Scientific Research Publishing

Open Journal of Statistics, 2025, 15(3), 294-311
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Kernel-Based Partial Conditional Mean Dependence

Zhentao Tian*, Zhongzhan Zhang

School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing, China
Email: *statisticandmath@163.com

How to cite this paper: Tian, Z.T. and Zhang, Z.Z. (2025) Kernel-Based Partial Conditional Mean Dependence. *Open Journal of Statistics*, **15**, 294-311.
https://doi.org/10.4236/ojs.2025.153015

Received: May 13, 2025
Accepted: June 20, 2025
Published: June 23, 2025

Abstract

We introduce the Kernel-based Partial Conditional Mean Dependence, a scalar-valued measure of conditional mean dependence of Y given X, while adjusting for the nonlinear dependence on Z. Here X, Y and Z are random elements from arbitrary separable Hilbert spaces. This measure extends the Kernel-based Conditional Mean Dependence. As the estimator of the measure is developed, the concentration property of the estimator is proved. Numerical results demonstrate the effectiveness of the new dependence measure in the context of dependence testing, highlighting their advantages in capturing nonlinear partial conditional mean dependencies.

Keywords

Partial Conditional Mean Dependence, Hilbert Space, High Dimension, Test of Independence

1. Introduction

Before constructing a regression model, it is important to determine whether the covariate X has an effect on the response Y. As pointed out by [1], in most cases, we are more concerned with the conditional mean of the response. Thus, the conditional mean dependence has received attention, which measures the departure of $E[Y \mid X]$ from $E[Y]$. When X has no effect on the conditional mean of Y, i.e., $E[Y \mid X] = E[Y]$, X should not be included in a conditional mean regression model. In practice, based on historical analysis or domain knowledge, some covariates related to response will be known. Our aim is to determine whether X has contribution on the conditional mean of Y after controlling the affect from the known variable Z.

Regarding partial dependence, work has been increasing recently. An intuitive approach to measure partial conditional mean dependence is

$$E\left[E\left(Y \mid X, Z\right)-E\left(Y \mid Z\right)\right]^2. \tag{1}$$

Based on a plug-in estimator of equation (1), [2] developed a partial conditional mean independence test. However, as pointed out by [3], under the null hypothesis, i.e., when quantity (1) equals zero, the test statistic has a degenerate distribution. To deal with the degenerate limit distribution, [3] developed a significance test based on the black-box learner, [4] proposed a general framework to evaluate feature importance, and [5] considered measuring the partial dependence based on the decomposition formula of the conditional variance. These methods combine machine learning with sample splitting. Therefore, to a certain extent, they suffer from the loss of power caused by sample splitting. Another issue is that they only consider scalar responses and cannot handle vector or functional responses. In the field of vector or functional data analysis, conditional mean regression is an important analytical tool (see [6] for a regression model with vector response, see [7]-[9]) for regression models with function response among others), and it is necessary to consider the partial conditional mean dependence for vector or functional response.

To our knowledge, among these tools for partial conditional mean dependence, the Partial Martingale Difference Divergence (pMDD), as introduced in [10], is currently the only one applicable to response variables in Hilbert space. pMDD is 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. It extends the martingale difference divergence (MDD) introduced in [11]. However, as shown in [12] [13], the performance of MDD suffers from the curse of dimensionality. Let $\left(X', Y'\right)$ be an independent copy of $\left(X, Y\right)$, and let $\mathrm{MDD}\left(Y \mid X\right)$ be the martingale difference divergence of Y given X. When $X=\left(X_1, \cdots, X_p\right) \in R^p$ and $Y \in R$, [12] shows that

$$\mathrm{MDD}\left(Y \mid X\right) \approx \frac{1}{\sqrt{\tau}} \sum_{i=1}^{p} \mathrm{cov}\left(X_i, Y\right),$$

where $\tau = E\left\|X - X'\right\|^2$, and $\mathrm{cov}\left(X_i, Y\right)$ is the covariance of X_i and Y. Since the covariance only captures the linear dependence, the martingale difference divergence may have less power when it is employed to detect nonlinear relationships, especially in the cases of high dimensions. pMDD, as an extension of MDD, will suffer from the curse of dimensionality for the same reason. This phenomenon can be found in the numerical results in Section 4.

In this paper, we introduce a new tool to measure the partial conditional mean dependence for vector or functional responses. The numerical experiments in [13] demonstrated the advantages of kernel-based conditional mean dependence over MDD in identifying nonlinear relationships and handling high-dimensional variables. This prompts us to develop a tool based on kernel methods for measuring partial conditional mean dependence. Our development follows that in [10], so we name our tool as Kernel-based Partial Conditional Mean Dependence. Simulation results show that Kernel-based Partial Conditional Mean Dependence has an

advantage over Partial Martingale Difference Divergence in identifying the nonlinear dependence of Y's conditional mean on X after controlling for Z.

The rest of the paper is organized as follows. In Section 2, we review the kernel-based conditional mean dependence measure. In Section 3, we explore the procedure of constructing Kernel-based Partial Conditional Mean Dependence, and give its sample analogy. A group of finite sample simulation studies is carried out in Section 4. In Section 5, some discussions are included. All technical proofs are presented in the **Appendix**.

2. Kernel-Based Conditional Mean Dependence

Before formally introducing the Kernel-based Partial Conditional Mean Dependence, it is necessary to review a tool for measuring conditional mean Dependence—Kernel-based Conditional Mean Dependence.

As proposed by [13], the kernel-based conditional mean dependence (KCMD) is defined as

$$\mathrm{KCMD}(Y \mid Z) = E\langle Y - EY, Y' - EY\rangle K_Z(Z, Z'),$$

where $\mathcal{Y}$, $\mathcal{Z}$ are separable Hilbert spaces, Y, Z are random elements valued in $\mathcal{Y}$, $\mathcal{Z}$, respectively. (Z', Y') is an independent copy of (Z, Y), $K_Z(\cdot,\cdot)$ is a characteristic kernel (For details on characteristic kernels, please refer to [14] [15]) defined on $\mathcal{Z}\times\mathcal{Z}$. We specifically point out that in this article, the kernel function $K_Z(\cdot,\cdot)$ used in KCMD is not fixed. Its subscript Z indicates that the form and domain of the kernel function depend on Z and the space in which Z is located. Through the paper, $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ represent inner products and norms respectively. Kernel-based conditional mean dependence (KCMD) is intended to measure departure from the relationship

$$E(Y \mid Z) = E(Y) \text{ almost surely}$$

for $Y \in \mathcal{Y}$ and $Z \in \mathcal{Z}$. Lemma 1 below summarizes the fundamental properties of KCMD.

