference divergence (MDD, hereafter) as an analog of distance covariance (Székely, Rizzo, and Bakirov 2007), martingale difference correlation and their sample versions, as well as their nice properties. In Section 3, we describe the MDC-based screening procedures and discuss their difference from DC-based ones. The sure screening consistency of conditional mean screening and conditional quantile screening is also presented. The results from simulation studies and empirical illustrations using two data examples are reported in Sections 4 and 5, respectively. Section 6 concludes and some technical proofs are presented in Section 6. The remaining technical proofs and additional simulation results are gathered in the supplementary material.

A word on notation. Let $i = \sqrt{-1}$ be the imaginary unit. The scalar product of vectors x and y is denoted by $< x, y >$. For a complex-valued function $f(\cdot)$, the complex conjugate of f is denoted by $\bar{f}$ and $|f|^2 = f\bar{f}$. Denote the Euclidean norm of $x = (x_1, \ldots, x_p) \in \mathbf{C}^p$ as $|x|_p$, where $|x|_p^2 = x_1\bar{x}_1 + \cdots + x_p\bar{x}_p$. A random vector $X \in \mathcal{L}^s$ if $\mathbb{E}|X|_p^s < \infty$. Let $c_p = \frac{\pi^{(1+p)/2}}{\Gamma((1+p)/2)}$. For a complex-valued function $\gamma : \mathbf{R}^q \to \mathbf{C}^p$, we define its norm as $\|\gamma(s)\|^2 = \int_{\mathbf{R}^q} |\gamma(s)|_p^2 (c_q |s|_q^{1+q})^{-1} ds$. The positive constant C is generic and its exact value may vary from place to place. The symbols $\xrightarrow[n\to\infty]{D}$ and $\xrightarrow[n\to\infty]{P}$ signify convergence in distribution and in probability, respectively.

2. MARTINGALE DIFFERENCE DIVERGENCE/ CORRELATION

To introduce the notion of MDD, we shall provide a brief discussion on the following three commonly studied relationships between two random vectors $U \in \mathbf{R}^q$ and $V \in \mathbf{R}$, where q is a fixed integer and does not depend on sample size. We shall restrict our attention to the case V is a scalar random variable for the convenience of notation. This will suffice for our feature screening application and a generalization to allow a random vector V is possible but is not pursued. Consider the following relationships,

$$\operatorname{cov}(U, V) = 0, \tag{2}$$

$$\mathbb{E}(V|U) = \mathbb{E}(V), \text{ or equivalently } \mathbb{E}((V - \mathbb{E}(V))|U) = 0 \text{ almost surely}, \tag{3}$$

$$U \text{ and } V \text{ are independent.} \tag{4}$$

Assume $U, V \in \mathcal{L}^2$, then (4)⇒(3)⇒(2). Thus the relationship (3) lies in between independence and uncorrelatedness. Throughout the article, we say that $V - \mathbb{E}(V)$ is martingale difference with respect to U (or sigma-field generated by U) if (3) holds. This definition of martingale difference for a pair of random vectors is consistent with the definition of martingale difference sequence in probability theory. It is not difficult to find examples from standard textbooks that there are random vectors such that (2) holds but (3) is violated, or (3) holds but (4) is not satisfied. Note that (4) is equivalent to that the conditional distribution of V given $U = u$ is independent of u, whereas (3) says that the conditional mean of V given $U = u$ is independent of u.

To measure the dependence between U and V, Székely, Rizzo, and Bakirov (2007) proposed distance covariance and distance correlation which have very nice properties. We refer the reader to Székely, Rizzo, and Bakirov (2007) and Székely and Rizzo (2009) for a complete discussion. The distance covariance between two random vectors $U \in \mathbf{R}^q$ and $V \in \mathbf{R}^r$ ($r = 1$ in the definition of MDD below) is defined as

$$\operatorname{dcov}(U, V) = \left[\frac{1}{c_r c_q} \int_{\mathbf{R}^{r+q}} \frac{|f_{U,V}(t, s) - f_U(t) f_V(s)|^2}{|t|_q^{1+q} |s|_r^{1+r}} dt ds \right]^{1/2},$$

where $f_{U,V}(t,s) = \mathbb{E}(e^{i(<t,U>+<s,V>)})$, $f_U(t) = \mathbb{E}(e^{i<t,U>})$, and $f_V(s) = \mathbb{E}(e^{i<s,V>})$ denote the joint and marginal characteristic functions. This motivates our definition of MDD.

Definition (Martingale difference divergence). For $U \in \mathbf{R}^q$ and $V \in \mathbf{R}^1$, the martingale difference divergence of V given U is the nonnegative number $\mathrm{MDD}(V|U)$ defined by

$$\mathrm{MDD}(V|U)^2 = \frac{1}{c_q}\int_{\mathbf{R}^q} \frac{|g_{V,U}(s) - g_V g_U(s)|^2}{|s|_q^{1+q}} ds,$$

where $g_{V,U}(s) = \mathbb{E}(Ve^{i<s,U>})$, $g_V = \mathbb{E}(V)$, and $g_U(s) = \mathbb{E}(e^{i<s,U>})$.

Since in general $\mathrm{MDD}(U|V) \neq \mathrm{MDD}(V|U)$, we name it divergence instead of distance. The number $\mathrm{MDD}(V|U)$ can be larger than 1 and it is not invariant to the scale change of U or V in that $\mathrm{MDD}(cV|U) = |c|\mathrm{MDD}(V|U)$ and $\mathrm{MDD}(V|cU) = \sqrt{|c|}\mathrm{MDD}(V|U)$ for any $c \in \mathbf{R}$. This motivates us to define the martingale difference correlation of V given U as a nonnegative number given by

$$\mathrm{MDC}(V|U)^2 = \begin{cases} \dfrac{\mathrm{MDD}(V|U)^2}{\sqrt{\mathrm{var(V)^2 dvar(U)^2}}} & \text{if } \mathrm{var}(V)^2\mathrm{dvar}(U)^2 > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathrm{dvar}(U) = \mathrm{dcov}(U,U)$ is the distance variance of U. It can be easily shown that $\mathrm{MDC}(V|U)$ is invariant to the scale change of U or V, like the Pearson correlation and distance correlation.

To introduce the properties of $\mathrm{MDD}(V|U)$ and $\mathrm{MDC}(V|U)$, we need more definitions from Székely and Rizzo (2009). If U is an $\mathbf{R}^q$-valued random variable, and $\mathcal{X}(s)$ is a random process (random field) defined for all $s \in \mathbf{R}^q$ and independent of U, define the $\mathcal{X}$-centered version of U by $U_{\mathcal{X}} = \mathcal{X}(U) - E[\mathcal{X}(U)|\mathcal{X}]$, whenever the conditional expectation exists. Furthermore, if V is an $\mathbf{R}^r$-valued random variable, and $\mathcal{Y}(t)$ is a random process defined for all $t \in \mathbf{R}^r$, then the $(\mathcal{X},\mathcal{Y})$ covariance of (U,V) is the nonnegative number whose square is: $\mathrm{cov}^2_{\mathcal{X},\mathcal{Y}}(U,V) = E(U_{\mathcal{X}}U'_{\mathcal{X}}V_{\mathcal{Y}}V'_{\mathcal{Y}})$. Throughout the article, (U',V'), (U'',V'') are iid copies of (U,V). Let $B(s), s \in \mathbf{R}^q$ denote a mean zero Gaussian process with covariance function $|s|_q + |t|_q - |s-t|_q$, and id denote the simple identity function. The expression E_U means that the expectation is taken with respect to U. We present some important properties of $\mathrm{MDD}(V|U)$ and $\mathrm{MDC}(V|U)$ as follows.

Theorem 1.

1. If $E(|V|^2 + |U|_q^2) < \infty$, then $\mathrm{MDD}(V|U) = \mathrm{cov}_{B,id}(U,V) = \{-E[(V-EV)(V'-EV')|U-U'|_q]\}^{1/2}$.
2. If $E(|V|^2 + |U|_q^2) < \infty$, then $0 \le \mathrm{MDC}(V|U) \le 1$ and $\mathrm{MDC}(V|U) = 0$ if and only if $E(V|U) = E(V)$ almost surely, that is, (1) holds.
3. If U and V are bivariate standard normal with correlation $\rho = \rho(U,V)$, then $\mathrm{MDC}(V|U)^2 = \rho^2/\sqrt{4(1-\sqrt{3}+\pi/3)} \approx 0.891\rho^2$.
4. $\mathrm{MDC}(a_1 + b_1 V|a_2 + b_2 CU) = \mathrm{MDC}(V|U)$ for any $a_1 \in \mathbf{R}^1$, $a_2 \in \mathbf{R}^q$, scalars b_1, b_2 and orthonormal matrix C.

Remark 1. In general, $\mathrm{MDC}(V|U) \neq \mathrm{MDC}(U|V)$ for $V \in \mathbf{R}^1$ and $U \in \mathbf{R}^1$, since the roles of V and U are asymmetric in the definition of MDC. Also $\mathrm{MDC}(V|V) \neq 1$ in general, so MDC does not inherit all the properties of distance correlation. For the purpose of high-dimensional screening, the second and third properties (i.e., MDC of a bivariate normal vector is a monotone function of the magnitude of correlation coefficient) in Theorem 1 are crucial, as pointed out in Li, Zhong, and Zhu (2012).

Given the iid observations $(U_k, V_k)_{k=1}^n$ from the joint distribution of (U,V), we define $\bar{V}_n = n^{-1}\sum_{k=1}^n V_k$, $a_{kl} = V_k V_l$, $\bar{a}_{k\cdot} = n^{-1}\sum_{l=1}^n a_{kl} = V_k\bar{V}_n$, $\bar{a}_{\cdot l} = n^{-1}\sum_{k=1}^n a_{kl} = \bar{V}_n V_l$, $\bar{a}_{\cdot\cdot} = n^{-2}\sum_{k,l=1}^n a_{kl} = \bar{V}_n\bar{V}_n$ and $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot} = (V_k - \bar{V}_n)(V_l - \bar{V}_n)$ for $k,l = 1,\ldots,n$. Similarly, let $b_{kl} = |U_k - U_l|_q$, $\bar{b}_{k\cdot} = n^{-1}\sum_{l=1}^n b_{kl}$, $\bar{b}_{\cdot l} = n^{-1}\sum_{k=1}^n b_{kl}$, $\bar{b}_{\cdot\cdot} = n^{-2}\sum_{k,l=1}^n b_{kl}$ and $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$, for $k, l = 1,\ldots,n$. Similar to the definition of the sample distance covariance/correlation, we define the sample martingale difference divergence/correlation.

Definition. The sample martingale difference divergence $\mathrm{MDD}_n(V|U)$ is the nonnegative number defined by

$$\mathrm{MDD}_n(V|U)^2 = -n^{-2}\sum_{k,l=1}^n A_{kl}B_{kl},$$

and the sample martingale difference correlation $\mathrm{MDC}_n(V|U)$ is the nonnegative number defined by

$$\mathrm{MDC}_n(V|U)^2 = \begin{cases} \dfrac{\mathrm{MDD}_n(V|U)^2}{\sqrt{\mathrm{var}_n(V)^2\cdot\mathrm{dvar}_n(U)^2}}, & \mathrm{var}_n(V)^2\mathrm{dvar}_n(U)^2 > 0; \\ 0, & \mathrm{var}_n(V)^2\mathrm{dvar}_n(U)^2 = 0, \end{cases}$$

where $\mathrm{var}_n(V) = n^{-1}\sum_{k=1}^n (V_k - \bar{V}_n)^2 = (n^{-2}\sum_{k,l=1}^n A_{kl}^2)^{1/2}$ is the sample variance and $\mathrm{dvar}_n(U) = (n^{-2}\sum_{k,l=1}^n B_{kl}^2)^{1/2}$ is the sample distance variance. At some places, we also use $\mathrm{MDC}_n((V_k)_{k=1}^n|(U_k)_{k=1}^n)$ to denote the sample MDC based on the data $(U_k, V_k)_{k=1}^n$.

Let $g^n_{V,U}(s) = n^{-1}\sum_{k=1}^n V_k \exp\{i<s,U_k>\}$, $g^n_V = \bar{V}_n$ and $g^n_U(s) = n^{-1}\sum_{k=1}^n \exp\{i<s,U_k>\}$. The following theorem shows that the definition for $\mathrm{MDD}_n(V|U)$ is natural in the sense that it equals to the empirical plug-in version of $\mathrm{MDD}(V|U)$.

Theorem 2. For a given random sample $(U_k, V_k)_{k=1}^n$ from the joint distribution of (U,V), we have that

$$\left\| g^n_{V,U}(s) - g^n_V g^n_U(s) \right\|^2 = -n^{-2}\sum_{k,l=1}^n A_{kl}B_{kl}.$$

By definition, $0 \le \mathrm{MDC}_n(V|U) \le 1$, where the latter inequality is a direct consequence of the Cauchy–Schwarz inequality. The following theorem states the consistency of $\mathrm{MDD}_n(V|U)$ ($\mathrm{MDC}_n(V|U)$) as an estimator of $\mathrm{MDD}(V|U)$ ($\mathrm{MDC}(V|U)$). It is analogous to Theorem 2 and Corollary 1 in Székely, Rizzo, and Bakirov (2007).

