Electronic Journal of Statistics
Vol. 9 (2015) 1492–1517
ISSN: 1935-7524
DOI: 10.1214/15-EJS1047

Partial martingale difference correlation[*]

Trevor Park, Xiaofeng Shao[†] and Shun Yao

University of Illinois at Urbana–Champaign
e-mail: thp2@illinois.edu*;* xshao@illinois.edu*;* shunyao2@illinois.edu

Abstract: We introduce the partial martingale difference correlation, a scalar-valued measure of conditional mean dependence of Y given X, adjusting for the nonlinear dependence on Z, where X, Y and Z are random vectors of arbitrary dimensions. At the population level, partial martingale difference correlation is a natural extension of partial distance correlation developed recently by Székely and Rizzo [14], which characterizes the dependence of Y and X, after controlling for the nonlinear effect of Z. It extends the martingale difference correlation first introduced in Shao and Zhang [10] just as partial distance correlation extends the distance correlation in Székely, Rizzo and Bakirov [13]. Sample partial martingale difference correlation is also defined building on some new results on equivalent expressions of sample martingale difference correlation. Numerical results demonstrate the effectiveness of these new dependence measures in the context of variable selection and dependence testing.

Keywords and phrases: Distance correlation, nonlinear dependence, partial correlation, variable selection.

Received February 2015.

1. Introduction

Measuring and testing (in)dependence and partial (in)dependence is important in many branches of statistics. To measure the dependence of two random vectors $X \in \mathbf{R}^p$ and $Y \in \mathbf{R}^q$, Székely, Rizzo and Bakirov [13] proposed distance covariance (dCov) and distance correlation (dCor), which have attracted a lot of attention lately; see related work by Śzekely and Rizzo [12], Li, Zhong and Zhu [7], Kong et al. [4], Lyons [8], Sejdinovick et al. [9], Sheng and Yin [11], Dueck et al. [2], Shao and Zhang [10], and Székely and Rizzo [14] among others for further extensions and applications of these concepts. In particular, Shao and Zhang [10] proposed the notion of martingale difference divergence (MDD, hereafter) and martingale difference correlation (MDC, hereafter) to measure the conditional mean (in)dependence of Y given X (Y is said to be conditionally mean independent of X provided that $E(Y|X) = E(Y)$ almost surely). Conditional mean dependence plays an important role in statistics. As Cook and Li [1] stated, "in many situations regression analysis is mostly concerned with inferring about the conditional mean of the response given the predictors,

[*] We want to thank the referees for their helpful comments.

[†] Shao would like to acknowledge partial financial support from an NSF grant.

and less concerned with the other aspects of the conditional distribution". In practice, it can occur that Z is a variable that has been known a priori to contribute to the variation of Y, and our main interest is to know if X contributes to the conditional mean of Y (i.e. if Y is conditional mean dependent on X) after adjusting for the (possibly nonlinear) effect of Z.

In this paper, our goal is to develop a scalar-valued measure of conditional mean independence of Y given X, controlling for the third random vector Z, where X, Y and Z are in arbitrary, not necessarily equal dimensions. Our development follows that in Székely and Rizzo [14], where the partial distance covariance and partial distance correlation coefficient (pdCov and pdCor, hereafter) are developed to measure the dependence of Y and X after removing their respective dependence on $Z \in \mathbf{R}^r$. Owing to this connection, we name them partial MDD and partial MDC (pMDD and pMDC, hereafter).

The main contribution of this paper is two-fold: (1) We discover an equivalent expression for MDD at both population and sample levels and an important connection to the distance covariance. The new expression makes it easier to prove a fundamental representation result concerning the sample MDD. A natural extension of MDD to allow for random vectors for both Y and X is also presented. (In Shao and Zhang [10], Y is restricted to be a one-dimensional random variable.) (2) We propose partial MDD and MDC as an extension of partial dCov and partial dCor at both population and sample levels. Our definition of partial MDD differs from that of partial dCov in that the role of Y and X are asymmetric in quantifying the conditional mean dependence of Y on X controlling for Z. Furthermore, we provide an unbiased estimator of the squared MDD using the $\mathcal{U}$-centering idea (Székely and Rizzo [14]) and find equivalent expressions for partial MDC at both population and sample levels. Numerical results are provided in Section 5 to show that a permutation-based test of zero partial MDC has accurate size and respectable power in finite sample. Also a data illustration is provided to demonstrate the effectiveness of pMDC-based forward variable selection approach as compared to pdCor-based counterpart. Section 6 concludes, and technical details are gathered in the Appendix 6.

2. Essentials of dCov and dCor

As proposed by Székely, Rizzo and Bakirov [13], the (population) dCov of two random vectors $X \in \mathbf{R}^p$ and $Y \in \mathbf{R}^q$ with finite first moments is $\mathcal{V}(X,Y)$, the non-negative square root of

$$\mathcal{V}^2(X,Y) \;=\; \int_{\mathbf{R}^{p+q}} |\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2 \frac{1}{c_p c_q \, |s|_p^{1+p}\, |t|_q^{1+q}}\, dt\, ds \tag{2.1}$$

where ϕ_X, ϕ_Y, and $\phi_{X,Y}$ are the individual and joint characteristic functions of X and Y, and $c_p = \pi^{(1+p)/2}/\Gamma((1+p)/2)$. Throughout the paper, $|\cdot|_p$ and $|\cdot|_q$ are the (possibly complex) Euclidean norms defined by, for example,

$$|x|_p \;=\; \sqrt{\overline{x}^T x} \;=\; \sqrt{x^H x}$$

where x^H denotes the conjugate transpose of $x \in \mathbf{C}^p$. In the special case of $\mathbf{C}^1$, we simply denote the modulus as $|x|$. The dCov characterizes independence in the sense that $\mathcal{V}(X,Y) = 0$ if and only if X and Y are independent.

When X and Y have finite *second* moments, it can be shown that

$$\begin{aligned} \mathcal{V}^2(X,Y) \;&=\; E(|X-X'|_p|Y-Y'|_q) \,+\, E(|X-X'|_p)E(|Y-Y'|_q) \\ &\quad -\, 2E(|X-X'|_p|Y-Y''|_q) \end{aligned} \tag{2.2}$$

where (X',Y') and (X'',Y'') are iid copies of (X,Y). (See Székely and Rizzo [12], Theorems 7 and 8.) The (population) dCor of X and Y is $\mathcal{R}(X,Y)$, defined as the nonnegative number satisfying

$$\mathcal{R}^2(X,Y) \;=\; \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y)}}$$

provided the denominator is positive, and zero otherwise. Both dCov and dCor have readily-computable sample analogues, which may be used in general tests for independence.

3. Some properties of MDD and MDC

Martingale difference divergence (MDD) and martingale difference correlation (MDC) are intended to measure departure from the relationship

$$E(Y|X) = E(Y) \quad \text{almost surely}$$

for $Y \in \mathbf{R}^q$ and $X \in \mathbf{R}^p$. Shao and Zhang [10] proposed these measures in the case $q = 1$, by extending the distance covariance and distance correlation proposed in Székely, Rizzo and Bakirov [13]. The following brief review of MDD and MDC also generalizes Shao and Zhang [10] by allowing for $q > 1$.

The martingale difference divergence $MDD(Y|X)$ for real random vectors $X \in \mathbf{R}^p$ and $Y \in \mathbf{R}^q$ is defined to be the nonnegative number satisfying

$$MDD(Y|X)^2 \;=\; \int_{\mathbf{R}^p} \frac{|E(Ye^{i\langle s,X\rangle}) \,-\, E(Y)\,E(e^{i\langle s,X\rangle})|_q^2}{c_p\,|s|_p^{1+p}}\,ds.$$

Compared to the expression of the squared distance covariance in (2.1), the squared MDD uses the same form of weighting function and thus forms a natural extension. In Shao and Zhang [10], it has been shown that the MDD and MDC inherit a number of useful properties of dCov and dCor, including the following:

Proposition 3.1. *If $E|Y|_q^2 + E|X|_p^2 < \infty$, then*

1. *letting (X',Y') be an iid copy of (X,Y),*

$$MDD(Y|X)^2 \;=\; -E[(Y-E(Y))^T(Y'-E(Y'))|X-X'|_p] \tag{3.1}$$

2. *$MDD(Y|X) = 0$ if and only if $E(Y|X) = E(Y)$ almost surely.*

In the case $q = 1$ these follow from Theorem 1 of Shao and Zhang [10], and extension to the case $q > 1$ is straightforward.

The martingale difference correlation $MDC(Y|X)$ is the nonnegative number satisfying

$$MDC(Y|X)^2 = \frac{MDD(Y|X)^2}{\sqrt{E[(Y-E(Y))^T(Y'-E(Y'))]^2\,\mathcal{V}^2(X,X)}}$$

when the denominator is positive, and zero otherwise. This coincides with the definition in Shao and Zhang [10] for the case $q = 1$. Following the same argument in the proof of Theorem 1 of Shao and Zhang [10], it can be shown that $MDC(Y|X) \in [0,1]$. Since there is no additional novelty, we skip the details.

3.1. Connection between MDD and dCov

In this subsection, we provide an alternative formulation of MDD, using the Laplace operator, which relates it more closely to the formula for distance covariance (Székely, Rizzo and Bakirov [13]). Furthermore, this new formulation makes it easier to prove a fundamental representation result concerning the empirical (sample) version of MDD.