Lemma 1. Suppose that $K_Z(\cdot,\cdot)$ is a positive definite and bounded characteristic kernel. Then $\mathrm{KCMD}(Y \mid Z)$ is well defined, and

a) $\mathrm{KCMD}(Y \mid Z) \geq 0$;

b) $\mathrm{KCMD}(Y \mid Z) = 0$ if and only if $E(Y \mid Z) = E(Y)$, almost surely.

Denote

$$\phi_Z(z, z') = K_Z(z, z') - EK_Z(z, Z') - EK_Z(Z, z') + EK_Z(Z, Z'), \tag{2}$$

$$\psi_Y(y, y') = \langle y - EY, y' - EY\rangle = \langle y, y'\rangle - E\langle y, Y\rangle - E\langle Y, y'\rangle + E\langle Y, Y'\rangle,$$

as shown in [13], one can give another expression of the KCMD as follows:

$$\mathrm{KCMD}(Y \mid Z) = E\left[\phi_Z(Z, Z')\psi_Y(Y, Y')\right]. \tag{3}$$

Using the expression, we can provide an unbiased estimator of $\mathrm{KCMD}(Y \mid Z)$. It is closely related to the so-called $\mathcal{U}$-centred matrix.

Let S_n denote the linear span of all $n \times n$ real valued, symmetric matrices with $n > 3$. It is easy to verify that any $A = (a_{ij}) \in S_n$ is a real valued, symmetric matrix. Define the $\mathcal{U}$-centered version of matrix A as $\tilde{A}$, the (i, j)-th elements of $\tilde{A}$ is

$$\tilde{A}_{ij} = \begin{cases} a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}, & i \neq j \\ 0, & i = j \end{cases},$$

$a_{i\cdot} = (\sum_{l=1}^{n} a_{il})/(n-2)$, $a_{\cdot j} = (\sum_{k=1}^{n} a_{kj})/(n-2)$ and $a_{\cdot\cdot} = (\sum_{k,l=1}^{n} a_{kl})/[(n-1)(n-2)]$. Let $H_n = \{\tilde{A} | A \in S_n\}$. We define the inner product of $\tilde{A}$ and $\tilde{B}$ in H_n as

$$\left(\tilde{A} \cdot \tilde{B}\right) = \frac{1}{n(n-3)} \sum_{i \neq j} \tilde{A}_{ij} \tilde{B}_{ij} \tag{4}$$

and $\left|\tilde{A}\right| = \left(\tilde{A} \cdot \tilde{A}\right)^{1/2}$ as the norm of $\tilde{A}$. Theorem 1 in [16] shows that the linear span of all matrices in H_n is a Hilbert space with inner product defined in (4). Using the $\mathcal{U}$-centered matrices, we can construct an estimator of $\mathrm{KCMD}(Y \mid Z)$. Given independent and identically distributed (i.i.d) observations $(Z_i, Y_i)_{i=1}^{n}$ from the joint distribution of (Z, Y), an unbiased estimator of $\mathrm{KCMD}(Y \mid Z)$ provided in [13] is defined as

$$\widehat{\mathrm{KCMD}}(Y \mid Z) = \frac{1}{n(n-3)} \sum_{i \neq j} \tilde{A}_{ij} \tilde{B}_{ij}. \tag{5}$$

Here, $\tilde{A}_{ij}$ and $\tilde{B}_{ij}$ are the $\mathcal{U}$-centred versions of matrixes A and B respectively, and the (i, j)-th elements of A is $a_{ij} = K_Z(Z_i, Z_j)$, the (i, j)-th elements of B is $b_{ij} = \left\|Y_i - Y_j\right\|^2 / 2$. [13] has shown that the estimator (5) is unbiased and admits a U-statistic expression

$$\widehat{\mathrm{KCMD}}(Y \mid Z) = \frac{1}{C_n^4} \sum_{i<j<s<t} h(V_i, V_j, V_s, V_t), \tag{6}$$

with the kernel

$$h(V_i, V_j, V_s, V_t) = \frac{1}{4!} \sum_{(u,v,q,r)}^{(i,j,s,t)} \left(a_{uv} b_{uv} - a_{uv} b_{uq} - a_{uv} b_{vr} + a_{uv} b_{qr}\right),$$

where $V_i = (Z_i, Y_i)$ and the sum is over all 4! permutations of (i, j, s, t).

3. Kernel-Based Partial Conditional Mean Dependence

In this section, we introduce the kernel-based partial conditional mean Dependence (partial KCMD), which can measure the conditional mean dependence of a response Y given a predictor variable X after controlling some variable Z, where X is an random element valued in the separable Hilbert space $\mathcal{X}$.

3.1. Population Partial KCMD

For any symmetric function $f(g, g')$ defined on $\mathcal{G} \times \mathcal{G}$, define $\mathbf{D}$ as a $\mathcal{U}$-centered operator, and

$$\mathbf{D}f(g, g') = f(g, g') - Ef(g, G') - Ef(G, g') + Ef(G, G'),$$

where G is a random element valued in Hilbert space $\mathcal{G}$, G' is an independent copy of G. The functional class $\mathcal{F}=\{\mathbf{D}f : f \text{ is a symmetric function}\}$ is a linear space. Define the inner product on $\mathcal{F}$ as

$$\mathbf{D}f_1 \circ \mathbf{D}f_2 = E\mathbf{D}f_1(G,G')\mathbf{D}f_2(G,G').$$

It can be verified that the map satisfies the conditions of an inner product. In addition, define the norm on $\mathcal{F}$ as $\|\mathbf{D}f\|=\sqrt{\mathbf{D}f \circ \mathbf{D}f}$. Take $\mathcal{G}=\mathcal{Z}\times\mathcal{Y}$. Let $K_Z((z,y),(z',y'))=K_Z(z,z')$ and $\phi_Z((z,y),(z',y'))=\phi_Z(z,z')$ for any $(z,y),(z',y')\in\mathcal{Z}\times\mathcal{Y}$. Then $\phi_Z\in\mathcal{F}$. Similarly, $\psi_Y\in\mathcal{F}$. Thus $\mathrm{KCMD}(Y\mid Z)$ in equation (3) can also be written as

$$\mathrm{KCMD}(Y\mid Z)=\phi_Z\circ\psi_Y. \tag{7}$$

This implies that the KCMD measures the conditional mean dependence of Y on Z through the inner product of ϕ_Z and ψ_Y in a linear space. The cosine value

$$\cos(\phi_Z,\psi_Y)=\frac{\phi_Z\circ\psi_Y}{\|\phi_Z\|\cdot\|\psi_Y\|}$$

measures the strength of conditional mean dependence.