Theorem 3. (Consistency) If $\mathbb{E}(|U|_q + |V|^2) < \infty$, then almost surely

$$\lim_{n\to\infty}\mathrm{MDD}_n(V|U) = \mathrm{MDD}(V|U) \text{and} \lim_{n\to\infty}\mathrm{MDC}_n(V|U) = \mathrm{MDC}(V|U).$$

Let $\Gamma(\cdot)$ denote a complex-valued zero-mean Gaussian random process with covariance function: $\mathrm{cov}_\Gamma(s, s_0) = \overline{F(s-s_0) - g_U(s-s_0)\cdot(E(V))^2} + \{E(V^2) + (E(V))^2\}g_U(s) \overline{g_U(s_0) - F(s)g_U(s_0) - g_U(s)F(s_0)}$, where $s, s_0 \in \mathbf{R}^q$ and $F(s) = E[V^2 \exp(i < U, s >)]$.

Theorem 4. (Weak convergence) Assume that $E(|U|_q + |V|^2) < \infty$. If $\mathrm{MDC}(V|U) = 0$, then (a),

$$n\mathrm{MDD}_n^2(V|U) \xrightarrow[n\to\infty]{D} ||\Gamma(s)||^2;$$

(b), under the additional assumption that $E(V^2|U) = E(V^2)$, $n\mathrm{MDD}_n^2(V|U)/S_n \xrightarrow[n\to\infty]{D} Q$, where Q is a nonnegative quadratic form of centered Gaussian random variable with $E(Q) = 1$ and $S_n = \frac{1}{n^2}\sum_k \sum_l |U_k - U_l|_q \frac{1}{n}\sum_k (V_k - \bar{V}_n)^2$.

If $\mathrm{MDC}(V|U) \neq 0$, then (c), $n\mathrm{MDD}_n^2(V|U)/S_n \xrightarrow[n\to\infty]{P} \infty$.

Theorem 4 is similar to Theorem 5 and Corollary 2 in Székely, Rizzo, and Bakirov (2007). It is worth noting that the MDD and its standardized version MDC are not the only way to quantify the departure from the martingale difference relationship (1). Our development is influenced by the distance covariance and distance correlation developed by Székely, Rizzo, and Bakirov (2007), since in the latter article, the distance covariance and distance correlation as well as their sample counterparts have been demonstrated to have a number of desirable properties. As expected, our MDD, MDC, and their sample counterparts inherit most of useful properties from distance covariance, distance correlation, and their sample versions, and thus it can be considered as a natural generalization.

The idea of quantifying the departure of conditional mean independence can be considerably generalized. For example, we can similarly define corresponding quantities that measure the departure from the relationship that U does not contribute to conditional quantiles of V. For a given quantile level $\tau \in (0, 1)$, we say that the conditional quantile of V given U does not depend on U at the τth quantile level if

$$Q_\tau(V|U = u) = Q_\tau(V) \text{ for any } u,$$

where $Q_\tau(V)$ and $Q_\tau(V|\cdot)$ denote the unconditional and conditional quantile of V at the τth quantile level, respectively. Note that the above conditional quantile independence between V and U at the quantile level τ is equivalent to $\mathbb{E}(\tau - \mathbf{1}(V - Q_\tau(V) \le 0)|U = u) = 0$ for any u, that is,

$$E(W|U) = E(W) = 0, \text{ where } W = \tau - \mathbf{1}(V - Q_\tau(V) \le 0),$$

which is further equivalent to $\mathrm{MDC}(W|U) = 0$. Similarly, we can define the metric that quantifies the departure from the relationship $\mathbb{E}(f(V, h(F_V))|U) = \mathbb{E}(f(V, h(F_V)))$, where f is a prespecified function, h is a functional, and F_V is the distribution function of V. For example, conditional quantile independence becomes a special case by letting $h(F_V) = Q_\tau(V)$ and $f(v, Q_\tau(V)) = \tau - \mathbf{1}(v \le Q_\tau(V))$.

3. MDC-BASED VARIABLE SCREENING

Let $Y \in \mathbf{R}^1$ be the univariate continuous response variable with support Ψ_Y and $X = \{X_j\}_{j=1}^p$ be the predictor variables. We consider the case when p greatly exceeds the sample size n, that is, high-dimensional setting. In Li, Zhong, and Zhu (2012), they used distance correlation as a marginal utility to perform independence ranking and feature screening. The main focus is on the conditional distribution function of Y given X, that is, $F_Y(y|X = x)$. Without specifying a model, they define the active and inactive predictor sets by

$$\mathcal{D}_F = \{j : F_Y(y|X) \text{ functionally depends on} X_j \text{ for some} y \in \Psi_Y\}$$

and

$$\mathcal{I}_F = \{j : F_Y(y|X) \text{ does not functionally depend on } X_j \text{ for any} y \in \Psi_Y\}.$$

As Cook and Li (2002) stated, in many situations regression analysis is mostly concerned with inferring about the conditional mean of the response given the predictors, and less concerned with the other aspects of the conditional distribution. Instead of focusing on $F_Y(y|X)$, we restrict our attention to the conditional mean $\mathbb{E}(Y|X)$ and define

$$\mathcal{D}_E = \{j : \mathbb{E}(Y|X) \text{ depends on} X_j\} \text{ and}$$
$$\mathcal{I}_E = \{j : \mathbb{E}(Y|X) \text{ does not depend on} X_j\}.$$

In general, it is easy to see that $\mathcal{D}_E \subset \mathcal{D}_F$, thus a direct comparison between DC-based and MDC-based screening procedures (MDC-SIS, hereafter) may not be meaningful, unless in the case $\mathcal{D}_E = \mathcal{D}_F$ or for some particular subsets of predictor variables in $\mathcal{D}_E \cap \mathcal{D}_F$. Write $X_{\mathcal{D}_E} = \{X_j : j \in \mathcal{D}_E\}$ and $X_{\mathcal{I}_E} = \{X_j : j \in \mathcal{I}_E\}$ and refer to $X_{\mathcal{D}_E}$ as an active predictor vector and its complement $X_{\mathcal{I}_E}$ as an inactive predictor vector. It implies that

$$\mathbb{E}(Y|X_{\mathcal{D}_E}) = \mathbb{E}(Y|X_{\mathcal{D}_E}, X_j),\ j \in \mathcal{I}_E.$$

For each X_j, $j = 1, \ldots, p$, and given the sample $(X_{1k}, \ldots, X_{pk}, Y_k)_{k=1}^n$, we define $\mathrm{MDD}^j = \mathrm{MDD}(Y|X_j)$, $\mathrm{MDC}^j = \mathrm{MDC}(Y|X_j)$ (corresponding to $q = 1$ in previous section), as well as their sample counterparts $\mathrm{MDD}_n^j = \mathrm{MDD}_n((Y_k)_{k=1}^n|(X_{jk})_{k=1}^n)$ and $\mathrm{MDC}_n^j = \mathrm{MDC}_n((Y_k)_{k=1}^n|(X_{jk})_{k=1}^n)$. For the ease of presentation, we write $\omega_j = (\mathrm{MDC}^j)^2$ and $\widehat{\omega}_j = (\mathrm{MDC}_n^j)^2$. At the population level, we use ω_j (or MDC^j) as the marginal utility to rank the importance of X_j, whereas at the sample level, we select a set of important predictors with large sample MDCs, that is, we define

$$\hat{\mathcal{D}}_E = \{j : \widehat{\omega}_j \ge cn^{-\kappa}, \text{for } 1 \le j \le p\},$$

where c and κ are prespecified threshold values defined in Assumption (A2).

In Zhu et al. (2011), another model-free feature-screening procedure SIRS was proposed and we provide a brief discussion on its difference from the DC and MDC-based counterparts. For two univariate random variables X (assumed to have mean zero) and Y, let $\Omega(y) = \mathrm{cov}(X, \mathbf{1}(Y < y))$ and $\gamma = \gamma(X, Y) = \mathbb{E}(\Omega^2(Y))$. For a set of predictor variables, $(X_1, \ldots, X_p)$, the marginal utility of SIRS is $\gamma_j = \mathbb{E}(\Omega_j^2(Y))$, where $\Omega_j(y) = \mathrm{cov}(X_j, \mathbf{1}(Y < y))$. Under suitable assumptions, it can be shown that $\gamma = 0$ if and only if $\mathbb{E}(X|Y) = \mathbb{E}(X)$ almost surely. Thus SIRS tries to screen out variables whose conditional mean given the response does not depend on the response. When X and Y are independent, both $\mathrm{MDC}(Y|X)$ and

$\gamma(X, Y)$ equal to zero, thus both MDC-SIS and SIRS can be regarded as a kind of independence screening procedure, although their interpretations are different and they may target different sets of variables (i.e., the active predictor vectors based on SIRS and MDC-SIS may be different). Also there seems no natural extension of SIRS to do conditional quantile screening, whereas the MDC-based conditional quantile screening can be readily done as described below.

The idea of focusing on the conditional mean of Y given X is not new. In the literature of sufficient dimension reduction, a central mean subspace is defined to be the intersection of dimension reduction spaces for conditional mean, as an analog of central subspace which aims for the minimal reduction of conditional distribution. As mentioned in Cook and Li (2002), "Pursuing the mean function through the central subspace can be inefficient because the scope of the statistical inquiry may be much larger than necessary." See Li, Cook, and Chiaromonte (2003) for methods of finding a central mean space. For sufficient dimensional reduction, it seems that the dimension p is typically fixed or is only allowed to grow slowly with sample size n. In our variable screening problem, the dimension can be ultrahigh relative to sample size, and we aim to reduce the dimensionality by screening out irrelevant variables.

Our idea can be generalized to screen out variables that do not contribute to conditional quantiles. Specifically, suppose we want to leave out variables that do not contribute to the conditional quantile of Y given X at the τth quantile level. Let $\mathcal{D}_{Q(\tau)} = \{j : Q_\tau(Y|X) \text{ depends on} X_j\}$ and $\mathcal{I}_{Q(\tau)} = \{j : Q_\tau(Y|X) \text{does not depend on} X_j\}$. Then

$$Q_\tau(Y|X_{\mathcal{D}_{Q(\tau)}}) = Q_\tau(Y|X_{\mathcal{D}_{Q(\tau)}}, X_j),\ j \in \mathcal{I}_{Q(\tau)}.$$

Thus the marginal utility is $\mathrm{MDC}(W|X_j)$, where $W = \tau - \mathbf{1}(Y \le Q_\tau(Y))$, and its sample estimate is $\mathrm{MDC}_n((\widehat{W}_k)_{k=1}^n|(X_{jk})_{k=1}^n)$, where $\widehat{W}_k = \tau - \mathbf{1}(Y_k \le \hat{Q}_\tau)$ and $\hat{Q}_\tau$ is the τth sample quantile of Y. Throughout the article, we use MDC-SISQ to denote MDC-based conditional quantile screening procedure. Like the DC-based screening, we do not assume a (parametric or nonparametric) model, so our screening procedure is model-free. What is different from Li, Zhong, and Zhu (2012) is that we focus on a specific aspect of the conditional distribution of the response given the covariates (e.g., its conditional mean), which is considerably less ambitious and is often directly linked to the second stage of constructing a parametric or nonparametric model for conditional mean or conditional quantiles.

Generally speaking, one can also apply the correlation-based or DC-based metric to the transformed data. In the case of conditional quantile screening, we can define the SISQ and DC-SISQ, which are correlation (distance correlation) based screening procedures for conditional quantile screening. Specifically, let $\mathrm{corr}_n(X_j, \widehat{W})$ ($dCor_n(X_j, \widehat{W})$) denote the sample correlation (distance correlation) based on the data $(X_{jk}, \widehat{W}_k)_{k=1}^n$. Then we rank the importance of predictor variables according to the magnitude of sample marginal utilities. It turns out that for conditional quantile screening, DC-SISQ and MDC-SISQ deliver the same ranking, as shown in the following proposition.

Proposition 1. For any variables (X, Y) with sample replicates $(X_k, Y_k)_{k=1}^n$, let $\widehat{W} = \tau - \mathbf{1}(Y \le \hat{Q}_\tau)$, where $\hat{Q}_\tau$ is the τth sample quantile for $\tau \in (0, 1)$. Then $\mathrm{MDC}_n(\widehat{W}|X) = dCor_n(X, \widehat{W})$ provided that $F_{Yn}(\hat{Q}_\tau) = \tau$, where F_{Yn} is the sample distribution function of the continuous response variable Y.

We proceed to show the sure screening consistency of our MDC-based procedure for conditional mean screening. To this end, we assume the following conditions.