Denote the gradient of a (possibly complex) function f of a real vector $x \in \mathbf{R}^p$ as $\nabla_x f$ and the Hessian as $H_x(f) = \nabla_x^2 f$. Define the Laplace operator (Laplacian) of f to be

$$\Delta_x f \;=\; \sum_{i=1}^{p} \frac{\partial^2 f}{\partial x_i^2} \;=\; \mathrm{trace}(H_x(f))$$

Proposition 3.2. *If Y has finite second moments, for any $s \in \mathbf{R}^p$,*

$$|E(Ye^{i\langle s,X\rangle}) - E(Y)\,E(e^{i\langle s,X\rangle})|_q^2 \;=\; \frac{1}{2}\Delta_t |\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2 \Big|_{t=0}. \qquad (3.2)$$

The proof of proposition 3.2 is in the appendix. It follows that

$$MDD(Y|X)^2 \;=\; \int_{\mathbf{R}^p} \left(\frac{1}{2}\,\Delta_t\,|\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2 \Big|_{t=0}\right) \frac{1}{c_p\,|s|_p^{1+p}}\,ds$$

which has a very similar form to the squared distance covariance

$$\mathcal{V}^2(X,Y) \;=\; \int_{\mathbf{R}^p} \left(\int_{\mathbf{R}^q} |\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2 \frac{1}{c_q\,|t|_q^{1+q}}\,dt \right) \frac{1}{c_p\,|s|_p^{1+p}}\,ds$$

Conceptually, $\mathcal{V}^2(X,Y)$ is weighted everywhere in both s and t, whereas $MDD(Y|X)^2$ is weighted everywhere in s, but depends only on particular behavior local to $t = 0$. This is conceptually sensible, because only first-moment information about Y is being used in $MDD(Y|X)^2$. Indeed, we can further relate this to the ordinary (squared) covariance. For example, when $p = q = 1$, it is true that (when X and Y have sufficiently many moments)

$$(E(YX) \;-\; E(Y)\,E(X))^2 \;=\; \frac{1}{2}\Delta_s\left(\frac{1}{2}\Delta_t\,|\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2 \Big|_{t=0}\right)\Big|_{s=0}$$

The new formulation (3.2) leads to a straightforward proof of a fundamental representation for *sample* MDD. Let $\phi^n_{X,Y}(s,t)$, $\phi^n_X(s)$, $\phi^n_Y(t)$, be the empirical characteristic functions, joint and marginal, based on averaging over a sample $(x_1, y_1), \ldots (x_n, y_n)$ of size n:

$$\phi^n_{X,Y}(s,t) \;=\; \frac{1}{n}\sum_{k=1}^{n} e^{i\langle s,x_k\rangle + i\langle t,y_k\rangle}$$

$$\phi^n_X(s) \;=\; \frac{1}{n}\sum_{k=1}^{n} e^{i\langle s,x_k\rangle} \qquad \phi^n_Y(t) \;=\; \frac{1}{n}\sum_{k=1}^{n} e^{i\langle t,y_k\rangle}$$

Let $MDD_n(Y|X)^2$ be the empirical squared martingale difference divergence, based on these empirical characteristic functions:

$$MDD_n(Y|X)^2 \;=\; \int_{\mathbf{R}^p} \left(\frac{1}{2}\,\Delta_t\,|\phi^n_{X,Y}(s,t) - \phi^n_X(s)\phi^n_Y(t)|^2\,\Big|_{t=0} \right) \frac{1}{c_p\,|s|_p^{1+p}}\,ds$$

This agrees with the definition in Shao & Zhang [10], according to Theorem 2 therein, along with the previous proposition (applied to the empirical characteristic functions).

Proposition 3.3.

$$MDD_n(Y|X)^2 \;=\; \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} A_{kl} B^{\circ}_{kl} \tag{3.3}$$

where A and B° are the double centered versions of the matrices that have elements

$$a_{kl} \;=\; |x_k - x_l|_p \qquad \text{and} \qquad b^{\circ}_{kl} \;=\; \frac{1}{2}|y_k - y_l|_q^2$$

respectively, that is $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$, $\bar{a}_{k\cdot} = n^{-1}\sum_{j=1}^{n} a_{kj}$, $\bar{a}_{\cdot l} = n^{-1}\sum_{j=1}^{n} a_{jl}$, $\bar{a}_{\cdot\cdot} = n^{-2}\sum_{k,l=1}^{n} a_{kl}$ and similarly for B°_{kl}.

It is straightforward to show that B° can be replaced with a matrix B^* that is the double centered version of the matrix with elements $b^*_{kl} = -y_k^T y_l$, which compares directly with the definition of $MDD_n(Y|X)$ by Shao and Zhang [10]. The advantage of (3.3) is that it can be proven more directly and it shows $MDD_n(Y|X)^2$ depends purely on Euclidean distances.

By expanding and canceling terms and using (3.1), it can be shown that (assuming finite *third* moments)

$$\begin{aligned} MDD(Y|X)^2 \;&=\; E\big(|X - X'|_p\,\tfrac{1}{2}|Y - Y'|_q^2\big) \;+\; E(|X - X'|_p)E\big(\tfrac{1}{2}|Y - Y'|_q^2\big) \\ &\quad -\; 2E\big(|X - X'|_p\,\tfrac{1}{2}|Y - Y''|_q^2\big) \end{aligned} \tag{3.4}$$

where (X', Y') and (X'', Y'') are iid copies of (X, Y). This compares with (2.2), thus making another connection between MDD and dCov.

3.2. Unbiased estimation of MDD

In general, $MDD_n(Y|X)^2$ is a biased estimator of $MDD(Y|X)^2$. When developing the partial distance covariance, Székely and Rizzo [14] introduced $\mathcal{U}$-centering, which seems essential and leads to unbiased estimator of squared distance covariance.

Let $A = (a_{ij})$ be a symmetric, real valued $n \times n$ matrix with zero diagonal, with $n > 3$. Define the $\mathcal{U}$-centered matrix $\widetilde{A}$ as having (i, j)-th entry

$$\widetilde{A}_{ij} \quad = \quad \begin{cases} a_{ij} - \frac{1}{n-2}\sum_{l=1}^{n} a_{il} - \frac{1}{n-2}\sum_{k=1}^{n} a_{kj} + \frac{1}{(n-1)(n-2)}\sum_{k,l=1}^{n} a_{kl}, & i \neq j \\ 0, & i = j \end{cases} \tag{3.5}$$

Let S_n denote the linear span of all $n \times n$ distance matrices of samples $\{x_1, \dots, x_n\}$. Any $A \in S_n$ is a real valued, symmetric matrix with zero diagonal. Let $H_n = \{\widetilde{A} | A \in S_n\}$. Following Székely and Rizzo [14], we define the inner product of $\widetilde{A}$ and $\widetilde{B}$ in H_n as

$$(\widetilde{A} \cdot \widetilde{B}) = \frac{1}{n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} \widetilde{B}_{ij} \tag{3.6}$$

and $|\widetilde{A}| = (\widetilde{A} \cdot \widetilde{A})^{1/2}$ as the norm of $\widetilde{A}$. Theorem 1 in Székely and Rizzo [14] shows that the linear span of all matrices in H_n is a Hilbert space with inner product defined in (3.6).

This inner product is useful because it defines an unbiased estimator of squared population dCov (see Proposition 1 of Székely and Rizzo [14]). Below we shall introduce an unbiased estimator for $MDD^2(Y|X)$, which is crucial for the development of partial MDD and partial MDC in Section 4.

Given a random sample $(x_1, y_1), \dots, (x_n, y_n)$ from the joint distribution of (X, Y), where $n > 3$, define

$$a_{ij} \;=\; |x_i - x_j|_p \qquad b^{\circ}_{ij} \;=\; \frac{1}{2}|y_i - y_j|_q^2 \qquad b^{*}_{ij} \;=\; -y_i^T y_j$$

Define $\widetilde{A}$, $\widetilde{B}^{\circ}$, and $\widetilde{B}^{*}$ to be the $\mathcal{U}$-centered matrices based on (a_{ij}), (b°_{ij}), and (b^{*}_{ij}). (Even though (b^{*}_{ij}) generally does not have a zero diagonal, the matrix $\widetilde{B}^{*}$ may still be formally defined as in (3.5). Obviously (b°_{ij}) has a zero diagonal, so it better fits the context defined by Székely & Rizzo.) Below we assert that $(\widetilde{A} \cdot \widetilde{B}^{\circ})$ is an unbiased estimate of the squared martingale difference divergence $MDD(Y|X)^2$, for Y and X in arbitrary dimensions.

Proposition 3.4.

$$E\big[(\widetilde{A} \cdot \widetilde{B}^{\circ})\big] \;=\; MDD(Y|X)^2$$

Perhaps surprisingly, $\widetilde{B}^{\circ}$ can be replaced with $\widetilde{B}^{*}$, and the result remains true. Indeed,

Proposition 3.5.

$$(\widetilde{A} \cdot \widetilde{B}^{*}) \;=\; (\widetilde{A} \cdot \widetilde{B}^{\circ})$$

Based on the $\mathcal{U}$-centering, we introduce the unbiased estimator of $MDD(Y|X)^2$ as

$$\widetilde{MDD}_n(Y|X)^2 := (\widetilde{A} \cdot \widetilde{B}^\circ) = \frac{1}{n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} \widetilde{B}^\circ_{ij}$$

and the sample MDC is defined by

$$\widetilde{MDC}_n(Y|X)^2 := \frac{\widetilde{MDD}_n(Y|X)^2}{|\widetilde{A}||\widetilde{B}^\circ|} = \frac{(\widetilde{A} \cdot \widetilde{B}^\circ)}{|\widetilde{A}||\widetilde{B}^\circ|}$$

4. Partial MDD and partial MDC

Following the development in Székely and Rizzo [14], we shall present population and sample versions of partial MDD and partial MDC below.

4.1. Population pMDD and pMDC

Let $C_Z = C_Z(Z, Z')$ denote the (random) double centered version of $c(z, z') = |z - z'|_r$ with respect to Z, where

$$\begin{aligned} C_Z(z, z') &= c(z, z') - \int_{\mathbf{R}^r} c(z, z') dF_Z(z') - \int_{\mathbf{R}^r} c(z, z') dF_Z(z) \\ &\quad + \int_{\mathbf{R}^r} \int_{\mathbf{R}^r} c(z, z') dF_Z(z) dF_Z(z'), \end{aligned}$$

provided that the integrals exist. Correspondingly, let A_X and B_Y denote the double centered version of $a(x, x') = |x-x'|_p$ and $b(y, y') = |y-y'|_q$, respectively. Székely and Rizzo [14] define

$$\mathcal{V}^2(X, Y) = E(A_X B_Y)$$

and indicate that this is equivalent to the original definition (2.1).

Let $b^\circ(y, y') = \frac{1}{2}|y - y'|_q^2$ for $y, y' \in \mathbf{R}^q$ and its double centered version with respect to Y by

$$\begin{aligned} B_Y^\circ(y, y') &= b^\circ(y, y') - \int_{\mathbf{R}^q} b^\circ(y, y') dF_Y(y') - \int_{\mathbf{R}^q} b^\circ(y, y') dF_Y(y) \\ &\quad + \int_{\mathbf{R}^q} \int_{\mathbf{R}^q} b^\circ(y, y') dF_Y(y') dF_Y(y) \\ &= \frac{1}{2}(y - y')^T(y - y') - \frac{1}{2}E_{Y'}\{(y - Y')^T(y - Y')\} \\ &\quad - \frac{1}{2}E_Y\{(Y - y')^T(Y - y')\} + \frac{1}{2}E[(Y - Y')^T(Y - Y')] \\ &= -(y - E(Y))^T(y' - E(Y)) \end{aligned}$$

after a straightforward calculation. Here E_Y denotes the expectation with respect to the random vector Y. Let $B_Y^\circ = B_Y^\circ(Y, Y')$. Note that $b^\circ(y, y')$ is a dissimilarity function.