Lemma 2. Suppose $\|\phi_Z\|\neq 0$ and $\|\psi_Y\|\neq 0$, then ψ_Y can be decomposed into two orthogonal parts,

$$\psi_Y=\frac{\|\psi_Y\|\cos(\phi_Z,\psi_Y)}{\|\phi_Z\|}\phi_Z+\left(\psi_Y-\frac{\|\psi_Y\|\cos(\phi_Z,\psi_Y)}{\|\phi_Z\|}\phi_Z\right). \tag{8}$$

The first term in (8) represents the part of Y's conditional mean that is affected by Z, the second term represents the part that can not be interpreted by Z, and in a sense it corresponds to $U=Y-E(Y\mid Z)$ since $E(U\mid Z)=0$. Next we define $W=(X,Z)$ and $\mathcal{W}=\mathcal{X}\times\mathcal{Z}$. One way to measure the additional contribution of X to the conditional mean of Y controlling for Z, is to measure $E(U\mid W)$.

Inspired by [10], we provide definitions for the Kernel-based Partial Conditional Mean Dependence and the Kernel-based Partial Conditional Mean Correlation. Define ϕ_W similar to ϕ_Z, replacing $K_Z(z,z')$ with $K_W(w,w')$ in ϕ_Z is ϕ_W.

Definition 1. The population partial KCMD of Y given X, after controlling for the effect of Z, i.e., $\mathrm{pKCMD}(Y\mid X;Z)$ is defined as

$$\mathrm{pKCMD}(Y\mid X;Z)=\phi_W\circ\left(\psi_Y-\frac{\|\psi_Y\|\cos(\phi_Z,\psi_Y)}{\|\phi_Z\|}\phi_Z\right).$$

If $\|\phi_Z\|=0$, then we define $\mathrm{pKCMD}(Y\mid X;Z)=\phi_W\circ\psi_Y$.

The population Kernel-based Partial Conditional Mean Correlation(pKCMC) is defined as

$$\mathrm{pKCMC}(Y\mid X;Z)=\frac{\mathrm{pKCMD}(Y\mid X;Z)}{\|\phi_W\|\cdot\left\|\psi_Y-\frac{\|\psi_Y\|\cos(\phi_Z,\psi_Y)}{\|\phi_Z\|}\phi_Z\right\|}.$$

If $\|\phi_W\| \cdot \left\| \psi_Y - \frac{\|\psi_Y\| \cos(\phi_Z, \psi_Y)}{\|\phi_Z\|} \phi_Z \right\| = 0$, then we define $\text{pKCMC}(Y \mid X; Z) = 0$.

After performing some straightforward calculations, we obtain an equivalent expression for pKCMD, which is given by

$$\begin{aligned} \text{pKCMD}(Y \mid X; Z) &= \phi_W \circ \psi_Y - \frac{(\phi_Z \circ \psi_Y)(\phi_W \circ \phi_Z)}{\|\phi_Z\|^2} \\ &= \text{KCMD}(Y \mid W) - \frac{\text{KCMD}(Y \mid Z)\,\text{HSIC}(Z, W)}{\text{HSIC}(Z, Z)}. \end{aligned} \tag{9}$$

The $\text{HSIC}(Z, W)$ is the Hilbert-Schmidt Independence Criterion(HSIC) between Z and W, it measures the dependence between these two random elements. The kernel functions used in $\text{HSIC}(Z, W)$ are $K_Z(\cdot,\cdot): \mathcal{Z} \times \mathcal{Z} \to R$ and $K_W(\cdot,\cdot): \mathcal{W} \times \mathcal{W} \to R$, which, like in KCMD, depend on the variables in their subscripts. The content about HSIC can be found in [15] [17]-[19] and so on. We reviewed the specific form of HSIC in the Appendix and derive the last equation of (9). When the conditional mean of Y does not depend on Z or Z is a constant, we have $\text{KCMD}(Y \mid Z) = 0$ or $\text{HSIC}(Z, Z) = 0$. As a result, we have $\text{pKCMD}(Y \mid X; Z) = \text{KCMD}(Y \mid W) = \text{KCMD}(Y \mid X)$.

3.2. Sample pKCMD

Given the sample $(X_i, Y_i, Z_i)_{i=1}^n$, we want to define sample partial KCMD, denoted as $\text{pKCMD}_n(Y \mid X; Z)$ as the sample analog of population partial KCMD. Let $W_i = (X_i, Z_i)$, and define $\tilde{C}$ be $n \times n$ matrix with entries $\tilde{C}_{ij}$,

$$\tilde{C}_{ij} = \begin{cases} c_{ij} - c_{i\cdot} - c_{\cdot j} + c_{\cdot\cdot}, & i \neq j \\ 0, & i = j \end{cases},$$

where $c_{ij} = K_W(W_i, W_j)$, and $c_{i\cdot}$, $c_{\cdot j}$ and $c_{\cdot\cdot}$ are defined similarly to $a_{i\cdot}$, $a_{\cdot j}$ and $a_{\cdot\cdot}$.