A1. There exists a positive constant s_0 such that for all $0 < s \le 2s_0$,

$$\sup_p \max_{1\le j\le p} E\left\{\exp\left(sX_j^2\right)\right\} < \infty, \text{ and } E\{\exp(sY^2)\} < \infty.$$

A2. The minimum MDC value of active predictors satisfies that

$$\min_{j\in\mathcal{D}_E} \omega_j \ge 2cn^{-\kappa}, \text{ for some constant } c > 0 \text{ and } 0 \le \kappa < \frac{1}{2}.$$

Under Assumption A2, we can separate the active variables from their inactive counterparts using the marginal utility at the population level. Since under A1, the sample marginal utilities can be shown to concentrate around their population counterparts uniformly with an overwhelming probability, the sure screening consistency thus follows. Below we provide a formal statement.

Theorem 5. Under Assumption A1, for any $0 < \gamma < 1/2 - \kappa$, there exist positive constants $c_1 > 0$ and $c_2 > 0$ such that

$$P\left\{\max_{1\le j\le p} |\widehat{\omega}_j - \omega_j| \ge cn^{-\kappa}\right\} \le O\big(p\big[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp\left(-c_2 n^{\gamma/2}\right)\big]\big) \quad (5)$$

Under conditions A1 and A2, we have that

$$P(\mathcal{D}_E \subset \hat{\mathcal{D}}_E) \ge 1 - O(s_n[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma/2})]),$$

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

Thus the sure screening property holds for MDC-SIS under reasonable moment assumptions on Y and X_j (see (A1)), which allow for Gaussian distribution. To balance the two terms in the right-hand side of (5), we can choose the optimal $\gamma = (2 - 4\kappa)/5$, which means that we can handle the NP-dimensionality of order $\log(p) = o(n^{(1-2\kappa)/5})$. With some stronger moment assumption on Y, for example, assume that $E(\exp(sY^4)) < \infty$ for all $s \in (0, 2s_0]$, we can handle $\log(p) = o(n^{(1-2\kappa)/3})$, which is the optimal rate obtained by Li, Zhong, and Zhu (2012) under sub-Gaussian type moment assumptions on Y and X_j.

Next we shall show the sure screening consistency of MDC-SISQ. Recall that $W_k = \tau - I(Y_k \le Q_\tau(Y))$ and $\widehat{W}_k = \tau - I(Y_k \le \widehat{Q}_\tau)$, where $\widehat{Q}_\tau$ is the τth sample quantile based on $(Y_k)_{k=1}^n$. Correspondingly, we define MDD, MDC, and their sample values based on $(X_{jk}, W_k)_{k=1}^n$ and $(X_{jk}, \widehat{W}_k)_{k=1}^n$ as $\mathrm{MDD}^j(W)$, $\mathrm{MDC}^j(W)$, $\mathrm{MDD}_n^j(W)$, $\mathrm{MDC}_n^j(W)$, $\mathrm{MDD}_n^j(\widehat{W})$, and $\mathrm{MDC}_n^j(\widehat{W})$. Write $\omega_j(W) = [\mathrm{MDC}^j(W)]^2$, $\widehat{\omega}_j(W) = [\mathrm{MDC}_n^j(W)]^2$, $\widehat{\omega}_j(\widehat{W}) = [\mathrm{MDC}_n^j(\widehat{W})]^2$. The following assumptions are needed.

B1. The cdf of the continuous response variable Y, F_Y is continuously differentiable in a small

neighborhood of $Q_\tau = Q_\tau(Y)$, say $[Q_\tau - \delta_0, Q_\tau + \delta_0]$, $\delta_0 > 0$. Let $G_1(\delta_0) = \inf_{y\in[Q_\tau-\delta_0, Q_\tau+\delta_0]} f_Y(y)$ and $G_2(\delta_0) = \sup_{y\in[Q_\tau-\delta_0, Q_\tau+\delta_0]} f_Y(y)$, where f_Y is the density function of Y. Assume that $0 < G_1(\delta_0) \le G_2(\delta_0) < \infty$.

B2. There exists a positive constant s_0 such that for all $0 < s \le 2s_0$, $\sup_p \max_{1\le j\le p} E[\exp(sX_j^2)] < \infty$.

B3. The minimum MDC value of active predictors satisfies that

$$\min_{j\in\mathcal{D}_{Q(\tau)}} \omega_j(W) \ge 2cn^{-\kappa},$$

for some constant $c > 0$ and $0 \le \kappa < \frac{1}{4}$.

Assumption is quite mild and it can be satisfied by most commonly used continuous distributions. Define the selected set of important predictor variables

$$\widehat{\mathcal{D}}_{Q(\tau)} = \{j : \widehat{\omega}_j(\widehat{W}) \ge cn^{-\kappa} \text{ for } 1 \le j \le p\},$$

where c and κ are specified in Assumption. As an intermediate step toward proving the sure screening consistency of MDC-SISQ, we obtain the following result, which may be of independent interest.

Proposition 2. Suppose the distribution function of Y satisfies Assumption B1, then there exists $\epsilon_0 > 0$ and $c_1 > 0$, such that for any $\epsilon \in (0, \epsilon_0)$,

$$P\left(\frac{1}{n}\sum_{l=1}^{n} |\widehat{W}_l - W_l| > \epsilon\right) \le 3\exp(-2nc_1\epsilon^2).$$

Theorem 6. Under Assumptions B1 and B2, for any $0 < \gamma < 1/2 - 2\kappa$ and $\kappa \in (0, 1/4)$, there exist positive constants c_1, c_2 such that for any $c > 0$,

$$P\left(\max_{1\le j\le p} |\widehat{\omega}_j(\widehat{W}) - \omega_j(W)| \ge cn^{-\kappa}\right) \le O\left(p\left[\exp\left\{-c_1 n^{1-2(2\kappa+\gamma)}\right\} + n\exp(-c_2 n^{\gamma})\right]\right).$$

With the additional Assumption B3, we can show that

$$P\left(\mathcal{D}_{Q(\tau)} \subseteq \widehat{\mathcal{D}}_{Q(\tau)}\right) \ge 1 - O\left(\widetilde{s_n}\left[\exp\left\{-c_1 n^{1-2(2\kappa+\gamma)}\right\} + n\exp(-c_2 n^{\gamma})\right]\right),$$

where $\widetilde{s_n}$ is the cardinality of $\mathcal{D}_{Q(\tau)}$.

The optimal order $\gamma = (1 - 4\kappa)/3$ can be chosen to balance the two exponents in the bound above, which makes it $O(\max(p, n)\exp\{-c_1 n^{(1-4\kappa)/3}\})$ for some $c_1 > 0$. Then our theory allows for the NP-dimensionality of order $\log(p) = o(n^{(1-4\kappa)/3})$. Note that the optimal order $\gamma = (1 - 4\kappa)/3$ is inferior to that corresponds to the DC-based screening, where $\gamma = (1 - 2\kappa)/3$. This is mainly due to a seemingly necessary technical consideration that is caused by the fact that the response $(W_l)_{l=1}^n$ are not observed but estimated by the plug-in version $(\widehat{W}_l)_{l=1}^n$. The asymptotic negligibility of estimation effect holds under a slightly more restrictive condition on the growth rate of p relative to n.

4. SIMULATION STUDIES

In this section, we assess the finite sample performance of the proposed MDC-based screening procedures in comparison with SIS- and DC-based counterparts, SIRS in Zhu et al. (2011), NIS in Fan, Feng, and Song (2011) as well as QaSIS in He, Wang, and Hong (2013) via Monte Carlo simulations. We repeat each experiment 500 times and consider three criteria for evaluating the performance following Li, Zhong, and Zhu (2012).

1. $\mathcal{S}$: the minimum model size to include all active predictors. We report the 5%, 25%, 50%, 75%, and 95% quantiles of $\mathcal{S}$ out of 500 replications.
2. $\mathcal{P}_s$: the proportion that an individual (active) predictor is selected for a given model size d in the 500 replications.
3. $\mathcal{P}_a$: the proportion that all active predictors are selected for a given model size d in the 500 replications.

The minimum model size $\mathcal{S}$ measures the ability of a screening procedure of including all the active predictors, whereas $\mathcal{P}_s$ and $\mathcal{P}_a$ serve a similar purpose but allow us to examine the screening performance for an individual predictor variable and all active predictors for a given model size d. For conditional quantile screening, the criterion $\mathcal{P}_s$ allows us to reveal possible heterogeneity in the data since the active sets may be different for different quantiles. A screening procedure is deemed competent if it yields an $\mathcal{S}$ value that is close to the true number of active predictors, and $\mathcal{P}_s$ and $\mathcal{P}_a$ values that are close to 1. It is our experience that these three criteria are in general correlated with each other. A small $\mathcal{S}$ tends to be associated with high proportions for $\mathcal{P}_a$ and $\mathcal{P}_s$. So we present the results based on only one or two criteria in some cases to conserve space.

4.1 Conditional Mean Screening

The examples presented in this section are designed to compare the finite sample performance of the MDC-SIS with the SIS, SIRS, NIS, and DC-SIS for the purpose of conditional mean screening.

Example 1. We adopt the simple linear model from Fan and Lv (2008): $Y = 5X_1 + 5X_2 + 5X_3 + \epsilon$. The predictor vector $(X_1, \ldots, X_p)$ is drawn from a multivariate normal distribution $N(0, \Sigma)$ whose covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$ has entries $\sigma_{ii} = 1$, $i = 1, \ldots, p$, and $\sigma_{ij} = \rho$, $i \ne j$. The error term ϵ is independently generated from the standard normal distribution. We consider several different combinations of (p, n, ρ), that is, $p = 100,\ 1000$, $n = 20,\ 50,\ 70$, and $\rho = 0,\ 0.1,\ 0.5,\ 0.9$.

In Table 1, we report the $\mathcal{P}_a$ values calculated from Example 1 for four methods: SIS, MDC-SIS, DC-SIS, and SIRS, with $d = n$. It can be seen that high collinearity (increasing ρ) as well as high dimensionality lead to worse performance for all four methods. It might not be surprising to see that SIS always performs the best, followed by MDC-SIS, and then DC-SIS and SIRS. The difference among them becomes smaller when the sample size gets larger for a fixed p. It suggests that although MDC-SIS, DC-SIS, and SIRS are all capable of detecting the linear relationship, MDC-SIS exhibits a better screening ability for linear models than DC-SIS and SIRS. The fact that SIS outperforms all other methods for normal linear models may be explained by the efficiency of Pearson correlation to characterize the dependence between two normal random variables.

In the supplementary material, we present the $\mathcal{P}_a$ values corresponding to $d = 5, 10, 15$ and $p = 1000$ in Table 11, those corresponding to $d = n$ and $p = 1000$ for models with varying

Table 1. $\mathcal{P}_a$ for Example 1 with $d = n$

p	n	Method	Results for the following values of ρ			
			$\rho = 0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
100	20	SIS	0.736	0.814	0.676	0.592
		DC-SIS	0.658	0.752	0.602	0.490
		MDC-SIS	0.708	0.770	0.630	0.528
		SIRS	0.666	0.742	0.522	0.198
	50	SIS	1	1	0.998	1
		DC-SIS	1	1	0.996	0.994
		MDC-SIS	1	1	0.994	0.994
		SIRS	1	1	0.994	0.894
1000	20	SIS	0.212	0.224	0.144	0.092
		DC-SIS	0.140	0.168	0.096	0.060
		MDC-SIS	0.174	0.212	0.116	0.062
		SIRS	0.112	0.132	0.068	0.008
	50	SIS	0.980	0.978	0.854	0.810
		DC-SIS	0.978	0.962	0.794	0.700
		MDC-SIS	0.978	0.972	0.822	0.718
		SIRS	0.966	0.954	0.728	0.226
	70	SIS	1	1	0.982	0.956
		DC-SIS	0.998	0.998	0.952	0.898
		MDC-SIS	1	0.996	0.964	0.920
		SIRS	0.998	0.998	0.926	0.468

signal-to-noise ratios (SNR, hereafter) in Table 12, and those for the case $p = 3000$ and $d = \lfloor n/\log(n) \rfloor$ in Table 13. The models with different SNRs were obtained by adjusting the constant c in the model $Y = c(5X_1 + 5X_2 + 5X_3) + \epsilon$, following example 1 in Zhu et al. (2011). Basically the same phenomenon occurs with the performance of methods ranked in the order of SIS, MDC-SIS, DC-SIS, and SIRS. It appears that for all three cases (with smaller d or for models with smaller SNR or for larger p) when the covariates are moderately or highly dependent (i.e., $\rho = 0.5, 0.9$), SIRS is substantially outperformed by the other three methods. When the signal-to-noise ratio is too small (say, SNR = 1) and the dependence is too strong among covariates (say, $\rho = 0.9$), none of the methods do well in terms of including the active predictor variables.