Then in view of (3.1), it is not hard to show that

$$MDD(Y|Z)^2 = E(B_Y^\circ C_Z)$$

To define the partial MDD at the population level, consider the usual L^2 space of random variables with finite second moment, having inner product $(U \cdot V) = E[UV]$. We assume A_X, B_Y°, and C_Z are in this space.

Let $\beta = \frac{MDD(Y|Z)^2}{\mathcal{V}^2(Z,Z)}$ and $\beta = 0$ if $\mathcal{V}^2(Z,Z) = 0$. We first define the projection of B_Y° onto the orthogonal complement of C_Z as $\widetilde{P}_{Z^\perp}(Y) = B_Y^\circ - \beta C_Z$ and $\widetilde{P}_{Z^\perp}(Y) = B_Y^\circ$ if $\mathcal{V}^2(Z,Z) = 0$. Then it is easy to see that

$$(\widetilde{P}_{Z^\perp}(Y) \cdot C_Z) = E(\widetilde{P}_{Z^\perp}(Y) C_Z) = 0$$

using the relation $\mathcal{V}^2(Z,Z) = E(C_Z^2)$ (see equation (4.2) of Székely and Rizzo [14]). Thus in a sense $\widetilde{P}_{Z^\perp}(Y)$ corresponds to $U := Y - E(Y|Z)$ since $E(U|Z) = E(U) = 0$. Next we define $W = (X^T, Z^T)^T \in \mathbf{R}^{p+r}$. One way to measure the additional contribution of X to the conditional mean of Y controlling for Z, is to measure $E(U|W)$.

Definition 4.1. The population partial MDD of Y given X, after controlling for the effect of Z, i.e., $pMDD(Y|X;Z)$ is defined as

$$pMDD(Y|X;Z) = (\widetilde{P}_{Z^\perp}(Y) \cdot D_W) = E[\widetilde{P}_{Z^\perp}(Y) D_W],$$

where D_W is the random double centered version of $d(w,w') = |w - w'|_{p+r}$ with respect to W. The population partial martingale difference correlation is defined as

$$pMDC(Y|X;Z) = \frac{(\widetilde{P}_{Z^\perp}(Y) \cdot D_W)}{\left((\widetilde{P}_{Z^\perp}(Y) \cdot \widetilde{P}_{Z^\perp}(Y)) \times \mathcal{V}^2(W,W)\right)^{1/2}}$$

If $(\widetilde{P}_{Z^\perp}(Y) \cdot \widetilde{P}_{Z^\perp}(Y)) \times \mathcal{V}^2(W,W) = 0$, then we define $pMDC(Y|X;Z) = 0$.

From Proposition 3.1, $E(U|W) = E(Y|X,Z) - E(Y|Z) = 0$ almost surely iff $MDD(U|W)^2 = 0$ iff $E[B_U^\circ D_W] = 0$. Since there is no random sample corresponding to U, there is no direct plug-in estimate for B_U°. Here we use $\widetilde{P}_{Z^\perp}(Y)$ as a surrogate. Further note that if $\beta = 0$, then $\widetilde{P}_{Z^\perp}(Y) = B_Y^\circ$ and

$$\begin{aligned} pMDC(Y|X;Z) &= \frac{E(B_Y^\circ D_W)}{\left(E(B_Y^\circ B_Y^\circ) \times \mathcal{V}^2(W,W)\right)^{1/2}} \\ &= \frac{MDD(Y|W)^2}{\left(E[(Y - E(Y))^T(Y' - E(Y))]^2 \times \mathcal{V}^2(W,W)\right)^{1/2}} \\ &= MDC(Y|W)^2. \end{aligned}$$

Remark 4.1. Alternatively, we could define $pMDD(Y|X;Z)$ as the difference between $MDD(Y|W)^2$ and $MDD(Y|Z)^2$, where the former measures the relationship $E((Y - E(Y))|W) = 0$ whereas the latter measures the relationship

$E((Y - E(Y))|Z) = 0$. The problem with this definition is that the interpretation is not as intuitive and straightforward as that based on the one defined above.

Analogous to Theorem 3 in Székely and Rizzo [14], we can also provide an alternative definition of $pMDC(Y|X;Z)$ below.

Proposition 4.1. *The following definition of population partial MDC is equivalent to Definition 4.1.*

$$pMDC(Y|X;Z) = \begin{cases} \frac{MDC(Y|W)^2 - MDC(Y|Z)^2 R^2(Z,W)}{\sqrt{1-MDC(Y|Z)^4}}, & MDC(Y|Z) \neq 1, \\ 0, & MDC(Y|Z) = 1. \end{cases} \tag{4.1}$$

Remark 4.2. It has been noted in Section 4.2 of Székely and Rizzo [14] that the partial distance correlation $R^*(X,Y;Z) = 0$ is not equivalent to the conditional independence between X and Y given Z, although both the conditional dependence measure and partial dependence measure capture overlapping aspects of dependence. In our setting, a natural notion that corresponds to conditional independence of X and Y given Z is the so-called conditional mean independence of Y given X conditioning on Z, i.e.,

$$E(Y|X,Z) = E(Y|Z),\ a.s. \qquad \text{or equivalently}$$
$$E((Y - E(Y|Z))|X,Z) = E(U|W) = 0,\ a.s.$$

That is, conditioning on Z, the variable X does not contribute to the (conditional) mean of Y. It can be expected that $pMDC(Y|X;Z) = 0$ is not equivalent to conditional mean independence of Y given X conditioning on Z. In particular, we revisit the example given in Section 4.2 of Székely and Rizzo [14]. Let Z_1, Z_2, Z_3 be iid standard normal random variables, $X = Z_1 + Z_3$, $Y = Z_2 + Z_3$ and $Z = Z_3$. Then X and Y are conditionally independent given Z, which implies that Y is conditionally mean independent of X given Z, but $pMDC(Y|X;Z) = 0.04805(0.0004) \neq 0$ based on numerical simulations with sample size 1000 and 10000 replications. On the other hand, it is also possible that $pMDD(Y|X;Z) = 0$ and yet Y is not conditionally mean independent of X conditioning on Z. For example, letting $X, Y \sim i.i.d.$ Bernoulli(0.5) and $Z \mid X, Y \sim \text{Bernoulli}\big(c\min(X,Y)\big)$, it is possible to numerically determine a zero value of $pMDD(Y|X;Z)$ at $c \approx 0.5857839$, for which we have

$$\begin{aligned} E(Y|X=0,Z=0) &= P(Y=1|X=0,Z=0) = 0.5 \\ E(Y|X=1,Z=0) &= P(Y=1|X=1,Z=0) \approx 0.2928945 \end{aligned}$$

Thus the mean of Y depends on X, even after conditioning on Z.

Recently, conditional distance correlation (Wang et al. [15]) was proposed to measure the dependence between Y and X conditioning on Z, and its extension to measure the conditional mean dependence of Y given X conditioning on Z would be very interesting. It is worth noting that our sample pMDC and pMDD can be easily calculated without any choice of a bandwidth parameter, whereas a scalar-valued metric that quantifies the conditional mean dependence of Y given X, conditioning on Z, presumably has to involve a bandwidth parameter, the choice of which can be difficult.

4.2. Sample pMDD and pMDC

Given the sample $(x_i^T, y_i^T, z_i^T)_{i=1}^n$, we want to define sample partial MDD and MDC, denoted as $pMDD_n(Y|X;Z)$ and $pMDC_n(Y|X;Z)$ as sample analogs of population partial MDD and partial MDC. Let $w_i = (x_i^T, z_i^T)^T$, $i = 1, \dots, n$ and let B°, C and D be $n \times n$ matrices with entries $B_{ij}^\circ = \frac{1}{2}|y_i - y_j|_q^2$, $C_{ij} = |z_i - z_j|_r$, and $D_{ij} = |w_i - w_j|_{p+r}$, respectively. We use $\widetilde{B}^\circ$, $\widetilde{C}$ and $\widetilde{D}$ to denote the $\mathcal{U}$-centered versions of B°, C and D, respectively. Then the sample analogs of B_Y°, β and C_Z are $\widetilde{B}^\circ$, $\frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C} \cdot \widetilde{C})}$, $\widetilde{C}$, respectively and the sample counterpart of $\widetilde{P}_{Z^\perp}(Y)$ is $\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C} \cdot \widetilde{C})}\widetilde{C}$.

Definition 4.2. Given a random sample from the joint distribution (X^T, Y^T, Z^T), the sample partial martingale difference divergence of Y given X, controlling for the effect of Z, is given by

$$pMDD_n(Y|X;Z) = \left(\left(\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C} \cdot \widetilde{C})}\widetilde{C}\right) \cdot \widetilde{D}\right)$$

assuming $(\widetilde{C} \cdot \widetilde{C}) \neq 0$ and $(\widetilde{B}^\circ \cdot \widetilde{D})$ otherwise. The sample martingale difference correlation $pMDC_n(Y|X;Z)$ is defined as

$$pMDC_n(Y|X;Z) = \frac{pMDD_n(Y|X;Z)}{\left|\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C} \cdot \widetilde{C})}\widetilde{C}\right| \left|\widetilde{D}\right|}$$

if $|\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C} \cdot \widetilde{C})}\widetilde{C}||\widetilde{D}| \neq 0$ and otherwise $pMDC_n(Y|X;Z) = 0$.

Proposition 4.2. *An equivalent computing formula for $pMDC_n(Y|X;Z)$ in Definition 4.2 is*

$$pMDC_n(Y|X;Z) = \begin{cases} \frac{\widetilde{MDC}_n(Y|W)^2 - \widetilde{MDC}_n(Y|Z)^2 \widetilde{R}_n^2(Z,W)}{\sqrt{1 - \widetilde{MDC}_n(Y|Z)^4}}, & \widetilde{MDC}_n(Y|Z) \neq 1 \\ 0, & \widetilde{MDC}_n(Y|Z) = 1 \end{cases} \tag{4.2}$$

Remark 4.3. In Shao and Zhang [10], the conditional quantile dependence of univariate Y given X at the τth level has been quantified using MDD and MDC by noting that $Q_\tau(Y|X) = Q_\tau(Y)$ almost surely if and only if $E(V_\tau|X) = E(V_\tau) = 0$ almost surely, where $Q_\tau(Y)$ and $Q_\tau(Y|X)$ are the unconditional and conditional τth quantiles of Y, and $V_\tau = \tau - \mathbf{1}(Y \leq Q_\tau(Y))$. Similarly, we can measure the so-called partial τth quantile dependence of Y given X, controlling for the effect of Z, by using $pMDD(V_\tau|X;Z)$ or $pMDC(V_\tau|X;Z)$. Their sample versions can be defined accordingly, and the details are omitted. It is worth noting that in a recent paper by Li et al. [6], quantile partial correlation was introduced to measure the conditional quantile dependence of a random variable Y given X in a linear way, controlling for the linear effect of Z. In contrast, our quantile partial MDC measures the nonlinear dependence of Y at τth quantile level on X after adjusting for the nonlinear effect of Z. Our sample pMDC can be readily calculated without fitting quantile regression models.