Definition 2. Given a random sample from the joint distribution (X, Y, Z), the sample partial kernel-based conditional mean dependence of Y given X, after controlling for the effect of Z, is given by

$$\text{pKCMD}_n(Y \mid X; Z) = (\tilde{C} \cdot \tilde{B}) - \frac{(\tilde{A} \cdot \tilde{B})(\tilde{A} \cdot \tilde{C})}{(\tilde{A} \cdot \tilde{A})},$$

assuming $(\tilde{A} \cdot \tilde{A}) \neq 0$ and $(\tilde{C} \cdot \tilde{B})$ otherwise. The sample partial kernel-based conditional mean correlation is given by

$$\text{pKCMC}_n(Y \mid X; Z) = \frac{|\tilde{A}|^2 \,\text{pKCMD}_n(Y \mid X; Z)}{|\tilde{C}| \cdot \left| (|\tilde{A}|)^2 \tilde{B} - (\tilde{A} \circ \tilde{B}) \tilde{A} \right|}$$

If $|\tilde{C}| \cdot \left| (|\tilde{A}|)^2 \tilde{B} - (\tilde{A} \circ \tilde{B}) \tilde{A} \right| = 0$, then we define $\text{pKCMC}_n(Y \mid X; Z) = 0$.

We next outline theoretical properties of the sample pKCMD. Analogous results hold for the sample pKCMC, which we omit discussing further here.

Theorem 1. If one of the following two conditions holds,

a) $K_Z(\cdot,\cdot)$ and $K_W(\cdot,\cdot)$ are bounded kernels, $E\|Y\|<\infty$;

b) $EK_Z^2(Z,Z')<\infty$, $EK_W^2(W,W')<\infty$, and $E\|Y\|^2<\infty$.

Then, as $n\to\infty$, we have $\text{pKCMD}_n(Y\mid X;Z)\to\text{pKCMD}(Y\mid X;Z)$ a.s..

We also show that $\text{pKCMD}_n(Y\mid X;Z)$ is concentrated. To obtain the bounds of the deviation $\text{pKCMD}_n(Y\mid X;Z)-\text{pKCMD}(Y\mid X;Z)$, we impose the following condition.

(C1) There exists a constant $s_0>0$ such that for all $0<s\le 2s_0$, $E\exp\left(s\|Y\|^2\right)<\infty$.

Condition (C1) follows immediately when Y is bounded uniformly, or when it has a Gaussian distribution. Condition (C1) is widely used in statistical research, for example, in [11] [20] [21], to analyze the theoretical properties of feature screening.

Theorem 2. If $K_Z(\cdot,\cdot)$ and $K_W(\cdot,\cdot)$ are bounded kernels, $E\|Y\|<\infty$, and Condition (C1) holds, then for any $\epsilon>0$, there exist constants $0<\beta<1/2$, $r_1>0$ and $r_2>0$ such that

$$P\left(\left|\text{pKCMD}_n(Y\mid X;Z)-\text{pKCMD}(Y\mid X;Z)\right|\ge\epsilon\right)$$
$$\le O\left(\exp\left(-r_1n^{1-2\beta}\epsilon^2\right)+n\exp\left\{-r_2n^{\beta}\right\}\right).$$

Take $\epsilon=\eta n^{-\gamma}$ with $0<\gamma<1/2$ and a constant $\eta>0$. According to Theorem 2, there exists $0<\beta+\gamma<1/2$, such that $P\left(\left|\text{pKCMD}_n(Y\mid X;Z)-\text{pKCMD}(Y\mid X;Z)\right|\ge\eta n^{-\gamma}\right)=o(1)$. This implies that $\text{pKCMD}_n(Y\mid X;Z)$ is concentrated, and the deviation between $\text{pKCMD}_n(Y\mid X;Z)$ and $\text{pKCMD}(Y\mid X;Z)$ is less than $\eta n^{-\gamma}$ with probability at least $1-O\left(\exp\left(-r_1n^{1-2\beta}\epsilon^2\right)+n\exp\left\{-r_2n^{\beta}\right\}\right)$.

4. Simulation

In the section, we examine tests of the null hypothesis of zero pKCMD. When calculating pKCMD, we need to choose kernel functions $K_Z(\cdot,\cdot)$ and $K_W(\cdot,\cdot)$. We use Gaussian kernels for $K_Z(\cdot,\cdot)$ and $K_W(\cdot,\cdot)$, defined as

$$K_Z(z_i,z_j)=\exp\left\{-\|z_i-z_j\|^2/\left(2\sigma_Z^2\right)\right\}$$

and

$$K_W(w_i,w_j)=\exp\left\{-\|w_i-w_j\|^2/\left(2\sigma_W^2\right)\right\},$$

respectively. For the choice of bandwidths σ_Z and σ_W for these kernels, we can use median heuristic ([13]). We compare our proposed method with pMDD introduced in [10]. We use permutation to obtain the critical values and take the permutation number as 300. The permutation method is described in Section 5 of [10].

The simulations take into account varying sample sizes and levels of dependence between the two random variables to evaluate the performance of tests. For each setting, the empirical sizes or powers of the tests (represented by the proportions

of rejections) are recorded through 1000 repetitions at different significance levels.

Example 1 Generate the i.i.d. sample of (X,Y,Z) from the following model: $X=(X_1,\cdots,X_p)$, $Z=(Z_1,\cdots,Z_p)$, $Y=\cos(Z)$, where $\cos(Z)=(\cos(Z_1),\cdots,\cos(Z_p))$. Consider two scenarios:

1) $X_i \sim N(0,1)$, and $Z_i \sim N(0,1)$.

2) $X_1,\cdots,X_p$ are independent and identically distributed (i.i.d.) random variables from the Cauchy distribution with location parameter 0 and scale parameter 1. $Z_1,\cdots,Z_p$ are i.i.d. random variables from the standard normal distribution.

In this example, X and Z are independent of each other, and Y depends only on Z. Thus, after controlling for the third random vector Z, the conditional mean of Y is independent of X. From Table 1, both methods can reasonably control the type-I error rate.

Table 1. Empirical size of the two tests for Example 1 with $n=50$ and $p=5$.

Scenario	Method	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$
(1)	pKCMD	0.014	0.050	0.098
	pMDD	0.011	0.048	0.105
(2)	pKCMD	0.009	0.051	0.091
	pMDD	0.010	0.059	0.111

Example 2 Generate the i.i.d. sample of (X,Y,Z) from the following model: $X=(X_1,\cdots,X_p)\sim N(\mathbf{0},I_p)$, $Z=(Z_1,\cdots,Z_p)\sim N(\mathbf{0},I_p)$, $Y=f(0.6X+Z)$, where $N(\mathbf{0},I_p)$ is a multivariate normal distribution with zero mean and identity covariance matrix I_p, and $f(x)=(f(x_1),\cdots,f(x_p))$ for any $x=(x_1,\cdots,x_p)$.