Example 2. In this example, we consider two nonlinear additive models, which have been analyzed in Meier, van de Geer, and Bühlmann (2009), and Fan, Feng, and Song (2011). Let $g_1(x) = x$, $g_2(x) = (2x-1)^2$, $g_3(x) = \sin(2\pi x)/(2 - \sin(2\pi x))$, and $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)$. The following cases are studied:

Case 2a: $Y = 5g_1(X_1) + 3g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sqrt{1.74}\epsilon$, where the covariates X_j, $j = 1, \ldots, p$ are simulated according to iid Unif(0,1), and ϵ is independent from the covariates and follows the standard normal distribution.

Case 2b: The covariates and the error term are simulated as in Case 2a, but the model structure is more involved with eight additional predictor variables. $Y = g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4) + 1.5g_1(X_5) + 1.5g_2(X_6) + 1.5g_3(X_7) + 1.5g_4(X_8) + 2g_1(X_9) + 2g_2(X_{10}) + 2g_3(X_{11}) + 2g_4(X_{12}) + \sqrt{0.5184}\epsilon$.

Tables 2 and 3 depict the simulation results of $\mathcal{S}$, $\mathcal{P}_s$, and $\mathcal{P}_a$ for Example 2 with $n = 400$, $p = 1000$, and $d = \lfloor n/\log(n) \rfloor$, 20, and 30. When the true model is nonlinear and the distribution of the variables is non-Gaussian, MDC-SIS performs significantly better than SIS and slightly outperforms DC-SIS in this particular example in terms of smaller minimum model size and higher proportion of including each active predictor variable. The performance of NIS of Fan, Feng, and Song (2011) and SIRS of Zhu et al. (2011) are both inferior to MDC-SIS in this case. For Case 2b, both MDC-SIS and DC-SIS behave comparably well, whereas SIS, NIS, and SIRS fail badly in identifying X_2, X_6, X_{10} for all ds under examination. Note that in this example, $D_E = D_F$, that is, the active predictor vectors for MDC-SIS and DC-SIS are identical.

In the supplementary material, we present the simulation results of $\mathcal{S}$, $\mathcal{P}_s$, and $\mathcal{P}_a$ for Example 2 with $n = 200$, $p = 2000$, and $d = \lfloor n/\log n \rfloor$ in Tables 14 and 15. Again for both cases, MDC-SIS and DC-SIS apparently outperform all the other three methods in terms of smaller minimum model size and higher proportion of including each active predictor variable.

Example 3. To demonstrate that DC-based and MDC-based feature screening methods aim at different active predictors, we also consider the following two additive models with heteroscedastic errors:

Case 3a: $Y = X_1 + X_2 + X_3^2 + \exp((X_4 + X_5)/2)\epsilon$,
Case 3b: $Y = X_1 + X_2 + X_3^2 + \exp(2(X_4 + X_5 + X_6)/5)\epsilon$,

where the predictor vector $(X_1, \ldots, X_p)$ is drawn from multivariate normal distribution with entries of covariance matrix being $\sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.5$ and the error term ϵ is generated independently from the predictor and follows

Table 2. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ for Example 2 with $n = 400$ and $p = 1000$

	Method	5%	25%	50%	75%	95%
2.a	SIS	49.80	262.00	489.00	753.75	950.15
	DC-SIS	6.00	19.00	43.00	94.00	270.75
	MDC-SIS	5.95	15.00	31.00	64.00	200.20
	NIS	12.00	213.75	503.00	790.25	974.15
	SIRS	52.90	252.75	524.00	762.25	947.20
2.b	SIS	413.60	642.75	799.00	921.00	983.05
	DC-SIS	48.00	116.00	224.50	391.75	758.15
	MDC-SIS	45.90	104.50	184.00	322.00	698.05
	NIS	632.90	845.25	937.50	981.00	999.00
	SIRS	447.90	671.00	819.00	912.25	976.05

the standard normal distribution. We set $n = 200$, $p = 2000$, and $d = 20, 30, \lfloor n/\log(n) \rfloor$ for these two cases. Note that the active sets for DC-SIS and MDC-SIS are different, where $D_F = \{1, 2, 3, 4, 5\}$ and $D_E = \{1, 2, 3\}$ for Case 3a, and $D_F = \{1, 2, 3, 4, 5, 6\}$ and $D_E = \{1, 2, 3\}$ for Case 3b.

As seen from Tables 4 and 5, DC-SIS, MDC-SIS, and NIS are all capable of capturing (X_1, X_2, X_3) with a very high percentage whereas SIS and SIRS have a relatively low percentage of including X_3. As expected, the MDC-SIS does not include X_4, X_5, and X_6 (only in Case 3b) with a high percentage, whereas DC-SIS does capture these variables in the active set D_F with a decent percentage. So the minimum model size for MDC-SIS is considerably smaller than that for DC-SIS, showing the efficiency of MDC-SIS over DC-SIS if we are only interested in the variables that contribute to conditional mean of the response variable.

4.2 Conditional Quantile Screening

In this section, we compare three different conditional quantile screening methods, SISQ, MDC-SISQ, and QaSIS (He,

Table 3. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ for Example 2 with $d_1 = 20, d_2 = 30, d_3 = \lfloor n/\log n \rfloor$, $n = 400$ and $p = 1000$

	Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	ALL
2.a	SIS	1	0.014	1	1	0.010	0.012	0.022	0.024	0.008	0.010	0.012	0.018	0.014
(d_1)	DC-SIS	1	0.264	1	1	0.014	0.014	0.018	0.026	0.012	0.010	0.014	0.016	0.254
	MDC-SIS	1	0.354	1	1	0.012	0.014	0.016	0.030	0.006	0.008	0.014	0.022	0.328
	NIS	1	0.066	1	1	0.018	0.018	0.014	0.018	0.020	0.014	0.020	0.014	0.066
	SIRS	1	0.016	1	1	0.014	0.014	0.022	0.018	0.014	0.012	0.014	0.012	0.012
2.a	SIS	1	0.028	1	1	0.020	0.022	0.034	0.036	0.026	0.018	0.018	0.028	0.026
(d_2)	DC-SIS	1	0.378	1	1	0.028	0.028	0.026	0.040	0.026	0.020	0.024	0.030	0.364
	MDC-SIS	1	0.482	1	1	0.020	0.022	0.030	0.038	0.028	0.020	0.018	0.032	0.474
	NIS	1	0.084	1	1	0.028	0.026	0.026	0.028	0.026	0.022	0.028	0.020	0.082
	SIRS	1	0.026	1	1	0.032	0.026	0.026	0.036	0.034	0.020	0.020	0.020	0.024
2.a	SIS	1	0.064	1	1	0.056	0.054	0.064	0.074	0.064	0.064	0.048	0.058	0.064
(d_3)	DC-SIS	1	0.640	1	1	0.068	0.056	0.050	0.076	0.072	0.068	0.048	0.052	0.640
	MDC-SIS	1	0.758	1	1	0.060	0.058	0.068	0.076	0.066	0.068	0.052	0.052	0.758
	NIS	1	0.132	1	1	0.076	0.052	0.068	0.066	0.066	0.064	0.066	0.044	0.132
	SIRS	1	0.07	1	1	0.06	0.072	0.05	0.074	0.07	0.068	0.042	0.062	0.07
2.b	SIS	0.610	0.016	0.722	0.374	0.950	0.022	0.992	0.820	0.998	0.022	1.000	0.988	0.000
(d_1)	DC-SIS	0.520	0.060	0.790	0.686	0.922	0.218	0.998	0.994	0.992	0.594	1.000	1.000	0.004
	MDC-SIS	0.572	0.060	0.844	0.736	0.948	0.242	1.000	0.996	0.996	0.676	1.000	1.000	0.006
	NIS	0.162	0.004	0.266	0.146	0.432	0.008	0.674	0.482	0.728	0.016	0.968	0.850	0.000
	SIRS	0.512	0.012	0.652	0.332	0.910	0.020	0.980	0.778	0.990	0.020	1.000	0.986	0.000
2.b	SIS	0.680	0.024	0.796	0.448	0.972	0.036	0.996	0.880	0.998	0.028	1.000	0.994	0.000
(d_2)	DC-SIS	0.610	0.094	0.866	0.774	0.954	0.338	1.000	0.996	0.992	0.754	1.000	1.000	0.008
	MDC-SIS	0.660	0.118	0.900	0.820	0.966	0.404	1.000	1.000	0.998	0.816	1.000	1.000	0.026
	NIS	0.208	0.006	0.318	0.192	0.498	0.012	0.726	0.526	0.774	0.020	0.980	0.862	0.000
	SIRS	0.582	0.026	0.736	0.396	0.944	0.026	0.994	0.844	0.992	0.040	1.000	0.990	0.000
2.b	SIS	0.786	0.070	0.878	0.618	0.988	0.064	1	0.944	1	0.068	1	1	0
(d_3)	DC-SIS	0.730	0.248	0.936	0.900	0.976	0.586	1	1	0.998	0.932	1	1	0.096
	MDC-SIS	0.778	0.278	0.944	0.948	0.984	0.678	1	1	1.000	0.962	1	1	0.126
	NIS	0.318	0.022	0.454	0.284	0.630	0.026	0.800	0.614	0.864	0.044	0.990	0.904	0.000
	SIRS	0.73	0.064	0.854	0.548	0.978	0.064	0.996	0.920	0.998	0.074	1	0.996	0.000

Table 4. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ for Example 3

	Method	5%	25%	50%	75%	95%
3.a	SIS	3.00	3.00	4.00	29.00	585.95
	DC-SIS	5.00	11.00	32.00	92.25	424.00
	MDC-SIS	3.00	3.00	3.00	3.00	9.05
	NIS	3.00	3.00	3.00	4.00	14.00
	SIRS	22.95	80.75	186.50	413.00	896.70
3.b	SIS	3.00	3.00	5.00	50.25	667.55
	DC-SIS	9.00	28.50	69.00	187.25	662.05
	MDC-SIS	3.00	3.00	3.00	3.00	17.00
	NIS	3.00	3.00	3.00	7.00	24.00
	SIRS	58.00	154.75	316.50	595.75	1296.05

Wang, and Hong 2013) for several heteroscedastic models with $(n, p) = (100, 3000)$ and $(200, 3000)$. We do not include DC-SISQ as it is equivalent to MDC-SISQ as shown in Proposition 1, but include DC-SIS and SIRS to see if there is any benefit of transforming Y for the purpose of conditional quantile screening.

Example 4.

Case 4a: $Y = X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5 + \exp(X_{20} + X_{21} + X_{22}) \cdot \epsilon$.

Case 4b: $Y = X_1 + 0.8\sin(|X_2|) + 0.6\exp(|X_3|) + 0.4X_4 + 0.2X_5 + \exp(X_{20} + X_{21} + X_{22}) \cdot \epsilon$.

Case 4c: $Y = X_1X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5 + \exp(X_{20} + X_{21} + X_{22}) \cdot \epsilon$.

In the above models, the error $\epsilon \sim N(0, 1)$ and is independent from the covariates. The predictor vector follows the multivariate normal distribution with the correlation structure described in Example 1 but with $\rho = 0.8$. All models in this example are heteroscedastic with the number of active variables being five at the median (i.e., $\tau = 0.5$) but eight for other τ's. Case 4a is adopted from an example used in Zhu et al. (2011) and also analyzed by He, Wang, and Hong (2013). Cases 4b and 4c are modified versions of Case 4a by including nonlinear structure and interaction terms. We report $\mathcal{S}$, $\mathcal{P}_s$, and $\mathcal{P}_a$ with $d = \lfloor n/\log n \rfloor$ for all three methods in Tables 6 and 7.