We also notice that both the population and sample pMDD (pMDC) could be negative, just like the pdCov (pdCor) proposed in Székely and Rizzo [14]. We illustrate this through the following example. Suppose that

$$X, Y \sim i.i.d.\ \text{Bernoulli}(0.5), \qquad Z = \min(X, Y)$$

Then, after straightforward but laborious computations, it can be shown that $pMDD(Y|X;Z) = \big(4 - 3\sqrt{2}\big)/144 \approx -0.001685$. Therefore, the sample counterpart may also be negative, since it is consistent.

5. Numerical studies

In the first three examples, we examine tests of the null hypothesis of zero pMDD. We compare our test with the partial distance covariance test (pdCov) by Székely and Rizzo [14] and partial correlation test (pcor), which is implemented as a t test as described in Legendre [5]. Both the pMDD test and pdCov test are implemented as permutation tests, in which we permute the sample X in order to approximate the sampling distribution of the test statistic under the null. Specifically, in each permutation test, we generate $R = 999$ replicates by permuting the sample X and calculate the observed test statistic T_0 with the original data and test statistic $T^{(i)}$ corresponding to the i-th permutation. The estimated p-value is computed as

$$\hat{p} = \frac{1 + \sum_{i=1}^{R} \mathbf{1}(T^{(i)} \geq T_0)}{R + 1}$$

where $\mathbf{1}$ is the indicator function. The significance level is α and we reject the null if $\hat{p} \leq \alpha$. The type I error rate and power are estimated via 10,000 Monte Carlo replications.

Example 5.1. The settings of this example are adapted from Examples 2–5 in Székely and Rizzo [14].

1.a: Let X, Y and Z be three independent random variables, each of which follows a standard normal distribution.

1.b: Replace X in 1.a with an independent standard lognormal random variable.

1.c: X, Y and Z are generated from a multivariate normal distribution with marginal distributions as standard normal, and the pairwise correlations as $\rho(X, Y) = \rho(Y, Z) = \rho(Z, X) = 0.5$.

1.d: Replace X in 1.c by a standard lognormal random variable such that the pairwise correlations are $\rho(\log X, Y) = \rho(Y, Z) = \rho(Z, \log X) = 0.5$.

Table 1

Type I error rate and power at nominal significance level α for Example 5.1

	n	α	pdCov	pMDD	pcor	α	pdCov	pMDD	pcor
1.a	10	0.05	0.0485	0.0502	0.0900	0.01	0.0978	0.1012	0.1432
	20	0.05	0.0534	0.0508	0.0659	0.01	0.1051	0.1003	0.1198
	30	0.05	0.0517	0.0514	0.0616	0.01	0.1027	0.1052	0.1162
	50	0.05	0.0511	0.0525	0.0608	0.01	0.1032	0.1023	0.1122
	100	0.05	0.0503	0.0514	0.0532	0.01	0.0990	0.1020	0.1058
1.b	10	0.05	0.0475	0.0481	0.0887	0.01	0.0962	0.0988	0.1438
	20	0.05	0.0549	0.0514	0.0688	0.01	0.1014	0.1010	0.1201
	30	0.05	0.0516	0.0501	0.0582	0.01	0.1028	0.1011	0.1127
	50	0.05	0.0537	0.0536	0.0577	0.01	0.1034	0.1014	0.1082
	100	0.05	0.0526	0.0510	0.0540	0.01	0.1000	0.0995	0.1032
1.c	10	0.05	0.2152	0.2002	0.2144	0.01	0.3256	0.3151	0.2995
	20	0.05	0.4626	0.4454	0.3390	0.01	0.5890	0.5723	0.4455
	30	0.05	0.6569	0.6245	0.4633	0.01	0.7605	0.7353	0.5780
	50	0.05	0.8717	0.8471	0.6819	0.01	0.9223	0.9036	0.7814
	100	0.05	0.9923	0.9874	0.9309	0.01	0.9962	0.9929	0.9633
1.d	10	0.05	0.2031	0.1823	0.1872	0.01	0.3151	0.2853	0.2622
	20	0.05	0.4290	0.3960	0.2434	0.01	0.5526	0.5228	0.3480
	30	0.05	0.6090	0.5646	0.3294	0.01	0.7214	0.6760	0.4382
	50	0.05	0.8379	0.8044	0.4820	0.01	0.8974	0.8662	0.6097
	100	0.05	0.9873	0.9738	0.7467	0.01	0.9930	0.9864	0.8335

Cases 1.a and 1.b demonstrate the size of different tests at nominal level α, whereas cases 1.c and 1.d show the power. Table 1 shows that partial correlation test's size (pcor) is inflated when sample size is relatively small for both normal and non-normal cases, and the size for pdCov and pMDD are reasonably close to the nominal level α. This is consistent with the findings in Székely and Rizzo [14]. For the power comparison in 1.c and 1.d, Table 1 also shows that the pdCov has the highest power, and pMDD's power is only slightly lower than pdCov. Both tests have superior power performance over pcor.

Example 5.2. This example examines the case of negative correlation.

2.a: X, Y and Z are generated from a multivariate normal distribution with marginal distributions as standard normal and the pairwise correlations as $\rho(X,Y) = \rho(Y,Z) = \rho(Z,X) = -0.48$.

2.b: Replace X in 2.a with a standard lognormal random variable such that the pairwise correlations are $\rho(\log X, Y) = \rho(Y,Z) = \rho(Z, \log X) = -0.48$.

2.c: X, Y and Z are generated from a multivariate t distribution with marginal distributions as student-t with degree of freedom three and the pairwise correlations as $\rho(X,Y) = \rho(Y,Z) = \rho(Z,X) = -0.48$.

From Table 2 we observe that overall pcor has the highest power; pMDD consistently outperform pdCov in all configurations and it is comparable to pcor when sample size is large.

Example 5.3. Generate $X = Z^2 + \epsilon_1$ and $Y = 2Z + \epsilon_2 X$, where Z, ϵ_1 and ϵ_2 are iid standard normals.

Table 2
Power at nominal significance level α for Example 5.2

case	n	α	pdCov	pMDD	pcor	α	pdCov	pMDD	pcor
2.a	10	0.05	0.279	0.406	0.993	0.10	0.394	0.544	0.996
	20	0.05	0.504	0.738	1.000	0.10	0.608	0.815	1.000
	30	0.05	0.644	0.874	1.000	0.10	0.723	0.914	1.000
	50	0.05	0.801	0.973	1.000	0.10	0.852	0.982	1.000
	100	0.05	0.949	0.999	1.000	0.10	0.962	0.999	1.000
2.b	10	0.05	0.266	0.366	0.938	0.10	0.376	0.508	0.966
	20	0.05	0.471	0.661	0.996	0.10	0.580	0.760	0.998
	30	0.05	0.611	0.805	1.000	0.10	0.695	0.866	1.000
	50	0.05	0.777	0.935	1.000	0.10	0.828	0.956	1.000
	100	0.05	0.934	0.995	1.000	0.10	0.951	0.996	1.000
2.c	10	0.05	0.290	0.400	0.979	0.10	0.400	0.519	0.987
	20	0.05	0.502	0.666	0.998	0.10	0.589	0.732	0.999
	30	0.05	0.629	0.785	1.000	0.10	0.698	0.828	1.000
	50	0.05	0.778	0.889	1.000	0.10	0.820	0.908	1.000
	100	0.05	0.926	0.969	1.000	0.10	0.940	0.975	1.000

Table 3
Probability of rejection for Example 5.3 (Negative partial MDD, zero partial correlation, positive partial distance correlation)

n	α	pdCov	pMDD	pcor	α	pdCov	pMDD	pcor
10	0.05	0.1052	0.0896	0.3026	0.10	0.1720	0.1377	0.3771
20	0.05	0.1662	0.0924	0.3445	0.10	0.2388	0.1284	0.4209
30	0.05	0.2016	0.0773	0.3585	0.10	0.2818	0.1034	0.4386
50	0.05	0.2774	0.0585	0.3892	0.10	0.3616	0.0743	0.4655
100	0.05	0.4239	0.0271	0.4126	0.10	0.4986	0.0332	0.4933

Numerical approximations based on $n = 1000$ and 10000 replications are 0.0253 (0.00014) for pdCor, -0.057 (0.0001) for pMDC, which indicated that after controlling the effect of Z, Y and X are still dependent with positive pdCor but pMDC is negative. It can be seen from Table 3 that for $n = 100$, the pdCov shows a substantial amount of rejections, whereas the rejection rate of pMDD is below the nominal level, which is consistent with the fact that $pMDC(Y|X;Z) < 0$ but $pdCor(Y,X|Z) > 0$. The population partial correlation can be easily calculated, that is, $pcor(Y,X|Z) = 0$. From Table 3, however, we see a big size distortion for pcor-based test. This is presumably because the joint distribution of (X,Y,Z) is not Gaussian.

Example 5.4. We consider the same prostate cancer data example used in Székely and Rizzo [14]. The response variable, *lpsa*, is log of the level of prostate specific antigen. Our goal is to predict the response based on one or more predictors from a total of eight predictor variables. For comparison purposes, we standardize each variable first as in Székely and Rizzo [14] and Hastie et al. [3] and use the 67 training data for variable selection. Then prediction error is further reported using the 30 testing data.