We consider the following four relationships for $f(x)$: a) $f(x)=x$, b) $f(x)=x^2$, c) $f(x)=\sin(x)$, and d) $f(x)=\cos(x)$.

Table 2. Empirical powers of the two tests for Example 2 with $p=2$.

Relationship	α	Method	$n=10$	$n=15$	$n=20$	$n=25$	$n=30$
$f(x)=x$	0.01	pKCMD	0.271	0.492	0.664	0.780	0.843
		pMDD	0.424	0.667	0.810	0.900	0.939
	0.05	pKCMD	0.470	0.669	0.801	0.878	0.908
		pMDD	0.606	0.784	0.893	0.943	0.972
	0.10	pKCMD	0.573	0.749	0.856	0.911	0.933
		pMDD	0.698	0.836	0.924	0.955	0.979
$f(x)=x^2$	0.01	pKCMD	0.050	0.102	0.158	0.288	0.421
		pMDD	0.028	0.051	0.055	0.085	0.111

Continued

	0.05	pKCMD	0.165	0.287	0.398	0.538	0.688
		pMDD	0.101	0.141	0.166	0.223	0.273
	0.10	pKCMD	0.270	0.420	0.551	0.696	0.810
		pMDD	0.180	0.238	0.284	0.346	0.420
	0.01	pKCMD	0.235	0.480	0.694	0.835	0.909
		pMDD	0.241	0.474	0.676	0.793	0.897
$f(x)=\sin(x)$	0.05	pKCMD	0.456	0.692	0.843	0.923	0.953
		pMDD	0.465	0.673	0.828	0.904	0.957
	0.10	pKCMD	0.583	0.778	0.905	0.949	0.966
		pMDD	0.602	0.772	0.886	0.940	0.972
	0.01	pKCMD	0.044	0.087	0.158	0.251	0.401
		pMDD	0.027	0.052	0.051	0.084	0.097
$f(x)=\cos(x)$	0.05	pKCMD	0.157	0.276	0.383	0.543	0.679
		pMDD	0.089	0.143	0.168	0.217	0.255
	0.10	pKCMD	0.254	0.416	0.542	0.707	0.816
		pMDD	0.166	0.230	0.279	0.356	0.428

This example compares the empirical powers of pKCMD and pMDD across different functional relationships, significance levels, and sample sizes. According to Table 2, for the linear function $f(x)=x$, pMDD consistently outperforms pKCMD, showing higher sensitivity. In the quadratic $f(x)=x^2$ and cosine $f(x)=\cos(x)$ relationships, pKCMD generally demonstrates superior power, especially at larger samples, indicating better non-linear effect detection. For $f(x)=\sin(x)$ relationship, both tests perform comparably well, with slight advantages for pMDD at lower significance levels. Overall, pMDD excels with linear relationships, while pKCMD is preferable for non-linear ones, particularly with larger sample sizes.

Example 3 Consider the model in Example 2, set $f(x)=\cos(x)$, $n=400$.

Table 3 compares the empirical powers of pKCMD and pMDD at varying significance levels α and varying dimension p, with $n=400$. pKCMD consistently outperforms pMDD across all settings, with its power decreasing more gradually as p increases. At lower significance levels ($\alpha=0.01$), pKCMD maintains high power even with $p=20$, while pMDD's power drops sharply. At higher α, both tests improve, but pKCMD remains superior, especially for larger p. Overall, according to the power shown in Table 3, pKCMD is more robust and powerful, particularly in high-dimensional contexts.

Table 3. Empirical powers of the two tests for Example 3 with $n = 400$.

α	Method	$p = 10$	$p = 20$	$p = 30$	$p = 40$	$p = 50$
0.01	pKCMD	1.000	0.960	0.670	0.399	0.281
	pMDD	0.933	0.385	0.183	0.101	0.091
0.05	pKCMD	1.000	0.995	0.863	0.638	0.506
	pMDD	0.986	0.647	0.413	0.262	0.222
0.10	pKCMD	1.000	0.999	0.937	0.761	0.662
	pMDD	0.998	0.771	0.552	0.394	0.336

Example 4 Consider two models with the formula $Y(t) = f(X(t)) + g(Z(t))$, where $Z(t)$ is generated by Wiener process. Two processes for X, including Ornstein-Uhlenbeck process (OU) and Gaussian process with exponential variogram (VP), are employed, and they are generated by rproc2f data function in R package fda.usc with default parameters. The models for generating $Y(t)$ are as follows:

1) $Y(t) = 1.5X^2(t) + Z(t)$, $t \in [0,1]$.
2) $Y(t) = 2\cos(X(t)) + 0.5Z^3(t)$, $t \in [0,1]$.

Table 4 and Table 5 reveal the performance of pKCMD and pMDD in detecting partial conditional mean dependencies under two distinct models of $Y(t)$. For the quadratic model in (1), pKCMD consistently outperforms pMDD across varying sample sizes and significance levels, with its empirical power improving significantly as the sample size increases. For instance, at $\alpha = 0.05$, pKCMD achieves a power of 0.943 and 0.952 for $n = 50$ with the OU and VP processes, respectively, while pMDD attains only 0.741 and 0.591. This indicates that pKCMD is more effective in capturing the quadratic relationship, and the choice between the OU and VP processes does not substantially affect the relative performance of the tests. For the cosine model in (2), similar trends are observed, with pKCMD maintaining higher power than pMDD across all conditions. The empirical power of both tests increases with the sample size. Although pKCMD shows slightly higher power for the OU process compared to the VP process at larger sample sizes, the difference is not substantial, suggesting that the tests' power is generally robust to the underlying stochastic process of $X(t)$.

Overall, our analysis shows that the tests based on Kernel-based Partial Conditional Mean Dependence perform effectively in most of the situations we examined. As the sample size grows, the power of these tests increases. When compared to pMDD, our pKCMD test proves to be more efficient at capturing nonlinear relationships. Importantly, even when the dimension of variables increases, the decline in our test's power is significantly slower compared to other tests.

Table 4. Empirical powers of the two tests for Example 4 (1).