As can be seen from Table 6, MDC-SISQ performs noticeably better than QaSIS for conditional median screening with the 5%, 25%, 50%, 75% quantiles of their minimum model sizes substantially smaller than the corresponding QaSIS-based ones for all cases, and the 95% quantiles of the minimum model sizes being comparable for Case 4b. Comparing to SISQ, MDC-SISQ does equally well in Case 4a, but significantly better in Cases 4b and 4c for conditional median screening. For conditional quantile screening at $\tau = 0.75$ and Cases 4a and 4b, the 5%, 25%, 50% quantiles of the minimum model sizes for

Table 5. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ for Example 3 with $d_1 = 20$, $d_2 = 30$, $d_3 = \lfloor n/\log n \rfloor$

	Method	X_1	X_2	X_3	X_4	X_5	X_6	ALL
3.a	SIS	0.998	0.998	0.704	0.240	0.108	0.016	0.698
(d_1)	DC-SIS	1.000	1.000	0.998	0.732	0.468	0.022	0.392
	MDC-SIS	0.998	1.000	0.972	0.264	0.092	0.010	0.970
	NIS	0.994	0.992	0.980	0.396	0.270	0.062	0.974
	SIRS	1.000	1.000	0.616	0.288	0.100	0.008	0.040
3.a	SIS	0.998	0.998	0.754	0.280	0.124	0.024	0.752
(d_2)	DC-SIS	1.000	1.000	1.000	0.790	0.550	0.042	0.482
	MDC-SIS	1.000	1.000	0.984	0.320	0.114	0.018	0.984
	NIS	0.994	0.996	0.992	0.468	0.324	0.080	0.988
	SIRS	1.000	1.000	0.676	0.348	0.154	0.020	0.080
3.a	SIS	0.998	0.998	0.774	0.306	0.140	0.030	0.774
(d_3)	DC-SIS	1.000	1.000	1.000	0.818	0.588	0.066	0.538
	MDC-SIS	1.000	1.000	0.992	0.336	0.124	0.022	0.992
	NIS	0.998	0.998	0.992	0.492	0.348	0.090	0.990
	SIRS	1.000	1.000	0.694	0.394	0.198	0.032	0.110
3.b	SIS	1.000	0.998	0.662	0.218	0.106	0.046	0.650
(d_1)	DC-SIS	1.000	1.000	0.998	0.716	0.590	0.272	0.190
	MDC-SIS	1.000	1.000	0.960	0.248	0.084	0.034	0.958
	NIS	0.974	0.978	0.960	0.352	0.302	0.196	0.936
	SIRS	1.000	1.000	0.612	0.272	0.136	0.066	0.006
3.b	SIS	1.000	1.000	0.708	0.268	0.128	0.066	0.702
(d_2)	DC-SIS	1.000	1.000	1.000	0.764	0.662	0.360	0.256
	MDC-SIS	1.000	1.000	0.972	0.282	0.118	0.046	0.972
	NIS	0.988	0.988	0.978	0.418	0.376	0.236	0.966
	SIRS	1.000	1.000	0.652	0.334	0.204	0.092	0.016
3.b	SIS	1.000	1.000	0.734	0.292	0.140	0.074	0.734
(d_3)	DC-SIS	1.000	1.000	1.000	0.800	0.690	0.412	0.318
	MDC-SIS	1.000	1.000	0.972	0.296	0.124	0.054	0.972
	NIS	0.988	0.994	0.988	0.448	0.404	0.254	0.978
	SIRS	1.000	1.000	0.674	0.370	0.246	0.098	0.020

Table 6. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ for Example 4

Settings	τ	Method	5%	25%	50%	75%	95%
				Case 4.a			
$\frac{n=100}{p=3000}$	0.5	SISQ	5.00	5.00	6.00	10.00	79.00
		MDC-SISQ	5.00	5.00	6.00	12.00	78.10
		QaSIS	9.00	20.00	42.00	86.00	272.00
	0.75	SISQ	23.00	117.75	364.00	986.75	2448.10
		MDC-SISQ	22.95	115.50	372.50	1112.75	2471.95
		QaSIS	92.00	202.75	392.50	786.00	1927.60
		DC-SIS	13.95	143.00	493.00	1083.25	1974.75
		SIRS	12.0	26.0	50.0	92.0	243.1
$\frac{n=200}{p=3000}$	0.5	SISQ	5.00	5.00	5.00	5.00	6.00
		MDC-SISQ	5.00	5.00	5.00	5.00	6.00
		QaSIS	5.00	6.00	8.00	10.00	17.00
	0.75	SISQ	9.00	15.00	34.00	154.25	1308.40
		MDC-SISQ	9.00	13.00	30.50	148.25	1108.10
		QaSIS	16.95	33.00	64.00	139.00	461.35
		DC-SIS	11.00	37.00	158.00	478.75	1535.90
		SIRS	8	9	11	14	25
				Case 4.b			
$\frac{n=100}{p=3000}$	0.5	SISQ	6.00	28.75	139.50	589.50	1742.25
		MDC-SISQ	6.00	17.00	78.00	272.50	1292.50
		QaSIS	19.00	54.00	144.00	393.50	1249.05
	0.75	SISQ	45.00	233.00	720.50	1505.25	2766.05
		MDC-SISQ	42.00	210.25	537.00	1389.25	2542.80
		QaSIS	257.00	537.75	924.00	1436.75	2310.10
		DC-SIS	156.85	770.25	1368.50	1963.75	2628.20
		SIRS	17.95	42.75	107.00	432.00	1572.25
$\frac{n=200}{p=3000}$	0.5	SISQ	5.00	6.00	14.00	59.25	478.75
		MDC-SISQ	5.00	5.00	7.00	20.00	225.60
		QaSIS	6.00	9.00	14.00	31.00	176.45
	0.75	SISQ	10.00	24.00	75.50	277.25	1195.65
		MDC-SISQ	9.00	20.75	52.00	167.75	741.05
		QaSIS	47.95	118.00	226.50	460.25	1177.80
		DC-SIS	65.95	438.75	887.50	1463.75	2394.60
		SIRS	8.00	10.00	14.00	44.25	297.00
				Case 4.c			
$\frac{n=100}{p=3000}$	0.5	SISQ	13.00	104.00	408.00	1189.00	2616.30
		MDC-SISQ	7.00	28.00	104.50	316.50	1138.05
		QaSIS	25.00	103.75	313.00	751.25	2064.00
	0.75	SISQ	294.80	1337.00	2234.00	2695.75	2956.10
		MDC-SISQ	175.40	493.75	1089.50	1895.75	2704.80
		QaSIS	272.00	563.00	1081.50	1721.00	2664.50
		DC-SIS	448.35	1231.25	1879.00	2352.50	2786.10
		SIRS	23.00	83.00	287.50	875.50	2386.25
$\frac{n=200}{p=3000}$	0.5	SISQ	6.95	27.75	117.00	618.25	1882.10
		MDC-SISQ	5.00	6.00	11.00	37.25	266.55
		QaSIS	7.00	15.00	37.50	125.50	763.05
	0.75	SISQ	97.70	803.00	1924.50	2661.00	2955.15
		MDC-SISQ	22.95	141.75	374.00	771.00	1840.90
		QaSIS	56.00	144.00	343.00	749.25	1797.25
		DC-SIS	324.6	936.5	1483.5	2037.5	2666.1
		SIRS	9.00	14.00	47.00	210.25	1288.45

MDC-SISQ and SISQ are substantially smaller than the corresponding QaSIS-based ones , and the 75% and 95% quantiles for MDC-SISQ and SISQ are mostly comparable to QaSIS-based ones and are slightly larger in a few settings. In Case 4c, SISQ does worse than both MDC-SISQ and QaSIS, whose performances are comparable. A comparison with DC-SIS and SIRS shows that for this particular example, SIRS is more effective in including all the predicator variables which are dependent on the response than DC-SIS in all cases. In terms of conditional median screening, MDC-SISQ delivers a model of smaller size than

Table 7. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ for Example 4 with $d = \lfloor n/\log n \rfloor$

Settings	τ	Method	X_1	X_2	X_3	X_4	X_5	X_{20}	X_{21}	X_{22}	ALL
						Case 4.a					
$\frac{n=100}{p=3000}$	0.5	SISQ	0.998	1.000	1.000	0.978	0.868	0.002	0.004	0.010	0.866
		MDC-SISQ	0.998	1.000	1.000	0.976	0.850	0.010	0.006	0.004	0.850
		QaSIS	0.848	0.900	0.852	0.656	0.312	0.576	0.672	0.568	0.284
	0.75	SISQ	0.840	0.880	0.828	0.722	0.524	0.332	0.360	0.308	0.042
		MDC-SISQ	0.844	0.878	0.832	0.698	0.512	0.354	0.386	0.330	0.046
		QaSIS	0.076	0.100	0.070	0.058	0.022	0.750	0.882	0.768	0.000
		DC-SIS	0.286	0.364	0.292	0.200	0.104	0.976	0.998	0.976	0.082
		SIRS	0.998	1.000	0.998	0.982	0.896	0.370	0.550	0.386	0.186
$\frac{n=200}{p=3000}$	0.5	SISQ	1.000	1.000	1.000	1.000	1.000	0.018	0.010	0.014	1.000
		MDC-SISQ	1.000	1.000	1.000	1.000	0.998	0.018	0.012	0.020	0.998
		QaSIS	1.000	1.000	1.000	1.000	0.990	0.644	0.750	0.654	0.990
	0.75	SISQ	0.996	1.000	0.996	0.986	0.952	0.686	0.752	0.676	0.524
		MDC-SISQ	0.998	1.000	0.996	0.988	0.950	0.714	0.776	0.682	0.536
		QaSIS	0.770	0.838	0.780	0.604	0.398	0.958	0.990	0.964	0.296
		DC-SIS	0.582	0.658	0.606	0.482	0.284	1.000	1.000	1.000	0.252
		SIRS	1.000	1.000	1.000	1.000	1.000	0.990	0.998	0.986	0.976
						Case 4.b					
$\frac{n=100}{p=3000}$	0.5	SISQ	0.854	0.626	0.390	0.462	0.354	0.010	0.010	0.016	0.220
		MDC-SISQ	0.902	0.776	0.750	0.570	0.374	0.010	0.018	0.018	0.286
		QaSIS	0.370	0.528	0.766	0.348	0.164	0.614	0.714	0.612	0.062
	0.75	SISQ	0.460	0.448	0.462	0.404	0.254	0.382	0.430	0.384	0.016
		MDC-SISQ	0.502	0.550	0.638	0.454	0.278	0.420	0.472	0.394	0.018
		QaSIS	0.012	0.014	0.010	0.010	0.006	0.738	0.878	0.762	0.000
		DC-SIS	0.050	0.054	0.054	0.040	0.022	0.976	0.998	0.982	0.006
		SIRS	0.832	0.626	0.444	0.512	0.370	0.662	0.812	0.662	0.090
$\frac{n=200}{p=3000}$	0.5	SISQ	0.996	0.954	0.790	0.886	0.764	0.026	0.038	0.028	0.676
		MDC-SISQ	1.000	0.988	0.992	0.954	0.830	0.028	0.026	0.030	0.820
		QaSIS	0.994	0.998	1.000	0.984	0.790	0.744	0.800	0.748	0.784
	0.75	SISQ	0.850	0.844	0.842	0.796	0.654	0.782	0.842	0.764	0.334
		MDC-SISQ	0.896	0.954	0.978	0.916	0.696	0.810	0.878	0.806	0.420
		QaSIS	0.196	0.272	0.312	0.190	0.130	0.972	0.982	0.968	0.020
		DC-SIS	0.176	0.150	0.150	0.108	0.064	1.000	1.000	1.000	0.024
		SIRS 0.994	0.964	0.858	0.896	0.786	0.996	1.000	1.000	0.720	
						Case 4.c					
$\frac{n=100}{p=3000}$	0.5	SISQ	0.096	0.284	0.704	0.762	0.612	0.014	0.018	0.016	0.074
		MDC-SISQ	0.328	0.550	0.836	0.826	0.634	0.010	0.020	0.010	0.214
		QaSIS	0.372	0.694	0.772	0.272	0.112	0.620	0.718	0.590	0.036
	0.75	SISQ	0.058	0.064	0.082	0.076	0.056	0.356	0.434	0.400	0.000
		MDC-SISQ	0.160	0.318	0.372	0.140	0.076	0.392	0.460	0.422	0.000
		QaSIS	0.012	0.020	0.018	0.004	0.006	0.722	0.866	0.746	0.000
		DC-SIS	0.014	0.022	0.022	0.012	0.004	0.978	0.998	0.982	0.000
		SIRS	0.152	0.340	0.718	0.762	0.612	0.716	0.854	0.716	0.044
$\frac{n=200}{p=3000}$	0.5	SISQ	0.330	0.658	0.986	0.992	0.964	0.034	0.030	0.034	0.318
		MDC-SISQ	0.790	0.946	0.998	0.998	0.964	0.042	0.026	0.032	0.750
		QaSIS	0.990	1.000	1.000	0.898	0.522	0.702	0.814	0.730	0.500
	0.75	SISQ	0.084	0.106	0.138	0.164	0.128	0.760	0.822	0.772	0.020
		MDC-SISQ	0.566	0.842	0.880	0.416	0.188	0.776	0.846	0.778	0.076
		QaSIS	0.238	0.370	0.388	0.200	0.108	0.958	0.984	0.972	0.022
		DC-SIS	0.034	0.060	0.094	0.048	0.040	1.000	1.000	1.000	0.000
		SIRS	0.472	0.728	0.972	0.988	0.950	1.000	1.000	1.000	0.464

both SIRS and DC-SIS, whereas SIRS seems more efficient to identify all the variables (X_1, X_2, X_3, X_4, X_5, X_{20}, X_{21}, X_{22}), which correspond to the active set of MDC-SISQ (0.75).