The pMDC-based variable selection is implemented as a simple forward selection style combining both partial MDC and MDC as described in Székely

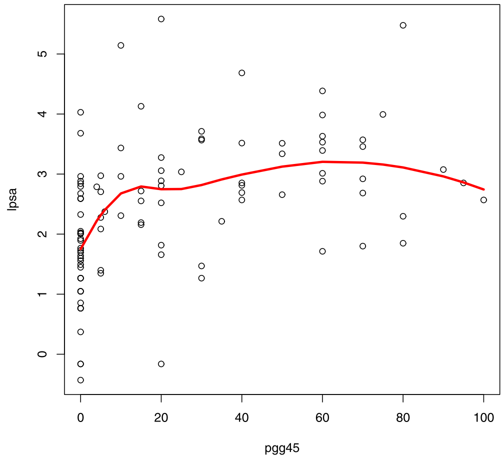

FIG 1. *Scatter plot of lpsa and pgg45 with loess smoother.*

and Rizzo [14] for pdCor. In the first step, we calculate the $MDC(y|x_i)$ and select x_j that has the largest martingale difference correlation. Then we compute $pMDC(y|x_i;x_j)$ for all the variables $x_i \neq x_j$ and also select the x_i for which $pMDC(y|x_i;x_j)$ is the largest. Define the vector w to be the variables that have already been included in the model through previous steps. We continue the procedure by selecting the next variable to be the one that has the largest $pMDC(y|x_i;w)$. The stopping rule for the variable selection procedure is at 5% significance level implemented as a permutation test. The models selected by pMDC, pdCor, best subset method (BIC-based) and lasso are listed below:

$$\begin{aligned}
&\text{pMDC:} && \text{lpsa} \sim \text{lcavol} + \text{lweight} + \text{pgg45} + \text{svi} + \text{lbph};\\
&\text{pdCor:} && \text{lpsa} \sim \text{lcavol} + \text{lweight} + \text{svi} + \text{gleason} + \text{lbph};\\
&\text{best subsets:} && \text{lpsa} \sim \text{lcavol} + \text{lweight};\\
&\text{lasso:} && \text{lpsa} \sim \text{lcavol} + \text{lweight} + \text{svi} + \text{lbph} + \text{pgg45};
\end{aligned}$$

The order for the forward selection with Mallow's C_p is *lcavol, lweight, svi, lbph, pgg45, age, lcp, gleason*. We can see that the selection results are to some extent similar for all the methods listed above, as all select the same top two variables. The top five variables selected by pMDC are the same as those selected from both LASSO and forward selection with C_p, except that *pgg45* comes third for pMDC while it comes fifth for LASSO and forward selection. Note that *pgg45* is the percent of gleason scores 4 or 5, which is highly correlated with variable *gleason*, the gleason score. The latter is selected to enter the model by pdCor. Also from the scatter plot (Figure 1) we can see a strong non-linear relationship between *lpsa* and *pgg45*, which could contribute to the mean of the response in a non-linear way.

Table 4
Prediction Errors

Error Metric	pMDC	pdCor	Best Subsets	LASSO
MSE	0.544	0.512	0.529	0.544
MAE	0.518	0.532	0.545	0.518

After selecting the variables, we fit a generalized additive model to evaluate and compare different model fits using the package *mgcv* in R. The adjusted R^2s from GAM models are:

$$\begin{aligned} \text{pMDC:} &\quad 69.8\%; \\ \text{pdCor:} &\quad 69.3\%; \\ \text{best subsets:} &\quad 65.9\%; \\ \text{lasso:} &\quad 69.8\%; \end{aligned}$$

The above results suggest that the additional variable *pgg45* can contribute to the conditional mean of the response in a non-linear way, which may not be detected by LASSO under linear model assumptions. In general, the pdCor selects the variable that has the strongest dependence after controlling for the effect of the previously selected variables, while this overall dependence is different from the conditional mean dependence. The variable that has the largest pdCor may not be the one that contributes the most to the conditional mean of Y. Therefore, it may obscure the variable which has the largest additional contribution to the conditional mean but with less overall dependence. The ranking of the variables delivered by pMDC seems to make more intuitive sense.

In addition, we notice both pdCor and LASSO select *svi* as the third variable to enter the model, whereas pMDC selects *pgg45*. To demonstrate the importance of the selected order, we fit the GAM model again with only the first three variables selected by different methods and report the adjusted R^2s as follows:

$$\begin{aligned} \text{pMDC:} &\quad \text{lpsa} \sim \text{lcavol} + \text{lweight} + \text{pgg45} \quad (R^2 = 68.2\%) \\ \text{pdCor \& lasso :} &\quad \text{lpsa} \sim \text{lcavol} + \text{lweight} + \text{svi} \qquad (R^2 = 67\%) \end{aligned}$$

Again, partial MDC based variable selection seems to deliver a more sensible order than its pdCor based counterpart. As all the models under consideration are for the conditional mean, pMDC can be more efficient to select the variables that enter into a conditional mean model.

Furthermore, we use the fitted GAM models mentioned above to forecast for the 30 testing data and report the mean square error (MSE) and mean absolute error (MAE) for the predictions in Table 4. From the results we can see, pdCor has the least MSE while pMDC and LASSO have the least MAE.

In the sequel, we further look into the variables that contribute to the conditional quantile of the response variable. This is based on the quantile pMDC variable selection, which is the same as pMDC variable selection except for the fact that we first apply a transformation to Y, i.e., $U_i = \tau - \mathbf{1}(Y_i - \hat{Q}_\tau(Y) \leq 0)$, where $\hat{Q}_\tau(Y)$ is the τ-th sample quantile of the response and $\tau = (0.25, 0.5, 0.75)$;

see Remark 4.3. The models selected by pMDC and τ-th quantile pMDC are as follows:

$$\begin{aligned}
&\text{pMDC:} && \text{lpsa} \sim \text{lcavol} + \text{lweight} + \text{pgg45} + \text{svi} + \text{lbph};\\
&\text{0.25-pMDC:} && \text{lpsa} \sim \text{lcavol} + \text{lweight} + \text{pgg45} + \text{lbph} + \text{gleason};\\
&\text{0.5-pMDC:} && \text{lpsa} \sim \text{lweight} + \text{lcavol} + \text{pgg45} + \text{svi};\\
&\text{0.75-pMDC:} && \text{lpsa} \sim \text{lcavol} + \text{svi};
\end{aligned}$$

It can be seen that different variables are selected for the conditional mean model and conditional quantile models at different quantile levels. The conditional quantile of Y seems to depend on different sets of covariates at different quantile levels.

6. Discussion

In this paper, we propose an extension of the martingale difference correlation introduced in Shao and Zhang [10] to partial martingale difference correlation following the Hilbert-space approach of Székely and Rizzo [14]. In the course of this extension, we provide an equivalent expression for MDD at population and sample levels, which facilitates our definition of partial martingale difference correlation. Although the definition is not unique, the proposed partial MDC has a natural interpretation, admits a neat equivalent expression, and can be easily estimated using existing $\mathcal{U}$-centering-based unbiased estimates of MDD and distance covariance. Our numerical simulation shows that the test of zero partial MDD has comparable power to the test of zero partial distance covariance. A data illustration demonstrates that it can be more effective to select variables that enter into the conditional mean model using pMDC-based forward selection approach than the pdCor-based counterpart.

Appendix

A.1. Properties of $\mathcal{U}$-centering and the distance inner product

We first generalize Definition 2 of Székely & Rizzo [14] to *any* $n \times n$ real matrix $A = (a_{ij})$. That is, let $\widetilde{A}$ have

$$\widetilde{A}_{ij} = \begin{cases} a_{ij} - \frac{1}{n-2}\sum_{\ell=1}^n a_{i\ell} - \frac{1}{n-2}\sum_{k=1}^n a_{kj} + \frac{1}{(n-1)(n-2)}\sum_{k=1}^n \sum_{\ell=1}^n a_{k\ell} & i \neq j \\ 0 & i = j \end{cases}$$

For $n \times n$ real matrix A define

$$\dot{A} = A - \frac{1}{n-2}AJ - \frac{1}{n-2}JA + \frac{1}{(n-1)(n-2)}JAJ$$

where $J = \mathbf{1}\mathbf{1}^T$ is the $n \times n$ matrix of ones. It is straightforward to verify that $\widetilde{A}$ is just $\dot{A}$ with the diagonal replaced by zeros.

Similarly, generalize the *expression* for the distance inner product

$$(\widetilde{A} \cdot \widetilde{B}) \ := \ \frac{1}{n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} \widetilde{B}_{ij}$$

to apply to all matrices A and B (though, of course, it is not a true inner product on the space of all matrices).

Let $\operatorname{vec}(A)$ be the usual vectorization of matrix A formed by stacking its columns.

Lemma A.1.

1. *If A is an $n \times n$ real matrix* *with all zeros on its diagonal*, *and B is* *any* *$n \times n$ real matrix, then*

$$n(n-3)(\widetilde{A} \cdot \widetilde{B}) \ = \ \operatorname{vec}(A)^T \operatorname{vec}\left(\widetilde{\widetilde{B}}\right)$$

2. *If D is any diagonal matrix,*

$$\widetilde{\widetilde{D}} = 0$$

3. *For any $n \times 1$ vector $\boldsymbol{a}$, if*

$$H_{\boldsymbol{a}} \ = \ \boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{a}^T$$

then

$$\widetilde{\widetilde{H}}_{\boldsymbol{a}} \ = \ 0$$

To prove (i): Clearly $\operatorname{vec}(\dot{B})$ is a linear transformation of $\operatorname{vec}(B)$. Let $\boldsymbol{S}$ be the matrix of this transformation:

$$\operatorname{vec}(\dot{B}) \ = \ \boldsymbol{S} \operatorname{vec}(B)$$

An explicit form for $\boldsymbol{S}$ is

$$\boldsymbol{S} \ = \ I \otimes I - \frac{1}{n-2} J \otimes I - \frac{1}{n-2} I \otimes J + \frac{1}{(n-1)(n-2)} J \otimes J$$

from which it is clear that $\boldsymbol{S}$ is symmetric.

Also, let $\boldsymbol{F}$ denote the matrix of the (linear) operator that sets the diagonal of a matrix to zero. That is,

$$\operatorname{vec}(B_{-D}) \ = \ \boldsymbol{F} \operatorname{vec}(B)$$

where B_{-D} is B with its diagonal set equal to zero. It is obvious that $\boldsymbol{F}$ is a diagonal matrix with ones and zeros on the diagonal, hence symmetric and idempotent.