α	Method	$n=30$		$n=50$	
		OU	VP	OU	VP
0.01	pKCMD	0.606	0.561	0.884	0.863
	pMDD	0.216	0.237	0.449	0.389
0.05	pKCMD	0.777	0.760	0.943	0.952
	pMDD	0.409	0.402	0.741	0.591
0.10	pKCMD	0.835	0.848	0.957	0.973
	pMDD	0.543	0.498	0.845	0.689

Table 5. Empirical powers of the two tests for Example 4 (2).

n	Method	$\alpha=0.01$		$\alpha=0.05$		$\alpha=0.10$	
		OU	VP	OU	VP	OU	VP
30	pKCMD	0.388	0.274	0.526	0.416	0.593	0.519
	pMDD	0.139	0.116	0.320	0.230	0.438	0.328
50	pKCMD	0.472	0.378	0.583	0.542	0.643	0.628
	pMDD	0.206	0.147	0.428	0.309	0.529	0.429
70	pKCMD	0.575	0.459	0.671	0.588	0.730	0.671
	pMDD	0.350	0.219	0.541	0.415	0.638	0.548
90	pKCMD	0.639	0.516	0.743	0.658	0.785	0.742
	pMDD	0.435	0.338	0.644	0.550	0.711	0.633

5. Real Data

In this section, we consider exploring the Tecator dataset contained in R package *fda.usc*. This dataset includes values of a 100-channel spectrum (of wavelength 850 - 1050 mm) of absorbance (Z), water content (X_1), fat content (Y), and protein content (X_2) for 215 meat samples. This dataset has been widely studied in functional data analysis, and this literature mainly focuses on characterizing the influence of functional covariate Z on scalar response Y. Our goal is to determine whether the other two variables, X_1 and X_2, have an impact on the conditional mean of the response after controlling for the influence of Z. For X_1, the p-values computed by pKCMD and pMDD all are 0.000, for X_2, we obtain the same results. These values means that when constructing a conditional mean regression model, X_1 and X_2 should also be considered. [22] has already considered the influence of X_1, X_2, and Z on Y through a semi-functional partial linear regression.

6. Conclusion

This paper introduces pKCMD, a novel test for detecting partial conditional mean dependencies in Hilbert spaces, extending existing measures. We derive equivalent expressions for pKCMD at both population and sample levels. Numerical experiments show that pKCMD consistently outperforms pMDD across sample sizes and significance levels, particularly in nonlinear relationships. From the simulation results, compared with pMDD, pKCMD is more robust and performs better in high-dimensional situations without more computational loss. Overall, pKCMD is a competitive, reliable method for analyzing partial conditional mean dependencies, especially in nonlinear settings. How to choose the optimal kernel function in dependency testing is an important issue. This is our future research direction. In addition, we are also considering extending pKCMD to other tasks, such as conditional independence test, goodness-of-fit test, and feature screening, to broaden its applicability.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cook, R.D. and Li, B. (2002) Dimension Reduction for Conditional Mean in Regression. *The Annals of Statistics*, **30**, 455-474. https://doi.org/10.1214/aos/1021379861

[2] Williamson, B.D., Gilbert, P.B., Carone, M. and Simon, N. (2020) Nonparametric Variable Importance Assessment Using Machine Learning Techniques. *Biometrics*, **77**, 9-22. https://doi.org/10.1111/biom.13392

[3] Dai, B., Shen, X. and Pan, W. (2024) Significance Tests of Feature Relevance for a Black-Box Learner. *IEEE Transactions on Neural Networks and Learning Systems*, **35**, 1898-1911. https://doi.org/10.1109/tnnls.2022.3185742

[4] Williamson, B.D., Gilbert, P.B., Simon, N.R. and Carone, M. (2022) A General Framework for Inference on Algorithm-Agnostic Variable Importance. *Journal of the American Statistical Association*, **118**, 1645-1658. https://doi.org/10.1080/01621459.2021.2003200

[5] Cai, L., Guo, X. and Zhong, W. (2024) Test and Measure for Partial Mean Dependence Based on Machine Learning Methods. *Journal of the American Statistical Association*, **120**, 833-845. https://doi.org/10.1080/01621459.2024.2366030

[6] Welsh, A.H. and Yee, T.W. (2006) Local Regression for Vector Responses. *Journal of Statistical Planning and Inference*, **136**, 3007-3031. https://doi.org/10.1016/j.jspi.2004.01.024

[7] Scheipl, F., Staicu, A. and Greven, S. (2015) Functional Additive Mixed Models. *Journal of Computational and Graphical Statistics*, **24**, 477-501. https://doi.org/10.1080/10618600.2014.901914

[8] Sun, X., Du, P., Wang, X. and Ma, P. (2018) Optimal Penalized Function-on-Function Regression under a Reproducing Kernel Hilbert Space Framework. *Journal of the American Statistical Association*, **113**, 1601-1611. https://doi.org/10.1080/01621459.2017.1356320

[9] Sun, Y. and Wang, Q. (2020) Function-on-Function Quadratic Regression Models. *Computational Statistics & Data Analysis*, **142**, Article ID: 106814.

https://doi.org/10.1016/j.csda.2019.106814

[10] Park, T., Shao, X. and Yao, S. (2015) Partial Martingale Difference Correlation. *Electronic Journal of Statistics*, **9**, 1492-1517. https://doi.org/10.1214/15-ejs1047

[11] 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. https://doi.org/10.1080/01621459.2014.887012

[12] Zhang, X., Yao, S. and Shao, X. (2018) Conditional Mean and Quantile Dependence Testing in High Dimension. *The Annals of Statistics*, **46**, 219-246. https://doi.org/10.1214/17-aos1548

[13] Lai, T., Zhang, Z. and Wang, Y. (2021) A Kernel-Based Measure for Conditional Mean Dependence. *Computational Statistics & Data Analysis*, **160**, Article ID: 107246. https://doi.org/10.1016/j.csda.2021.107246

[14] Fukumizu, K., Gretton, A., Schölkopf, B. and Sriperumbudur, B.K. (2009) Characteristic Kernels on Groups and Semigroups. In: *Advances in Neural Information Processing Systems*, Vol. 21, Curran Associates, 473-480.