Table 7 shows that when $\tau = 0.5$, QaSIS is somehow capable of preserving the inactive variables (X_{20}, X_{21}, X_{22}) in the heteroscedastic error part, which do not enter into the conditional median. At $\tau = 0.75$, the variables in the heteroscedastic

error part do enter into the conditional quantile, and the performance for MDC-SISQ is inferior to QaSIS in this regard. On the other hand, for other variables $(X_1, \ldots, X_5)$, MDC-SISQ's performance is superior to QaSIS in Cases 4a and 4b, and the performance of QaSIS improves substantially when n increases from 100 to 200, indicating that QaSIS works more reliably with a large sample size, which is presumably related to the nonparametric marginal estimation involved. For Case 4c, MDC-SISQ outperforms QaSIS in the case $n = 100$, whereas MDC-SISQ and QaSIS are comparable when $n = 200$. Although SISQ and MDC-SISQ deliver comparable results under the linear model in Case 4a, MDC-SISQ has some edge over SISQ in dealing with the nonlinear cases. In particular, MDC-SISQ has a higher proportion of including $\{X_2, X_3\}$ in Case 4b and $\{X_1, X_2\}$ in Case 4c, which enter into the conditional quantile of Y in a nonlinear fashion. As far as the percentage of including all active predictors is concerned, MDC-SISQ and SISQ perform comparably in Case 4a, for which QaSIS is the worst. The DC-SIS seems to pick up the three variables in the heteroscedastic errors very effectively, but are unable to include the variables $(X_1, \ldots, X_5)$ especially for Cases 4b and 4c. By contrast, SIRS tends to include all the variables $(X_1, \ldots, X_5, X_{20}, \ldots, X_{22})$ quite effectively, especially when $n = 200$.

Tables 16–21 in the supplementary material present the results for Case 4(a–c) with varying signal-to-noise ratios. Since the results are qualitatively similar, we do not provide a detailed comment. Upon the suggestion of an associate editor, we further compare the performance of several screening procedures using all the examples in He, Wang, and Hong (2013) as listed below.

Example HWH1: (additive model, $n = 400$, $p = 1000$). This example is adapted from Fan, Feng, and Song (2011). Case 1a: $Y = 5g_1(X_1) + 3g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sqrt{(1.74)}\epsilon$, where the vector of covariates X is generated from the multivariate normal distribution $\mathrm{N}(0, \Sigma)$ with $\sigma_{ij} = \rho^{|i-j|}$ and the expressions of g_j, $j = 1, 2, 3, 4$ can be found from Example 2. In Case 1a, we consider $\rho = 0$; Case 1b: same as Case 1a except that $\rho = 0.8$; Case 1c: same as Case 1b except that $\epsilon \sim$ Cauchy.

Example HWH2: (index model, $n = 200$, $p = 2000$). This example is adapted from Zhu et al. (2011). The random data are generated from

$$Y = 2(X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5) + \exp(X_{20} + X_{21} + X_{22}) \cdot \epsilon.$$

This model is heteroscedastic: the number of active variables is five at the median but eight elsewhere.

Example HWH3: (a more complex structure, $n = 400$, $p = 5000$). Case 3a: $Y = 2(X_1^2 + X_2^2) + \exp((X_1 + X_2 + X_{18} + X_{19} + \cdots + X_{30})/10) \cdot \epsilon$, where $\epsilon \sim N(0, 1)$, and X follows the multivariate normal distribution with the correlation structure described in Case 1b. In this case, the number of active variables is 2 at the median but is 15 elsewhere. Case 3b: Same as Case 3a, but with $2(X_1^2 + X_2^2)$ replaced by $2((X_1 + 1)^2 + (X_2 + 2)^2)$.

As can be seen from Tables 22–24 in the supplementary material, the performance of MDC-SISQ and QaSIS are quite comparable. In Case 1a, all screening procedures deliver similar performance; in Case 1b, MDC-SISQ, NIS, DC-SIS, and QaSIS outperform the other two; in Case 1c, MDC-SISQ, DC-SIS, and QaSIS are the top three. For Case 2, when $\tau = 0.5$, SISQ, MDC-SISQ, and QaSIS perform equally well, whereas for $\tau = 0.75$, QaSIS outperforms. SIRS does the best in this case as an independence screening procedure. In Case 3a, MDC-SISQ appears to top the other two for $\tau = 0.5$, whereas for $\tau = 0.75$, QaSIS does comparably well and both are superior to DC-SIS, NIS, and SIRS. In Case 3b, SISQ, MDC-SISQ, and QaSIS perform equally well when $\tau = 0.5$, whereas QaSIS has some advantage over SISQ and MDC-SISQ when $\tau = 0.75$. SIRS is the best among the three independence screening procedures without any transformation of response variables.

Based on the above results, the finite sample performance of MDC-SIS and MDC-SISQ is quite encouraging. For the normal linear models presented in Example 1, MDC-SIS is inferior to SIS but outperforms DC-SIS and SIRS. For nonparametric models presented in Example 2, MDC-SIS and DC-SIS demonstrate superior performance over SIRS and NIS. For Example 3, where the active set for MDC-SIS and DC-SIS (SIRS) are different, MDC-SIS appears to be more efficient than DC-SIS in terms of having a smaller minimum model size. In the case of conditional quantile screening, MDC-SISQ seems favorable to SISQ and comparable to QaSIS in most cases. Again MDC-SISQ can be more efficient than DC-SIS, NIS, and SIRS if the active set for MDC-SISQ is smaller than that for DC-SIS and others. It is worth noting that the computational cost for MDC-SISQ is much cheaper than QaSIS, since the latter involves fitting marginal spline-based quantile regression models and is quite computationally expensive.

An inherent limitation of the marginal screening method is that it may miss important predictors that have zero marginal utility with the response. This is no exception for MDC-based screening procedures. One way to fix this issue is to develop an iterative procedure similar to Fan, Samworth, and Wu (2009) and Zhu et al. (2011). In an unreported simulation study, we applied iterated MDC mimicking iterated feature screening as described in Section 2.5 of Zhu et al. (2011) but we found that the performance largely depends on the chosen tuning parameters (i.e., the number of variables selected at each stage), which seems to be a difficult practical issue.

5. DATA ILLUSTRATIONS

In this section, we illustrate the proposed methods by empirical analysis of two real datasets.

5.1 Cardiomyopathy Microarray Data

The cardiomyopathy microarray data are generated from a transgenic mouse model of dilated cardiomyopathy by Redfern et al. (2000), and have been analyzed by Hall and Miller (2009) using the generalized correlation-based screening and Li, Zhong, and Zhu (2012) with DC-SIS. Our particular interest is to identify the most influential genes for overexpression of Ro1 through DNA-array analysis, which helps to discover more diagnostic and therapeutic targets for the syndrome. A full listing of these gene expression data can be found in the supplementary material of Redfern et al. (2000) at *www.pnas.org*, which consists of 30 arrays over 6319 genes from hearts expressing Ro1.

Table 8. Adjusted R^2 and the deviance explained (%) for three methods

Method		SIS	
Rank	Msa.2877.0	Msa.964.0	Msa.2134.0
Single	80.0/81.4		
Double(1,2)		84.2/85.8	
Method		DC-SIS	
Rank	Msa.2134.0	Msa.2877.0	Msa.26025.0
Single	65.7/69.8		
Double(1,2)		96.8/98.3	
Method		MDC-SIS	
Rank	Msa.2877.0	Msa.2134.0	Msa.741.0
Single	80.0/81.4		
Double(1,2)		96.8/98.3	

In our analysis, we treat the Ro1 expression level as the response, and other gene expression levels as predictors. All the gene expression levels have been standardized to have mean 0 and variance 1. The SIS procedure ranks Msa.2877.0, Msa.964.0, and Msa.2134.0 at the top, while the DC-SIS procedure delivers a different ranking order and chooses Msa.2134.0, Msa.2877.0, and Msa.26025.0 as the top three. Our finding is different from both of them as we rank Msa.2877.0, Msa.2134.0, and Msa.741.0 at the top. Notice that there are two overlapping genes among these three methods, but the ranking orders are different. We compare the performance of three feature screening methods following the same manner described in Li, Zhong, and Zhu (2012). For each approach, we fit a nonparametric additive model to their selected genes by including top one or two genes. Due to the small sample size, we do not consider the model with more than two predictor variables. The adjusted R^2 and the explained deviance (the proportion of the null deviance explained by the proposed model, with a larger value indicating better performance) are reported in Table 8 for each fitted model.

We can see from Table 8 that MDC-SIS compares favorably with SIS and DC-SIS, in terms of delivering a more sensible ranking to genes with higher adjusted R^2 and the deviance explained. In particular, SIS and MDC-SIS outperform DC-SIS in terms of the adjusted R^2 and the deviance explained for the nonparametric model with the top gene selected as the predictor variable; MDC-SIS and DC-SIS have the same top two predictor variables and the fitted additive model seems superior to the one with the top two predictor variables selected by using SIS.

5.2 Affymetric GeneChip Rat Genome 230 2.0 Array Data

In this section, we apply our proposed methods to microarray data reported in Scheetz et al. (2006). The later authors carried out expression quantitative trait locus (eQTL) mapping experiment on 120 12-week-old male F2 rats to gain a broad perspective of gene regulation in the mammalian eye and to identify genetic variation relevant to human eye disease. This gene expression dataset consists of 120 arrays, each of which contains 31,042 probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array), which is analyzed on a logarithmic scale. The complete gene expression matrix is available at Gene Expression Omnibus (*http://www.ncbi.nlm.nih.gov/geo*) with accession number GSE5680. This dataset has been analyzed by a few statisticians, including Fan, Feng, and Song (2011), Wang, Wu, and Li (2012), and Huang, Horowitz, and Wei (2010), among others.

Table 9. Results of gene screening using 3 feature screening methods

	$d = \lfloor \frac{n}{\log n} \rfloor$			$d = n$		
Probe ID	SIS	DC-SIS	MDC-SIS	SIS	DC-SIS	MDC-SIS
1371755_at						
1372928_at					√	√
1373534_at		√	√	√	√	√
1373944_at					√	√
1374669_at					√	√
1376686_at	√	√	√	√	√	√
1376747_at	√	√	√	√	√	√
1377880_at			√	√	√	√
1378590_at	√	√	√	√	√	√

"√" represents this gene has been selected.

We focus on 18,976 probes which are considered to be expressed in animals' eyes and have sufficient variation over gene expression levels. More details related to these two standards can be found in Scheetz et al. (2006). Among these 18,976 genes, gene TRIM 32 has been recently identified and validated to cause Bardet–Biedl syndrome (Chiang et al. 2006), which is a multisystem human disease. Of particular interest is to identify which gene expressions are related to that of gene TRIM 32, the probe ID of which is 1389163_at. In this way, additional disease-causing genes can be discovered based on the principle that pairwise correlation of gene expression reveals biologically relevant functional relationships.

Before the analysis, we standardized each probe to have mean 0 and variance 1. According to Fan, Feng, and Song (2011), nine genes have been selected as the most relevant ones by INIS-penGAM, which is a two-stage variable selection method. Since variable screening methods have a different aim, we cannot compare MDC-SIS with INIS-penGAM directly. Instead we compare the performance of SIS, MDC-SIS, and DC-SIS by examining whether they are capable of including these nine genes when d is smaller than or comparable to n. The results are reported in Table 9 with the first column containing the probe IDs for the nine genes. It can be seen that MDC-SIS and DC-SIS include more covariates than SIS does under both scenarios, which can be interpreted by their capability of detecting nonlinear dependence relationships. Also we see that MDC-SIS includes one more gene than DC-SIS when $d = \lfloor \frac{n}{\log(n)} \rfloor$. For some reason, the first gene 1371755_at is not included by all three methods and is likely to be identified if we apply an iterative procedure as INIS-penGAM.

Following the suggestion of a reviewer, we apply the screening procedures SIS, MDC-SIS, DC-SIS, and NIS to reduce the dimensionality of covariates to $d = \lfloor n/\log(n) \rfloor$, followed by variable selection within the framework of nonparametric additive models using penGAM, which was proposed by Meier, van de Geer, and Bühlmann (2009). We use six-fold cross-validation to select the tuning parameters inside the penGAM algorithm. In

Table 10. Model fitting and prediction for the second application

Method	All data # of probes	All data RSS	Random partition Ave. # of probes	Random partition Ave. PE
SIS-penGAM	17	0.169	14.54 (2.00)	0.458 (0.239)
MDC-penGAM	14	0.156	13.94 (1.81)	0.487 (0.262)
DC-penGAM	14	0.160	14.08 (2.08)	0.504 (0.247)
NIS-penGAM	16	0.177	15.35 (2.11)	0.546 (0.225)

the second and third columns of Table 10, we report the number of probes (All data # of probes) selected by each method, as well as their corresponding residual sum of squares (All data RSS). The MDC-based procedure delivers the smallest RSS value and selects smaller number of variables than SIS and NIS-based counterparts. To further evaluate the performance of these methods, we also conduct 100 random partitions, each of which contains 100 observations as the training data and the other 20 as the testing data. A five-fold cross-validation is applied to the training data to select the tuning parameters. We compute the number of probes selected based on the training data and the mean-squared prediction errors using the testing data. This process is repeated 100 times. The fourth and fifth columns of Table 10 present the average values and their associated standard deviations (in the parentheses) over 100 replications. Somewhat surprisingly, SIS-based procedure delivers the smallest prediction error followed by MDC-based counterpart, although none of the procedures significantly outperform the other in view of the standard deviations. We thus conclude that the performance of MDC-based procedure is quite competitive. We also tried different d or different ways of selecting tuning parameters in penGAM and obtained quantitatively similar results (not shown).