It follows that

$$\operatorname{vec}(\widetilde{B}) \ = \ \boldsymbol{F}\boldsymbol{S} \operatorname{vec}(B)$$

Then the distance inner product may be written

$$(\widetilde{A} \cdot \widetilde{B}) \ = \ \frac{1}{n(n-3)} \operatorname{vec}(\widetilde{A})^T \boldsymbol{F} \operatorname{vec}(\widetilde{B})$$

and thus

$$\begin{aligned} n(n-3)(\widetilde{A}\cdot\widetilde{B}) &= \operatorname{vec}(\widetilde{A})^T \boldsymbol{F}\operatorname{vec}(\widetilde{B}) \;=\; \operatorname{vec}(A)^T \boldsymbol{SFFFS}\operatorname{vec}(B) \\ &= \operatorname{vec}(A)^T \boldsymbol{SFS}\operatorname{vec}(B) \;=\; \operatorname{vec}(A)^T \boldsymbol{FSFS}\operatorname{vec}(B) \\ &= \operatorname{vec}(A)^T \operatorname{vec}\left(\widetilde{\widetilde{B}}\right) \end{aligned}$$

because A already has a zero diagonal (so $\boldsymbol{F}\operatorname{vec}(A) = \operatorname{vec}(A)$), and using the symmetry of $\boldsymbol{S}$ and $\boldsymbol{F}$.

To prove (ii): By linearity, it suffices to prove the proposition for the matrices

$$E_i = \boldsymbol{e}_i\boldsymbol{e}_i^T$$

which are all zero except for a one in the ith diagonal position. Now,

$$\begin{aligned} \dot{E}_i &= E_i - \frac{1}{n-2}E_iJ - \frac{1}{n-2}JE_i + \frac{1}{(n-1)(n-2)}JE_iJ \\ &= E_i - \frac{1}{n-2}(\boldsymbol{e}_i\boldsymbol{1}^T + \boldsymbol{1}\boldsymbol{e}_i^T) + \frac{1}{(n-1)(n-2)}J \end{aligned}$$

Then (removing the diagonal)

$$\widetilde{E}_i \;=\; 0 - \frac{1}{n-2}(\boldsymbol{e}_i\boldsymbol{1}^T + \boldsymbol{1}\boldsymbol{e}_i^T - 2\boldsymbol{e}_i\boldsymbol{e}_i^T) + \frac{1}{(n-1)(n-2)}(J-I) \tag{A.1}$$

Now

$$\widetilde{E}_i\boldsymbol{1} \;=\; -\frac{1}{n-2}(n\boldsymbol{e}_i + \boldsymbol{1} - 2\boldsymbol{e}_i) + \frac{1}{(n-1)(n-2)}(n\boldsymbol{1} - \boldsymbol{1}) \;=\; -\boldsymbol{e}_i$$

and similarly

$$\boldsymbol{1}^T\widetilde{E}_i \;=\; -\boldsymbol{e}_i^T$$

It follows that

$$\begin{aligned} \dot{\widetilde{E}}_i &= \widetilde{E}_i - \frac{1}{n-2}(-\boldsymbol{e}_i)\boldsymbol{1}^T - \frac{1}{n-2}\boldsymbol{1}(-\boldsymbol{e}_i^T) + \frac{1}{(n-1)(n-2)}\boldsymbol{1}(-1)\boldsymbol{1}^T \\ &= \widetilde{E}_i + \frac{1}{n-2}(\boldsymbol{e}_i\boldsymbol{1}^T + \boldsymbol{1}\boldsymbol{e}_i^T) - \frac{1}{(n-1)(n-2)}J \end{aligned}$$

Using (A.1), this becomes

$$\dot{\widetilde{E}}_i \;=\; \frac{2}{n-2}\boldsymbol{e}_i\boldsymbol{e}_i^T - \frac{1}{(n-1)(n-2)}I$$

and this is clearly a diagonal matrix, so (removing the diagonal)

$$\widetilde{\widetilde{E}}_i \;=\; 0$$

and the proof is complete.

To prove (iii): Since

$$H_{\boldsymbol{a}} \;=\; \sum_{i=1}^{n} a_i(\boldsymbol{e}_i\boldsymbol{1}^T + \boldsymbol{1}\boldsymbol{e}_i^T)$$

it suffices by linearity to prove the proposition for the matrices

$$H_i \;=\; \boldsymbol{e}_i \mathbf{1}^T + \mathbf{1}\boldsymbol{e}_i^T$$

For such matrices,

$$H_i J \;=\; n\boldsymbol{e}_i \mathbf{1}^T + J \qquad \text{and} \qquad J H_i \;=\; J + n\mathbf{1}\boldsymbol{e}_i^T$$

and so

$$\begin{aligned}\dot{H}_i &= H_i - \frac{1}{n-2} H_i J - \frac{1}{n-2} J H_i + \frac{1}{(n-1)(n-2)} J H_i J \\ &= H_i - \frac{1}{n-2}(nH_i + 2J) + \frac{1}{(n-1)(n-2)} 2nJ \\ &= -\frac{2}{n-2} H_i + \frac{2}{(n-1)(n-2)} J\end{aligned}$$

and setting the diagonal to zero gives

$$\widetilde{H}_i \;=\; -\frac{2}{n-2}(H_i - 2\boldsymbol{e}_i\boldsymbol{e}_i^T) + \frac{2}{(n-1)(n-2)}(J - I) \;=\; 2\widetilde{E}_i$$

according to equation (A.1). It then follows that

$$\widetilde{\widetilde{H}}_i \;=\; 2\widetilde{\widetilde{E}}_i \;=\; 0$$

according to (ii), and the proof is complete. □

A.2. Proofs of propositions

Proof of Proposition 3.2. For any fixed $s \in \mathbf{R}^q$, let

$$\begin{aligned}h(t) &= |\phi_{U,V}(s,t) - \phi_U(s)\phi_V(t)|^2 \\ &= (\phi_{U,V}(s,t) - \phi_U(s)\phi_V(t))\,\overline{(\phi_{U,V}(s,t) - \phi_U(s)\phi_V(t))}\end{aligned}$$

which is at least twice differentiable under the assumption that V has finite second moments. Then

$$\begin{aligned}\nabla_t h(t) &= (\nabla_t \phi_{U,V}(s,t) - \phi_U(s)\nabla_t\phi_V(t))\,\overline{(\phi_{U,V}(s,t) - \phi_U(s)\phi_V(t))} \\ &\quad + (\phi_{U,V}(s,t) - \phi_U(s)\phi_V(t))\,\overline{(\nabla_t\phi_{U,V}(s,t) - \phi_U(s)\nabla_t\phi_V(t))}\end{aligned}$$

and

$$\begin{aligned}H_t(h)(t) &= \nabla_t^2 h(t) \\ &= (\nabla_t^2 \phi_{U,V}(s,t) - \phi_U(s)\nabla_t^2\phi_V(t))\,\overline{(\phi_{U,V}(s,t) - \phi_U(s)\phi_V(t))} \\ &\quad + \overline{(\nabla_t\phi_{U,V}(s,t) - \phi_U(s)\nabla_t\phi_V(t))}\,(\nabla_t\phi_{U,V}(s,t) - \phi_U(s)\nabla_t\phi_V(t))^T\end{aligned}$$

$$
\begin{aligned}
&+ \left(\nabla_t \phi_{U,V}(s,t) - \phi_U(s)\nabla_t \phi_V(t)\right) \overline{\left(\nabla_t \phi_{U,V}(s,t) - \phi_U(s)\nabla_t \phi_V(t)\right)^T} \\
&+ \left(\phi_{U,V}(s,t) - \phi_U(s)\phi_V(t)\right) \overline{\left(\nabla_t^2 \phi_{U,V}(s,t) - \phi_U(s)\nabla_t^2 \phi_V(t)\right)}
\end{aligned}
$$

Since $\phi_{U,V}(s,0) - \phi_U(s)\phi_V(0) \;=\; E(e^{i\langle s,U\rangle}) - \phi_U(s)\cdot 1 \;=\; 0$, we have

$$
H_t(h)(0) \;=\; \overline{z}\, z^T + z\, \overline{z}^T \;=\; \overline{z\, z^H} + z\, z^H
$$

where z^H is the conjugate transpose of z, and

$$
\begin{aligned}
z \;&=\; \nabla_t \phi_{U,V}(s,0) - \phi_U(s)\nabla_t \phi_V(0) \;=\; E(e^{i\langle s,U\rangle}\, iV) - \phi_U(s)E(iV) \\
&=\; i\big(E(Ve^{i\langle s,U\rangle}) - E(V)\phi_U(s)\big)
\end{aligned}
$$

since the derivatives may be taken under the expectation. Thus

$$
\begin{aligned}
\Delta_t h(0) \;&=\; \mathrm{trace}(H_t(h)(0)) \;=\; \overline{\mathrm{trace}(z\, z^H)} \;+\; \mathrm{trace}(z\, z^H) \\
&=\; \overline{z^H z} \;+\; z^H z \;=\; 2 z^H z \\
&=\; 2\,|E(Ve^{i\langle s,U\rangle}) \;-\; E(V)\, E(e^{i\langle s,U\rangle})|_r^2
\end{aligned}
$$

and the result follows. □

Proof of Proposition 3.3.

$$
\begin{aligned}
&|\phi^n_{U,V}(s,t) - \phi^n_U(s)\phi^n_V(t)|^2 \\
&\qquad =\; (\phi^n_{U,V}(s,t) - \phi^n_U(s)\phi^n_V(t))\overline{(\phi^n_{U,V}(s,t) - \phi^n_U(s)\phi^n_V(t))} \\
&\qquad =\; \phi^n_{U,V}(s,t)\overline{\phi^n_{U,V}(s,t)} \;+\; \phi^n_U(s)\phi^n_V(t)\overline{\phi^n_U(s)\phi^n_V(t)} \\
&\qquad\qquad -\; \phi^n_{U,V}(s,t)\overline{\phi^n_U(s)\phi^n_V(t)} \;-\; \overline{\phi^n_{U,V}(s,t)}\phi^n_U(s)\phi^n_V(t)
\end{aligned}
\tag{A.2}
$$

Define

$$
\alpha_{kl} \;=\; 1 - e^{i\langle s,\, U_k - U_l\rangle} \qquad\qquad \beta_{kl} \;=\; 1 - e^{i\langle t,\, V_k - V_l\rangle}
$$

Then, in this notation, the first term of (A.2) becomes

$$
\begin{aligned}
\phi^n_{U,V}(s,t)\overline{\phi^n_{U,V}(s,t)} \;&=\; \frac{1}{n^2}\sum_{k=1}^n\sum_{l=1}^n (1-\alpha_{kl})(1-\beta_{kl}) \\
&=\; 1 \;-\; \frac{\alpha_{..}}{n^2} \;-\; \frac{\beta_{..}}{n^2} \;+\; \frac{1}{n^2}\sum_{k=1}^n\sum_{l=1}^n \alpha_{kl}\beta_{kl}
\end{aligned}
$$