[15] Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. (2005) Measuring Statistical Dependence with Hilbert-Schmidt Norms. In: Jain, S., Simon, H.U. and Tomita, E., Eds., *Lecture Notes in Computer Science*, Springer, 63-77. https://doi.org/10.1007/11564089_7

[16] Székely, G.J. and Rizzo, M.L. (2014) Partial Distance Correlation with Methods for Dissimilarities. *The Annals of Statistics*, **42**, 2382-2412. https://doi.org/10.1214/14-aos1255

[17] Albert, M., Laurent, B., Marrel, A. and Meynaoui, A. (2022) Adaptive Test of Independence Based on HSIC Measures. *The Annals of Statistics*, **50**, 858-879. https://doi.org/10.1214/21-aos2129

[18] Balasubramanian, K., Sriperumbudur, B. and Lebanon, G. (2013) Ultrahigh Dimensional Feature Screening via RKHS Embeddings. *Proceedings of the* 16*th International Conference on Artificial Intelligence and Statistics*, Vol. 31, 126-134.

[19] Manfoumbi Djonguet, T.K., Mbina Mbina, A. and Nkiet, G.M. (2024) Testing Independence of Functional Variables by an Hilbert-Schmidt Independence Criterion Estimator. *Statistics & Probability Letters*, **207**, Article ID: 110016. https://doi.org/10.1016/j.spl.2023.110016

[20] Li, R., Zhong, W. and Zhu, L. (2012) Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association*, **107**, 1129-1139. https://doi.org/10.1080/01621459.2012.695654

[21] Wu, Y. and Yin, G. (2015) Conditional Quantile Screening in Ultrahigh-Dimensional Heterogeneous Data. *Biometrika*, **102**, 65-76. https://doi.org/10.1093/biomet/asu068

[22] Aneiros-Pérez, G. and Vieu, P. (2006) Semi-Functional Partial Linear Regression. *Statistics & Probability Letters*, **76**, 1102-1110. https://doi.org/10.1016/j.spl.2005.12.007

[23] Serfling, R.J. (1980) Approximation Theorems of Mathematical Statistics. Wiley. https://doi.org/10.1002/9780470316481

[24] Song, L., Smola, A., Gretton, A., Borgwardt, K. and Bedo, J. (2012) Feature Selection via Dependence Maximization. *Journal of Machine Learning Research*, **13**, 1393-1434.

Appendix

Appendix A1. Hilbert-Schmidt Independent Criterion and Equation (9)

Definition A1. (HSIC) The Hilbert-Schmidt Independent Criterion of random elements $Z \in \mathcal{Z}$ and $W \in \mathcal{W}$ is defined as

$$\begin{aligned}\mathrm{HSIC}(Z,W) = {} & E\left[K_Z(Z,Z')K_W(W,W')\right] + E\left[K_Z(Z,Z')\right]E\left[K_W(W,W')\right] \\ & - 2E\left\{E\left[K_Z(Z,Z')\,|\,Z\right]E\left[K_W(W,W')\,|\,W\right]\right\},\end{aligned}$$

where (Z',W') is an independent copy of (Z,W), $K_Z(\cdot,\cdot)$ and $K_W(\cdot,\cdot)$ are two kernel functions.

Note that in Definition A1, the kernel functions $K_Z(\cdot,\cdot)$ and $K_W(\cdot,\cdot)$ are mutable and depend on Z and W, respectively. HSIC was developed to test the independence of Z and W, and has many good properties. a) It is non negative; b) If $K_Z(\cdot,\cdot)$ and $K_W(\cdot,\cdot)$ are characteristic, then $\mathrm{HSIC}(Z,W)=0$ if and only if Z and W are independent. Using the symbol ϕ_Z defined in (2), symbol ϕ_W is similar to ϕ_Z, but the kernel function used in $K_W(\cdot,\cdot)$ is ϕ_W, we have

$$\begin{aligned}&\phi_Z \circ \phi_W \\ &= E\phi_Z(Z,Z')\phi_W(W,W') \\ &= E\Big(\left\{K_Z(Z,Z') - E\left[K_Z(Z,Z')\,|\,Z\right] - E\left[K_Z(Z,Z')\,|\,Z'\right] + EK_Z(Z,Z')\right\} \\ &\quad \times\left\{K_W(W,W') - E\left[K_W(W,W')\,|\,W\right] - E\left[K_W(W,W')\,|\,W'\right] + EK_W(W,W')\right\}\Big) \\ &= E\left[K_Z(Z,Z')K_W(W,W')\right] + E\left[K_Z(Z,Z')\right]E\left[K_W(W,W')\right] \\ &\quad - 2E\left\{E\left[K_Z(Z,Z')\,|\,Z\right]E\left[K_W(W,W')\,|\,W\right]\right\} \\ &= \mathrm{HSIC}(Z,W),\end{aligned}$$

thus (9) holds.

Appendix A2. The Proofs of Main Results

Proof of Lemma 1. The Lemma 1 is the Proposition 1 in [13]. □

Proof of Lemma 2. Because

$$\frac{\|\psi_Y\|\cos(\phi_Z,\psi_Y)}{\|\phi_Z\|}\phi_Z(z,z') \in \mathcal{F},$$

and

$$\begin{aligned}&\left(\psi_Y - \frac{\|\psi_Y\|\cos(\phi_Z,\psi_Y)}{\|\phi_Z\|}\phi_Z\right)((y,z),(y',z')) \\ &= \psi_Y(y,y') - \frac{\|\psi_Y\|\cos(\phi_Z,\psi_Y)}{\|\phi_Z\|}\phi_Z(z,z') \in \mathcal{F},\end{aligned}$$

their inner product is

$$\begin{aligned}
&\frac{\|\psi_Y\|\cos(\phi_Z,\psi_Y)}{\|\phi_Z\|}\phi_Z\circ\left(\psi_Y-\frac{\|\psi_Y\|\cos(\phi_Z,\psi_Y)}{\|\phi_Z\|}\phi_Z\right)\\
&=\frac{\|\psi_Y\|\cos(\phi_Z,\psi_Y)}{\|\phi_Z\|}(\phi_Z\circ\psi_Y)-\frac{\|\psi_Y\|^2\left[\cos(\phi_Z,\psi_Y)\right]^2}{\|\phi_Z\|^2}(\phi_Z\circ\phi_Z)\\
&=\|\psi_Y\|^2\cos(\phi_Z,\psi_Y)\times\frac{(\phi_Z\circ\psi_Y)}{\|\phi_Z\|\cdot\|\psi_Y\|}-\|\psi_Y\|^2\left[\cos(\phi_Z,\psi_Y)\right]^2\\
&=0,
\end{aligned}$$