We also apply conditional quantile screening methods SISQ and MDC-SISQ on the same dataset, to examine the heterogeneity in this dataset. Comparing the top $d = \lfloor \frac{n}{\log n} \rfloor$ probes selected at different quantiles by MDC-SISQ, that is, MDC-SISQ(0.3), MDC-SISQ(0.5), MDC-SISQ(0.7), where MDC-SISQ(τ) denotes the MDC-SISQ applied at the τth quantile level, only one probe (1372453_at) is selected at all three quantile levels. A similar phenomenon has also been observed with SISQ, which demonstrates the heterogeneity in the data. A natural next step is to fit a linear or nonparametric additive quantile regression model to a moderate number of predictor variables kept by MDC-SISQ. One can do variable selection by including LASSO or SCAD penalty (Wang, Wu, and Li 2012), the details of which are omitted as it is beyond the scope of this article.

6. DISCUSSION AND CONCLUSIONS

In this article, we propose a new metric, namely the martingale difference correlation, to measure the degree of conditional mean independence of Y given X. The martingale difference correlation is a natural extension of the distance correlation proposed recently by Székely, Rizzo, and Bakirov (2007), which was used to measure the dependence between Y and X. It turns out that the martingale difference correlation and its sample analog inherit a number of nice properties of the distance correlation and its sample counterpart. Motivated by the recent work of Li, Zhong, and Zhu (2012), we further propose to use the martingale difference correlation as a marginal utility to do high-dimensional variable screening. For conditional mean screening, the MDC-based screening method has a natural interpretation in that we screen out the variables that do not contribute to the conditional mean of Y given X marginally. Furthermore, the MDC-based screening procedure can be easily extended to do conditional quantile screening, thus making it broadly applicable to variable screening for high-dimensional heterogenous data. Theoretically, we show the sure screening consistency for MDC-based conditional mean and quantile screening under suitable conditions. Simulation results show the usefulness of the MDC-based screening procedures which can outperform existing counterparts.

To conclude, we reiterate several appealing features of the MDC-based screening procedures. First, it is model-free, like DC-SIS of Li, Zhong, and Zhu (2012). However, unlike the DC-based procedure, which screens out the variables that are marginally independent of the distribution of the response Y, we focus on a specific aspect of the conditional distribution of Y given X, say conditional mean or quantiles of Y given X. This is directly linked to the second stage after variable screening: model fitting and variable selection. For example, to select variables entered into the quantile regression models at different quantile levels in the second stage, it may be more efficient to use variables kept by MDC-based conditional quantile screening procedures at these quantile levels in the first stage than the ones suggested by DC-based screening since the active set corresponding to the latter may contain some redundant variables that do not contribute to the modeling of a particular conditional quantile; see Section 4.2. Second, the MDC-based screening procedures are conceptually simple, convenient to implement with no tuning parameters or nonparametric model fitting involved. The calculation of sample MDC is straightforward, similar to sample DC. Third, as we mentioned at the end of Section 2, it can be easily extended to do screening for other aspects of the conditional distribution of the response given covariates, besides what we mentioned in the article. This hopefully provides some flexibility for the user with little sacrifice of efficiency to explore the conditional distribution of high-dimensional data. Of course, its theoretical and finite sample performance remains unexplored and will be an interesting topic for future research. In summary, it seems fair to view MDC-based procedures as a useful complement but not a competitor to the growing class of screening methods. Since the existing mode-free screening methods are either correlation-based or dependence-based, the proposed MDC-based procedure can hopefully be recommended to the practitioner as an additional tool for the analysis and modeling of high-dimensional data.

SUPPLEMENTARY MATERIALS

The supplementary material contains some additional simulation results and proofs of Theorems 3–6.

TECHNICAL APPENDIX I

This appendix contains the proofs of Theorem 1, Theorem 2, Proposition 1, and Proposition 2.

Proof of Theorem 1.

1. To prove the first statement, we shall prove the following two assertions separately:

$$\begin{aligned}\mathrm{cov}^2_{B,id}(U,V) &= E(U_B U'_B V_{id} V'_{id}) \\ &= -E[(V-EV)(V'-EV')|U-U'|_q]; \end{aligned} \tag{A.3}$$

$$\mathrm{MDD}^2(V|U) = -E[(V-EV)(V'-EV')|U-U'|_q]. \tag{A.4}$$

First we show that $E(U_B U'_B V_{id} V'_{id})$ is nonnegative and finite. Following the argument in the proof of Theorem 7 in Székely and Rizzo (2009), we have that

$$\begin{aligned} E[U_B U'_B V_{id} V'_{id}] &= E[E(U_B U'_B V_{id} V'_{id}|B)] \\ &= E[E(U_B V_{id}|B)E(U'_B V'_{id}|B)] \\ &= E[E(U_B V_{id}|B)]^2 \geq 0, \end{aligned}$$

and

$$\begin{aligned} E[U_B U'_B V_{id} V'_{id}] &\leq \frac{1}{2}(E|U_B U'_B|^2 + E|V_{id} V'_{id}|^2) \\ &\leq \frac{1}{4}(E|U_B|^4 + E|U'_B|^4) + \frac{1}{2}[E(V-EV)^2]^2 \\ &< \infty, \end{aligned}$$

where we have applied the inequality $ab \leq \frac{1}{2}(a^2+b^2), a, b \in \mathbf{R}$ twice and the fact that $E|U_B|^4 < \infty$ (see page 1262 of Székely and Rizzo (2009)). Thus, $E[U_B U'_B V_{id} V'_{id}]$ is always nonnegative and finite.

Next, to show (A.1), we have that from the proof of Theorem 7 in Székely and Rizzo (2009),

$$\begin{aligned} E[U_B U'_B|U,U',V,V'] &= E_{U'}|U-U'|_q + E_U|U'-U|_q \\ &\quad - |U-U'|_q - E|U-U'|_q, \\ E[V_{id} V'_{id}|U,U',V,V'] &= (V-EV)(V'-EV'), \end{aligned}$$

which leads to

$$\begin{aligned} &\mathrm{cov}^2_{B,id}(U,V) \\ &\quad = E[U_B U'_B V_{id} V'_{id}] \\ &\quad = E[E(U_B U'_B V_{id} V'_{id}|U,U',V,V')] \\ &\quad = E[E(U_B U'_B|U,U',V,V')E(V_{id} V'_{id}|U,U',V,V')] \\ &\quad = E[(V-EV)(V'-EV') * (E_{U'}|U-U'|_q \\ &\qquad + E_U|U'-U|_q - |U-U'|_q - E|U-U'|_q)] \\ &\quad = -E[(V-EV)(V'-EV')|U-U'|_q]. \end{aligned}$$

We proceed to show (A.2). By definition, $\mathrm{MDD}^2(V|U) = \frac{1}{c_q}\int_{\mathbf{R}^q} \frac{|g_{V,U}(s)-g_V g_U(s)|^2}{|s|_q^{1+q}} ds$. We write

$$\begin{aligned} |g_{V,U}(s) - g_V g_U(s)|^2 &= |g_{V,U}(s)|^2 - g_{V,U}(s) g_V \bar{g}_U(s) \\ &\quad - \bar{g}_{V,U}(s) g_V g_U(s) + g_V^2 |g_U(s)|^2, \end{aligned}$$

where

$$\begin{aligned} &|g_{V,U}(s)|^2 \\ &\quad = E[V e^{i<s,U>}]E'[V' e^{-i<s,U'>}] = E[VV' e^{i<s,U-U'>}] \\ &\quad = -E[VV'(1-\cos<s,U-U'>)] + E(VV') + A_1, \\ &g_{V,U}(s) g_V \bar{g}_U(s) \\ &\quad = E[V e^{i<s,U>}] \cdot E'(V') \cdot E''[e^{-i<s,U''>}] \\ &\quad = -E[VV'(1-\cos<s,U-U''>)] + E(VV') + A_2, \\ &\bar{g}_{V,U}(s) g_V g_U(s) \\ &\quad = E[V e^{-i<s,U>}] \cdot E'(V') \cdot E''[e^{i<s,U''>}] \\ &\quad = -E[VV'(1-\cos<s,U''-U>)] + E(VV') + A_3, \\ &g_V^2 |g_U(s)|^2 \\ &\quad = E[VV']E[e^{i<s,U-U'>}] \\ &\quad = -E[VV']E(1-\cos<s,U-U'>) + E(VV') + A_4, \end{aligned}$$

with A_1, A_2, A_3, A_4 representing terms that vanish when the integral is evaluated and E', E'' stands for the expectation with respect to (U', V') and (U'', V''), respectively.
Integrating each item above separately and summing up, we obtain

$$\begin{aligned} \mathrm{MDD}^2(V|U) &= -E[VV'|U-U'|_q] + E[VV'|U-U''|_q] \\ &\quad + E[VV'|U''-U|_q] - E[VV'E|U-U'|_q] \\ &= -E[(V-E(V))(V'-E(V'))|U-U'|_q], \end{aligned}$$

which completes our proof for the first property.

2. In view of the expressions of $\mathrm{MDD}(V|U)$, $\mathrm{var}(V)$ and $\mathrm{dvar}(U)$, that is,

$$\begin{aligned} \mathrm{MDD}(V|U)^2 &= E[(V-EV)(V'-EV') \\ &\quad * (E_U|U-U'|_q + E_{U'}|U-U'|_q \\ &\quad - |U-U'|_q - E|U-U'|_q)] \\ \mathrm{var}(V) &= E(V-EV)^2 \\ \mathrm{dvar}(U)^2 &= E(E_U|U-U'|_q + E_{U'}|U-U'|_q \\ &\quad - |U-U'|_q - E|U-U'|_q)^2 \end{aligned}$$

the assertion that $\mathrm{MDC}(V|U) \leq 1$ follows from an application of the Cauchy–Schwarz inequality. Furthermore, it is trivial to see that $\mathrm{MDC}(V|U) \geq 0$. Also, $E(V|U) = E(V)$ if and only if the numerator $\mathrm{MDD}^2(V|U) = ||g_{U,V}(s) - g_V g_U(s)||^2$ of $\mathrm{MDC}^2(V|U)$ is 0, which is equivalent to the fact that $g_{V,U}(s) = g_V g_U(s)$ for almost all s.

3. We introduce the function:

$$F(\rho) = \int_{-\infty}^{\infty} |g_{U,V}(s) - g_U(s) g_V|^2 \frac{ds}{s^2} = \int_{-\infty}^{\infty} |E(V e^{isU})|^2 \frac{ds}{s^2}.$$

If U and V are bivariate standard normal with correlation ρ, then $\mathrm{MDD}^2(V|U) = F(\rho)/c_1 = F(\rho)/\pi$. According to the proof of Theorem 7 in Székely et al. (2007), $\mathrm{dvar}(U)^2 = 4(1-\sqrt{3}+\pi/3)/\pi$. Hence,

$$\mathrm{MDC}^2(V|U) = \frac{\mathrm{MDD}^2(V|U)}{\sqrt{\mathrm{var}(V)^2 \mathrm{dvar}(U)^2}} = \frac{F(\rho)}{\sqrt{4\pi(1-\sqrt{3}+\pi/3)}}.$$

Furthermore, by a straightforward calculation, we can derive that $E\{\exp(isU)V\} = \rho si \exp\{-\frac{1}{2}s^2\}$, which yields that

$$\begin{aligned} F(\rho) &= \int_{-\infty}^{\infty} |\rho si \exp(-0.5s^2)|^2 \frac{ds}{s^2} = \rho^2 \int_{-\infty}^{\infty} \exp(-s^2) ds \\ &= \rho^2 \sqrt{\pi} \end{aligned}$$

Thus the formulas for $\mathrm{MDC}^2(V|U)$ follows.