(where a dotted subscript represents summation over the associated index, as usual). The second term of (A.2) becomes

$$
\begin{aligned}
\phi^n_U(s)\phi^n_V(t)\overline{\phi^n_U(s)\phi^n_V(t)} \;&=\; \frac{1}{n^2}\sum_{k=1}^n\sum_{l=1}^n (1-\alpha_{kl}) \;\cdot\; \frac{1}{n^2}\sum_{k=1}^n\sum_{l=1}^n (1-\beta_{kl}) \\
&=\; \left(1 - \frac{\alpha_{..}}{n^2}\right)\left(1 - \frac{\beta_{..}}{n^2}\right) \;=\; 1 \;-\; \frac{\alpha_{..}}{n^2} \;-\; \frac{\beta_{..}}{n^2} \;+\; \frac{\alpha_{..}}{n^2}\frac{\beta_{..}}{n^2}
\end{aligned}
$$

For the third term of (A.2),

$$\begin{aligned}\phi^n_{U,V}(s,t)\overline{\phi^n_U(s)\phi^n_V(t)} &= \frac{1}{n^3}\sum_{k=1}^n\sum_{l=1}^n\sum_{m=1}^n(1-\alpha_{kl})(1-\beta_{km}) \\ &= 1 - \frac{\alpha_{\cdot\cdot}}{n^2} - \frac{\beta_{\cdot\cdot}}{n^2} + \frac{1}{n^3}\sum_{k=1}^n\sum_{l=1}^n\sum_{m=1}^n\alpha_{kl}\beta_{km}\end{aligned}$$

For the fourth term, because the conjugates of α_{kl} and β_{km} are

$$\overline{\alpha_{kl}} = \alpha_{lk} \qquad \text{and} \qquad \overline{\beta_{kl}} = \beta_{lk}$$

we have

$$\begin{aligned}\overline{\phi^n_{U,V}(s,t)}\phi^n_U(s)\phi^n_V(t) &= \overline{\phi^n_{U,V}(s,t)\overline{\phi^n_U(s)\phi^n_V(t)}} \\ &= 1 - \frac{\alpha_{\cdot\cdot}}{n^2} - \frac{\beta_{\cdot\cdot}}{n^2} + \frac{1}{n^3}\sum_{k=1}^n\sum_{l=1}^n\sum_{m=1}^n\alpha_{lk}\beta_{mk}\end{aligned}$$

Substituting all of these into (A.2) and canceling terms gives

$$\begin{aligned}&|\phi^n_{U,V}(s,t) - \phi^n_U(s)\phi^n_V(t)|^2 \\ &\quad = \frac{1}{n^2}\sum_{k=1}^n\sum_{l=1}^n\alpha_{kl}\beta_{kl} + \frac{\alpha_{\cdot\cdot}}{n^2}\frac{\beta_{\cdot\cdot}}{n^2} - \frac{1}{n^3}\sum_{k=1}^n\sum_{l=1}^n\sum_{m=1}^n(\alpha_{kl}\beta_{km}+\alpha_{lk}\beta_{mk})\end{aligned} \tag{A.3}$$

Now, from the fundamental Lemma of Székely, Rizzo, and Bakirov [13],

$$\int_{\mathbf{R}^q}\frac{\alpha_{kl}}{c_q|s|_q^{1+q}}\,ds = |U_k - U_l|_q = a_{kl}$$

Also,

$$\begin{aligned}\frac{1}{2}\Delta_t\beta_{kl}\Big|_{t=0} &= -\frac{1}{2}\sum_{i=1}^r\big(i(V_{ki}-V_{li})\big)^2 e^{i\langle t,\,V_k-V_l\rangle}\Big|_{t=0} \\ &= \frac{1}{2}\sum_{i=1}^r(V_{ki}-V_{li})^2 = \frac{1}{2}|V_k-V_l|_r^2 = b^\circ_{kl}\end{aligned}$$

Applying both of these operations (the integral and the Laplacian) to (A.3), and using their linearity, gives

$$\begin{aligned}&\int_{\mathbf{R}^q}\left(\frac{1}{2}\,\Delta_t\,|\phi^n_{U,V}(s,t)-\phi^n_U(s)\phi^n_V(t)|^2\Big|_{t=0}\right)\frac{1}{c_q\,|s|_q^{1+q}}\,ds \\ &\quad = \frac{1}{n^2}\sum_{k=1}^n\sum_{l=1}^n a_{kl}b^\circ_{kl} + \frac{a_{\cdot\cdot}}{n^2}\frac{b^\circ_{\cdot\cdot}}{n^2} - \frac{1}{n^3}\sum_{k=1}^n\sum_{l=1}^n\sum_{m=1}^n(a_{kl}b^\circ_{km}+a_{lk}b^\circ_{mk}) \\ &\quad = \frac{1}{n^2}\sum_{k=1}^n\sum_{l=1}^n a_{kl}b^\circ_{kl} + \frac{a_{\cdot\cdot}}{n^2}\frac{b^\circ_{\cdot\cdot}}{n^2} - 2\,\frac{1}{n^3}\sum_{k=1}^n\sum_{l=1}^n\sum_{m=1}^n a_{kl}b^\circ_{km} \\ &\quad = S_1 + S_2 - 2S_3\end{aligned}$$

using the fact that $a_{lk} = a_{kl}$ and $b^\circ_{lk} = b^\circ_{kl}$. The rest follows from the identity

$$\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} A_{kl}B^\circ_{kl} \;=\; S_1 \,+\, S_2 \,-\, 2S_3$$

which follows in precisely the same way as in Theorem 1 of Székely, Rizzo, and Bakirov [13].

Incidentally, the proof above could be adapted as an alternative to the proof of Theorem 1 of Székely, Rizzo, and Bakirov [13]. □

Proof of Proposition 3.4. The proof follows Appendix A.1 of Székely and Rizzo [14]. Letting (X, Y), (X', Y') and (X'', Y'') be iid from the population distribution, define

$$\begin{aligned}
\alpha &:= E(a_{kl}) \;=\; E(|X - X'|_p) \qquad \beta \;:=\; E(b^\circ_{kl}) \;=\; E\big(\tfrac{1}{2}|Y - Y'|_q^2\big) \qquad k \neq l\\
\delta &:= E(a_{kl}b^\circ_{kj}) \;=\; E\big(|X - X'|_p\, \tfrac{1}{2}|Y - Y''|_q^2\big) \qquad j, k, l \text{ distinct}\\
\gamma &:= E(a_{kl}b^\circ_{kl}) \;=\; E\big(|X - X'|_p\, \tfrac{1}{2}|Y - Y'|_q^2\big) \qquad k \neq l
\end{aligned}$$

Then, according to (3.4),

$$MDD(Y|X)^2 \;=\; \gamma + \alpha\beta - 2\delta$$

By the very same derivation as in Appendix A.1 of Székely and Rizzo [14], it follows that

$$E\big[(\widetilde{A} \cdot \widetilde{B}^\circ)\big] \;=\; \gamma + \alpha\beta - 2\delta$$

(Indeed, their derivation uses only the symmetry of the matrices $A = (a_{kl})$ and $B = (b_{kl})$ and the fact that they have zero diagonals, so B may be replaced with $B^\circ = (b^\circ_{kl})$.) The result follows immediately. □

Proof of Proposition 3.5. Since

$$(y_i - y_j)^T(y_i - y_j) \;=\; |y_i|_q^2 + |y_j|_q^2 - 2y_i^T y_j$$

we can write

$$B^\circ \;=\; H_{\boldsymbol{a}} + B^*$$

where

$$H_{\boldsymbol{a}} \;=\; ((H_{\boldsymbol{a}})_{ij}), \qquad\qquad (H_{\boldsymbol{a}})_{ij} \;=\; \frac{1}{2}|y_i|_q^2 + \frac{1}{2}|y_j|_q^2$$

Note that

$$H_{\boldsymbol{a}} \;=\; \boldsymbol{a}\boldsymbol{1}^T + \boldsymbol{1}\boldsymbol{a}^T$$

where $\boldsymbol{a}$ is the $n \times 1$ vector whose ith element is $\frac{1}{2}|y_i|_q^2$.

By the linearity of $\mathcal{U}$-centering, and from Lemma A.1, we have

$$\widetilde{\widetilde{B^\circ}} \;=\; \widetilde{\widetilde{H}}_{\boldsymbol{a}} + \widetilde{\widetilde{B^*}} \;=\; \widetilde{\widetilde{B^*}}$$

Then, using Lemma A.1,

$$\begin{aligned} n(n-3)(\widetilde{A}\cdot\widetilde{B}^{*}) &= \operatorname{vec}(A)^{T}\operatorname{vec}\left(\widetilde{\widetilde{\widetilde{B}^{*}}}\right) \\ &= \operatorname{vec}(A)^{T}\operatorname{vec}\left(\widetilde{\widetilde{\widetilde{B}^{\circ}}}\right) = n(n-3)(\widetilde{A}\cdot\widetilde{B}^{\circ}) \end{aligned}$$

and the proof is complete. □

Proof of Proposition 4.1. We consider the following cases:

- Case 1: If Z is constant a.s. then $\widetilde{P}_{Z^\perp}(Y) = B_Y^\circ - \beta C_Z = B_Y^\circ$.

$$pMDC(Y|X;Z) = \frac{(B_Y^\circ \cdot D_W)}{((B_Y^\circ \cdot B_Y^\circ)\cdot \mathcal{V}^2(W,W))^{1/2}} = MDC(Y|W)^2$$

  and (4.1) $= MDC(Y|W)^2$ as well, since $MDC(Y|Z)^2 = 0$ by definition.
- Case 2: If Y is constant almost surely and Z is not almost surely constant, then $MDC(Y|W)^2 = 0$ and $MDC(Y|Z)^2 = 0$ by definition, so (4.1) $= 0$. We also have $\beta = \frac{MDD(Y|Z)^2}{\mathcal{V}^2(Z,Z)} = 0$, $\widetilde{P}_{Z^\perp}(Y) = B_Y^\circ$. So $pMDC(Y|X;Z) = 0$ by Definition 4.1.
- Case 3: If Z and Y are not almost surely constant, but $|\widetilde{P}_{Z^\perp}(Y)| = 0$.

$$\begin{aligned} (\widetilde{P}_{Z^\perp}(Y)\cdot\widetilde{P}_{Z^\perp}(Y)) &= E(B_Y^\circ B_Y^\circ) + \frac{MDD(Y|Z)^4}{\mathcal{V}^4(Z,Z)}E(C_Z C_Z) \\ &\quad - 2\frac{MDD(Y|Z)^4}{\mathcal{V}^2(Z,Z)} \\ &= E(B_Y^\circ B_Y^\circ) - \frac{MDD(Y|Z)^4}{\mathcal{V}^2(Z,Z)} \end{aligned}$$

  If $|\widetilde{P}_{Z^\perp}(Y)| = 0$, $MDD(Y|Z)^2 = \sqrt{E(B_Y^\circ B_Y^\circ)\mathcal{V}^2(Z,Z)}$, $MDC(Y|Z)^2 = 1$. Then both (4.1) and $pMDC(Y|X;Z)$ are zero.
- Case 4: If Z and Y are not almost surely constants, and $|\widetilde{P}_{Z^\perp}(Y)| > 0$. Then

$$E(\widetilde{P}_{Z^\perp}(Y)D_W) = E(B_Y^\circ D_W) - \frac{MDD(Y|Z)^2}{\mathcal{V}^2(Z,Z)}E(C_Z D_W)$$

  And also

$$\begin{aligned} (\widetilde{P}_{Z^\perp}(Y)\cdot\widetilde{P}_{Z^\perp}(Y)) &= E(B_Y^\circ B_Y^\circ) - \frac{MDD(Y|Z)^4}{\mathcal{V}^2(Z,Z)} \\ &= E(B_Y^\circ B_Y^\circ)(1 - \frac{MDD(Y|Z)^4}{E(B_Y^\circ B_Y^\circ)\mathcal{V}^2(Z,Z)}) \\ &= E(B_Y^\circ B_Y^\circ)(1 - MDC(Y|Z)^4) \end{aligned}$$

  Hence we have

$$pMDC(Y|X;Z) = \frac{(\widetilde{P}_{Z^\perp}(Y)\cdot D_W)}{((\widetilde{P}_{Z^\perp}(Y)\cdot\widetilde{P}_{Z^\perp}(Y))\cdot\mathcal{V}^2(W,W))^{1/2}}$$

$$
\begin{aligned}
&= \frac{E(B_Y^\circ D_W) - \frac{MDD(Y|Z)^2}{\mathcal{V}^2(Z,Z)} E(C_Z D_W)}{\sqrt{E(B_Y^\circ B_Y^\circ)\mathcal{V}^2(W,W)}\sqrt{1-MDC(Y|Z)^4}} \\
&= \frac{\frac{E(B_Y^\circ D_W)}{\sqrt{E(B_Y^\circ B_Y^\circ)\mathcal{V}^2(W,W)}}}{\sqrt{1-MDC(Y|Z)^4}} \\
&\quad - \frac{\frac{MDD(Y|Z)^2}{\sqrt{E(B_Y^\circ B_Y^\circ)\mathcal{V}^2(Z,Z)}} \frac{E(C_Z D_W)}{\sqrt{\mathcal{V}^2(Z,Z)\mathcal{V}^2(W,W)}}}{\sqrt{1-MDC(Y|Z)^4}} \\
&= \frac{MDC(Y|W)^2 - MDC(Y|Z)^2 R^2(Z,W)}{\sqrt{1-MDC(Y|Z)^4}}
\end{aligned}
$$

Thus, in all cases Definition 4.1 and (4.1) coincide. □

Proof of Proposition 4.2. We consider the following cases:

- Case 1: If $(z_i)_{i=1}^n$ are all equal, then $C_{ij} = |z_i - z_j|_r = 0$. Therefore $\beta_n = \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})} = 0$ and $\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}\widetilde{C} = \widetilde{B}^\circ$.

  $$pMDC_n(Y|X;Z) = \frac{(\widetilde{B}^\circ \cdot \widetilde{D})}{|\widetilde{B}^\circ||\widetilde{D}|} = \widetilde{MDC}_n(Y|W)^2$$

  Since $|\widetilde{C}| = 0$, we also have $\widetilde{MDC}_n(Y|Z)^2 = 0$ by definition. So (4.2) = $\widetilde{MDC}_n(Y|W)^2$
- Case 2: If $(y_i)_{i=1}^n$ are all equal and $(z_i)_{i=1}^n$ are not, then we have $B_{ij}^\circ = \frac{1}{2}|y_i - y_j|_q^2 = 0$ and $|\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}\widetilde{C}| = 0$. So $pMDC_n(Y|X;Z) = 0$ by definition. Also since $|\widetilde{B}^\circ| = 0$, $\widetilde{MDC}_n(Y|W)^2 = 0$ and $\widetilde{MDC}_n(Y|Z)^2 = 0$, so (4.2) $= 0$ by definition.
- Case 3: If $(z_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are not all equal, but $|\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}\widetilde{C}| = 0$.

  $$
  \begin{aligned}
  (\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}\widetilde{C} \cdot \widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}\widetilde{C}) &= (\widetilde{B}^\circ \cdot \widetilde{B}^\circ) + \frac{(\widetilde{B}^\circ \cdot \widetilde{C})^2}{(\widetilde{C}\cdot\widetilde{C})} - 2\frac{(\widetilde{B}^\circ \cdot \widetilde{C})^2}{(\widetilde{C}\cdot\widetilde{C})} \\
  &= (\widetilde{B}^\circ \cdot \widetilde{B}^\circ) - \frac{\widetilde{MDD}_n(Y|Z)^4}{(\widetilde{C}\cdot\widetilde{C})}
  \end{aligned}
  $$

  If $|\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}\widetilde{C}| = 0$, then $\widetilde{MDD}_n(Y|Z)^2 = |\widetilde{B}^\circ||\widetilde{C}|$, which implies that $\widetilde{MDC}_n(Y|Z)^2 = 1$. Then both (4.2) and $pMDC_n(Y|X;Z)$ are zero.
- Case 4: If $(z_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are not constants, and $|\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}\widetilde{C}| > 0$. Then

  $$pMDD_n(Y|X;Z) \quad = \quad (\widetilde{B}^\circ \cdot \widetilde{D}) - \frac{(\widetilde{B}^\circ \cdot \widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}(\widetilde{C}\cdot\widetilde{D})$$

$$= \widetilde{MDD}_n(Y|W)^2 - \frac{\widetilde{MDD}_n(Y|Z)^2}{(\widetilde{C}\cdot\widetilde{C})}(\widetilde{C}\cdot\widetilde{D})$$

And also

$$\begin{aligned}\left(\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ\cdot\widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}\widetilde{C}\cdot\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ\cdot\widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}\widetilde{C}\right) &= (\widetilde{B}^\circ\cdot\widetilde{B}^\circ) - \frac{\widetilde{MDD}_n(Y|Z)^4}{(\widetilde{C}\cdot\widetilde{C})}\\ &= (\widetilde{B}^\circ\cdot\widetilde{B}^\circ)\left[1 - \frac{\widetilde{MDD}_n(Y|Z)^4}{(\widetilde{B}^\circ\cdot\widetilde{B}^\circ)(\widetilde{C}\cdot\widetilde{C})}\right]\\ &= (\widetilde{B}^\circ\cdot\widetilde{B}^\circ)(1 - \widetilde{MDC}_n(Y|Z)^4)\end{aligned}$$

Hence we have

$$\begin{aligned} pMDC_n(Y|X;Z) &= \frac{pMDD_n(Y|X;Z)}{|\widetilde{B}^\circ - \frac{(\widetilde{B}^\circ\cdot\widetilde{C})}{(\widetilde{C}\cdot\widetilde{C})}\widetilde{C}||\tilde{D}|}\\ &= \frac{\widetilde{MDD}_n(Y|W)^2 - \frac{\widetilde{MDD}_n(Y|Z)^2}{(\widetilde{C}\cdot\widetilde{C})}(\widetilde{C}\cdot\widetilde{D})}{|\widetilde{B}^\circ||\widetilde{D}|\sqrt{1-\widetilde{MDC}_n(Y|Z)^4}}\\ &= \frac{\frac{\widetilde{MDD}_n(Y|W)^2}{|\widetilde{B}^\circ||\widetilde{D}|} - \frac{\widetilde{MDD}_n(Y|Z)^2}{|\widetilde{B}^\circ||\widetilde{C}|}\frac{(\widetilde{C}\cdot\widetilde{D})}{|\widetilde{C}||\widetilde{D}|}}{\sqrt{1-\widetilde{MDC}_n(Y|Z)^4}}\\ &= \frac{\widetilde{MDC}_n(Y|W)^2 - \widetilde{MDC}_n(Y|Z)^2\widetilde{R}_n^2(Z,W)}{\sqrt{1-\widetilde{MDC}_n(Y|Z)^4}}\end{aligned}$$

Thus, in all cases Definition 4.2 and (4.2) coincide. □

References

[1] Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, **30**, 455–474. MR1902895

[2] Dueck, J., Edelmann, D., Gneiting, T., and Richards, D. (2014). The affinely invariant distance correlation. *Bernoulli*, **20**, 2305–2330. MR3263106

[3] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of Statistical Learning*, 2nd ed, Springer, New York. MR2722294

[4] Kong, J., Klein, B. E. K., Klein, R., Lee, K., and Wahba, G. (2012). Using distance correlation and SS-AVOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proceeding of the National Academy of Sciences*, **109(50)**, 20352–20357.

[5] Legendre, P. (2000). Comparison of permutation methods for the partial correlation and partial Mantel tests. *J. Statist. Compu. and Simu.*, **67**, 37–73. MR1815171

[6] Li, G., Li, Y., and Tsai, C.-L. (2014). Quantile correlations and quantile autoregressive modeling. *Journal of the American Statistical Association*, to appear. MR3338500

[7] Li, R., Zhong, W., and Zhu, L. (2012). Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association*, **107**, 1129–1139. MR3010900

[8] Lyons, R. (2013). Distance covariance in metric spaces. *Annals of Probability*, **41**, 3284–3305. MR3127883

[9] Sejdinovick, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, **41**, 2263–2291. MR3127866

[10] Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high dimensional variable screening. *Journal of the American Statistical Association*, **109**, 1302–1318. MR3265698

[11] Sheng, W. and Yin, X. (2013). Direction estimation in single-index models via distance covariance. *Journal of Multivariate Analysis*, **122**, 148–161. MR3189314

[12] Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, **3**, 1236–1265. MR2752127

[13] Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing independence by correlation of distances. *Annals of Statistics*, **35(6)**, 2769–2794. MR2382665

[14] Székely, G. J. and Rizzo, M. L. (2014). Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, to appear. MR3269983

[15] Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, to appear.