4. The proof is straightforward by the definition of $\mathrm{MDC}(V|U)$ and is omitted. □

Proof of Theorem 2. The argument is similar to that in the proof of Theorem 1 of Székely et al. (2007). For simplicity, we only consider the case $q = 1$. Note that the integrand in $\|g^n_{V,U}(s) - g^n_V g^n_U(s)\|^2$ involves $|g^n_{V,U}(s)|^2$, $|g^n_V g^n_U(s)|^2$ and $\overline{g^n_{V,U}(s)} g^n_V g^n_U(s)$. For the first, we have

$$g^n_{V,U}(s)\overline{g^n_{V,U}(s)} = -n^{-2}\sum_{k,l=1}^n V_k V_l(1 - \cos((U_k - U_l)s)) + \bar{V}_n^2 + W_1,$$

where W_1 represents terms that vanish when the integral in $\|\cdot\|^2$ is evaluated. The second expression is $g^n_V g^n_U(s)\overline{g^n_V g^n_U(s)} = -n^{-2}\sum_{k,l=1}^n V_k V_l n^{-2}\sum_{k,l=1}^n\{1 - \cos((U_k - U_l)s)\} + \bar{V}_n^2 + W_2$ and the third expression becomes $\overline{g^n_{V,U}(s)} g^n_V g^n_U(s) = -n^{-3}\sum_{k,l,h=1}^n V_k V_h\{1 - \cos((U_h - U_l)s)\} + \bar{V}_n^2 + W_3$, where W_2 and W_3 stand for terms that vanish when the integral is evaluated. The case $q > 1$ can be treated along the same line. Then by Lemma 1 of Székely et al. (2007), we can derive $\|g^n_{V,U}(s) - g^n_V g^n_U(s)\|^2 = -S_{1n} - S_{2n} + 2S_{3n}$, where

$$S_{1n} = n^{-2}\sum_{k,l=1}^n V_k V_l |U_k - U_l|_q,$$

$$S_{2n} = n^{-2}\sum_{k,l=1}^n V_k V_l n^{-2}\sum_{k,l=1}^n |U_k - U_l|_q, \text{ and}$$

$$S_{3n} = n^{-3}\sum_{k,l,h=1}^n V_k V_h |U_h - U_l|_q.$$

Finally, the verification of the algebraic identity $-S_{1n} - S_{2n} + 2S_{3n} = -n^{-2}\sum_{k,l=1}^n A_{kl}B_{kl} =: \mathrm{MDD}_n^2(V|U)$ follows by a straightforward check. □

Proof of Proposition 1. Note that the sample distance covariance for $(X, \widehat{W})$ is $\mathrm{dcov}_n^2(X, \widehat{W}) = \frac{1}{n^2}\sum_{k,l=1}^n A_{kl}B_{kl}$, where $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$ with $a_{kl} = |\widehat{W}_k - \widehat{W}_l|$, $\bar{a}_{k\cdot} = \frac{1}{n}\sum_l a_{kl}$, $\bar{a}_{\cdot l} = \frac{1}{n}\sum_k a_{kl}$, $\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l} a_{kl}$, and $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$ with $b_{kl} = |X_k - X_l|$, $\bar{b}_{k\cdot} = \frac{1}{n}\sum_l b_{kl}$, $\bar{b}_{\cdot l} = \frac{1}{n}\sum_k b_{kl}$, $\bar{b}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l} b_{kl}$. Further note that $\mathrm{MDD}_n^2(\widehat{W}|X) = \frac{1}{n^2}\sum_{k,l=1}^n -\tilde{A}_{kl}B_{kl}$, where $\tilde{A}_{kl} = (\widehat{W}_k - \overline{W}_n)(\widehat{W}_l - \overline{W}_n)$, and $\overline{W}_n = n^{-1}\sum_{k=1}^n(\tau - I(Y_k \le \hat{Q}_\tau)) = 0$.

For $\mathrm{dcov}_n^2(X, \widehat{W})$,

$$\begin{aligned}
a_{kl} &= |I(Y_k \le \hat{Q}_\tau) - I(Y_l \le \hat{Q}_\tau)| = I(Y_k \le \hat{Q}_\tau)I(Y_l > \hat{Q}_\tau) \\
&\quad + I(Y_k > \hat{Q}_\tau)I(Y_l \le \hat{Q}_\tau), \\
\bar{a}_{k\cdot} &= I(Y_k \le \hat{Q}_\tau)\frac{1}{n}\sum_l I(Y_l > \hat{Q}_\tau) + I(Y_k > \hat{Q}_\tau)\frac{1}{n}\sum_l I(Y_l \le \hat{Q}_\tau), \\
&= I(Y_k \le \hat{Q}_\tau)(1-\tau) + I(Y_k > \hat{Q}_\tau)\tau, \\
\bar{a}_{\cdot l} &= I(Y_l \le \hat{Q}_\tau)(1-\tau) + I(Y_l > \hat{Q}_\tau)\tau, \text{ and } \bar{a}_{\cdot\cdot} = 2(1-\tau)\tau,
\end{aligned}$$

which yields that

$$\begin{aligned}
A_{kl} &= a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot} \\
&= I(Y_k \le \hat{Q}_\tau)(1 - I(Y_l \le \hat{Q}_\tau)) + I(Y_l \le \hat{Q}_\tau)(1 - I(Y_k \le \hat{Q}_\tau)) \\
&\quad - I(Y_k \le \hat{Q}_\tau)(1-\tau) - (1 - I(Y_k \le \hat{Q}_\tau))\tau \\
&\quad - I(Y_l \le \hat{Q}_\tau)(1-\tau) - (1 - I(Y_l \le \hat{Q}_\tau))\tau + 2(1-\tau)\tau \\
&= -2I(Y_k \le \hat{Q}_\tau)I(Y_l \le \hat{Q}_\tau) + 2\tau I(Y_k \le \hat{Q}_\tau) + 2\tau I(Y_l \le \hat{Q}_\tau) \\
&\quad - 2\tau^2 = -2\tilde{A}_{kl}.
\end{aligned}$$

So $\mathrm{MDD}_n^2(\widehat{W}|X) = 2\mathrm{dcov}_n^2(X, \widehat{W})$. Applying the above equality, that is, $A_{kl} = -2\tilde{A}_{kl}$, we obtain

$$\begin{aligned}
\mathrm{dvar}_n^2(\widehat{W}) &= \frac{1}{n^2}\sum_{k,l=1}^n A_{kl}^2 = \frac{4}{n^2}\sum_{k,l=1}^n \tilde{A}_{kl}^2 = \frac{4}{n^2}\sum_{k,l=1}^n \widehat{W}_k^2\widehat{W}_l^2 \\
&= [2\mathrm{var}_n(\widehat{W})]^2.
\end{aligned}$$

The result then follows from the definitions of $\mathrm{MDC}_n(\widehat{W}|X)$ and $dCor_n(X, \widehat{W})$. □

Proof of Proposition 2. Note that $|W_l - \widehat{W}_l| = |I(Y_l \le Q_\tau) - I(Y_l \le \widehat{Q}_\tau)| = I(\widehat{Q}_\tau < Y_l \le Q_\tau) + I(Q_\tau < Y_l \le \widehat{Q}_\tau)$. Then $P(\frac{1}{n}\sum_{l=1}^n |\widehat{W}_l - W_l| > \epsilon) \le P(\frac{1}{n}\sum_{l=1}^n |\widehat{W}_l - W_l| > \epsilon, |\widehat{Q}_\tau - Q_\tau| \le \delta) + P(|\widehat{Q}_\tau - Q_\tau| > \delta) =: I_1 + I_2$. For I_2, we apply Serfling (1980) Theorem 2.3.2 and get $I_2 = P(|\widehat{Q}_\tau - Q_\tau| > \delta) \le 2\exp(-2nL(\delta)^2)$, where $L(\delta) = \min\{F_Y(Q_\tau + \delta) - \tau, \tau - F_Y(Q_\tau - \delta)\}$. Under the Assumption (B1), we have $G_1(\delta_0)\delta \le L(\delta) \le G_2(\delta_0)\delta$. Let $\delta = \min(\epsilon/\{4G_2(\delta_0)\}, \delta_0)$, which equals to $\epsilon/\{4G_2(\delta_0)\}$ if $\epsilon < \epsilon_0 = 4G_2(\delta_0)\delta_0$. Then for $\epsilon \in (0, \epsilon_0)$,

$$I_2 \le 2\exp(-2nG_1(\delta_0)^2\delta^2) \le 2\exp\left(-2n\frac{G_1(\delta_0)^2}{16G_2(\delta_0)^2}\epsilon^2\right).$$

Setting $P_{Q_\tau} = P(|Y_l - Q_\tau| \le \delta)$, we can find a bound for I_1 as follows,

$$\begin{aligned}
I_1 &\le P\left(\frac{1}{n}\sum_{l=1}^n I(|Y_l - Q_\tau| \le \delta) > \epsilon\right) \\
&= P\left(\frac{1}{n}\sum_{l=1}^n I(|Y_l - Q_\tau| \le \delta) - P_{Q_\tau} > \epsilon - P_{Q_\tau}\right).
\end{aligned}$$

By Hoeffding's inequality, $I_1 \le \exp(-2(\epsilon - P_{Q_\tau})^2 n)$. Since $P_{Q_\tau} \le 2\delta G_2(\delta_0) \le \epsilon/2$ when $\epsilon \in (0, \epsilon_0)$, then $I_1 \le \exp(-2n\epsilon^2/4)$. Together with the bound for I_2,

$$P\left(\frac{1}{n}\sum_{l=1}^n |\widehat{W}_l - W_l| > \epsilon\right) \le 3\exp(-2c_1 n\epsilon^2) \text{ for some } c_1 > 0.$$

□

[Received April 2013. Revised January 2014.]

REFERENCES

Chiang, A., Beck, J., Yen, H., Tayeh, M., Scheetz, T., Swiderski, R., Nishimura, D., Braun, T., Kim, K., Huang, J., et al. (2006), "Homozygosity Mapping With Snp Arrays Identifies Trim32, an e3 Ubiquitin Ligase, as a Bardet-Biedl Syndrome Gene (bbs11)," *Proceedings of the National Academy of Sciences*, 103, 6287–6292. [1314]

Cook, R. D., and Li, B. (2002), "Dimension Reduction for Conditional Mean in Regression," *The Annals of Statistics*, 30, 455–474. [1305,1306]

Delaigle, A., and Hall, P. (2012), "Effect of Heavy Tails on Ultra High Dimensional Variable Ranking Methods," *Statistica Sinica*, 22, 909–932. [1302]

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557. [1302,1307,1308,1313,1314]

Fan, J., Feng, Y., and Wu, Y. (2010), "High-Dimensional Variable Selection for Cox's Proportional Hazards Model" (A Festscrift for Lawrence D. Brown), *IMS Collections*, 6, 70–86. [1302]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space" (with discussions and rejoinder), *Journal of the Royal Statistical Society,* Series B, 70, 849–911. [1302,1307]

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh Dimensional Feature Selection: Beyond The Linear Model," *Journal of Machine Learning Research*, 10, 2013–2038. [1313]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1302]

Gorst-Rasmussen, A., and Scheike, T. (2013), "Independent Screening for Single-Index Hazard Rate Models With Ultra-High Dimensional Features," *Journal of the Royal Statistical Society,* Series B, 75, 217–245. [1302]

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. [1302,1313]

He, X., Wang, L., and Hong, H. G. (2013), "Quantile-Adaptive Model-Free Variable Screening for High-Dimensional Heterogeneous Data," *The Annals of Statistics*, 41, 342–369. [1303,1307,1310,1313]

Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313. [1314]

Li, B., Cook, R. D., and Chiaromonte, F. (2003), "Dimension Reduction for the Conditional Mean in Regressions With Categorical Predictors," *The Annals of Statistics*, 31, 1636–1668. [1306]

Li, P., Peng, H., Zhang, J., and Zhu, L. (2012), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, 40, 1846–1877. [1302]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [1304,1305,1306,1307,1313,1315]

Meier, L., van de Geer, S., and Bühlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [1308,1314]

Redfern, C. H., Degtyarev, M. Y., Kwa, A. T., Salomonis, N., Cotte, N., Nanevicz, T., Fidelman, N., Desai, K., Vranizan, K., Lee, E. K., Coward, P., Shah, N., Warrington, J. A., Fishman, G. I., Bernstein, D., Baker, A. J., and Conklin, B. R. (2000), "Conditional Expression of a Gi-Coupled Receptor Causes Ventricular Conduction Delay and a Lethal Cardiomyopathy," *Proceedings of the National Academy of Sciences*, 97, 4826–4831. [1313]

Scheetz, T., Kim, K., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T., et al. (2006), "Regulation of Gene Expression in the Mammalian Eye and Its Relevance to Eye Disease," *Proceedings of the National Academy of Sciences*, 103, 14429–14434. [1314]

Székely, G. J., and Rizzo, M. L. (2009), "Brownian Distance Covariance," *The Annals of Applied Statistics*, 3, 1236–1265. [1303,1304,1316]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [1302,1303,1304,1305,1315,1316,1317]

Wang, L., Wu, Y., and Li, R. (2012), "Quantile Regression for Analyzing Heterogeneity in Ultra-High Dimension," *Journal of the American Statistical Association*, 107, 214–222. [1314,1315]

Zhao, D., and Li, Y. (2012), "Principled Sure Independence Screening for Cox Models With Ultra-High-Dimensional Covariate," *Journal of Multivariate Analysis*, 105, 397–411. [1302]

Zhu, L., Li, L., Li, R., and Zhu, L. (2011), "Model-Free Feature Screening for Ultrahigh- Dimensional Data," *Journal of the American Statistical Association*, 106, 1464–1474. [1302,1305,1307,1308,1310,1313]