thus these two parts in (8) are orthogonal. □

Proof of Theorem 1. a) When $K_Z(\cdot,\cdot)$ and $K_W(\cdot,\cdot)$ are bounded kernels. We suppose that $K_Z(\cdot,\cdot)<M<\infty$ and $K_W(\cdot,\cdot)<M<\infty$. First, we consider $(\tilde{A}\cdot\tilde{B})$ in $\mathrm{pKCMD}_n(Y\mid X;Z)$. Because $(\tilde{A}\cdot\tilde{B})$ is an unbiased estimate of $\phi_Z\circ\psi_Y$ and can be written as an U-statistic by equation (6), and

$$\begin{aligned}
E\left|h(V_1,V_2,V_3,V_4)\right|&\le E\left|a_{uv}\left(b_{uv}-b_{uq}-b_{vr}+b_{qr}\right)\right|\\
&\le ME\left|b_{uv}-b_{uq}-b_{vr}+b_{qr}\right|\\
&\le 4ME\left|b_{uv}\right|\\
&\le 4ME\|Y\|E\|Y'\|\\
&<\infty.
\end{aligned}$$

By the strong law of large number for U-statistics,

$$(\tilde{A}\cdot\tilde{B})\to\phi_Z\circ\psi_Y,a.s..$$

Similarly, we have

$$(\tilde{C}\cdot\tilde{B})\to\phi_W\circ\psi_Y,a.s.,$$

$$(\tilde{A}\cdot\tilde{C})\to\phi_Z\circ\phi_W,a.s.,$$

$$(\tilde{A}\cdot\tilde{A})\to\phi_Z\circ\psi_Z,a.s..$$

According to the continuous mapping theorem, we have $\mathrm{pKCMD}(Y\mid X;Z)\to\mathrm{pKCMD}(Y\mid X;Z)$ a.s..

b) For general kernel.

$$\begin{aligned}
E\left|h(V_1,V_2,V_3,V_4)\right|&\le E\left|a_{uv}\left(b_{uv}-b_{uq}-b_{vr}+b_{qr}\right)\right|\\
&\le Ea_{uv}b_{uv}+Ea_{uv}b_{uq}+Ea_{uv}b_{vr}+Ea_{uv}b_{qr}\\
&\le 2Ea_{uv}^2+2Eb_{uv}^2\\
&\le 2EK_Z^2(Z_u,Z_v)+2\left(E\|Y\|^2\right)^2\\
&\left(\text{because } EK_Z^2(Z,Z')<\infty \text{ and } E\|Y\|^2<\infty\right)\\
&<\infty.
\end{aligned}$$

Similar to (a), we can also prove that $\mathrm{pKCMD}(Y\mid X;Z)\to\mathrm{pKCMD}(Y\mid X;Z)$ a.s.. This completes the proof. □

Proof of Theorem 2. Denote $w=\phi_Z\circ\psi_Y$ and $\hat{w}=\widehat{\mathrm{KCMD}}(Y\mid Z)$. By employing the Markov inequality, we obtain that for any $\epsilon>0$ and $t>0$,

$$\begin{aligned} P\left(\hat{w}-w>\epsilon\right) &= P\left(\exp\left(\hat{w}t-wt\right)>\exp\left(\epsilon t\right)\right) \\ &\le \exp\left(-\epsilon t\right)E\exp\left(\hat{w}t-wt\right) \\ &= \exp\left(-\epsilon t-wt\right)E\exp\left(\hat{w}t\right). \end{aligned}$$

Note that $\hat{w}$ admits a U-statistic expression

$$\hat{w}=\frac{1}{\binom{n}{4}}\sum_{i<j<s<t}h\left(V_i,V_j,V_s,V_t\right),$$

with the kernel $h\left(V_i,V_j,V_s,V_t\right)=\frac{1}{4!}\sum_{(u,v,q,r)}^{(i,j,s,t)}\left(a_{uv}b_{uv}-a_{uv}b_{uq}-a_{uv}b_{vr}+a_{uv}b_{qr}\right)$, where $V_i=\left(Z_i,Y_i\right)$ and the sum is over all 4! permutations of $\left(i,j,s,t\right)$. Following [23] (Section 5.1.6), we write

$$\hat{w}_k=\left(n!\right)^{-1}\sum_{n!}\sum_{i=1}^{m}h\left(V_{(4i-3)},V_{(4i-2)},V_{(4i-1)},V_{(4i)}\right)\Big/m,$$

where where $\sum_{n!}$ denotes the summation over all possible permutations of $\left(1,\cdots,n\right)$, $V_{(i)}$ is the i-th element under the permutation, $m=\lfloor n/4\rfloor$ is the integer part of $n/4$. Denote $h_i=h\left(V_{(4i-3)},V_{(4i-2)},V_{(4i-1)},V_{(4i)}\right)$, write $\hat{w}=\hat{w}_1+\hat{w}_2$, where

$$\hat{w}_1=\left(n!\right)^{-1}\sum_{n!}\sum_{i=1}^{m}h_iI\left(\left|h_i\right|\le M_0\right)\Big/m,$$

$$\hat{w}_2=\left(n!\right)^{-1}\sum_{n!}\sum_{i=1}^{m}h_iI\left(\left|h_i\right|> M_0\right)\Big/m,$$

and $M_0>0$. Correspondingly, its population counterpart can also be decomposed as $w=Eh_1I\left(\left|h_1\right|\le M_0\right)+Eh_1I\left(\left|h_1\right|> M_0\right)=w_1+w_2$.

Jensen inequality yields

$$\begin{aligned} E\exp\left(\hat{w}_1t\right) &= E\exp\left(t\left(n!\right)^{-1}\sum_{n!}\sum_{i=1}^{m}h_iI\left(\left|h_i\right|\le M_0\right)\Big/m\right) \\ &\le E\exp\left(\frac{t}{m}\sum_{i=1}^{m}h_iI\left(\left|h_i\right|\le M_0\right)\right) \\ &= E^m\exp\left(\frac{t}{m}h_iI\left(\left|h_i\right|\le M_0\right)\right), \end{aligned}$$

so

$$P\left(\hat{w}_1-w_1>\epsilon\right)\le\exp\left(-\epsilon t\right)E^m\exp\left(\frac{t}{m}h_iI\left(\left|h_i\right|\le M_0\right)-\frac{t}{m}w_1\right).$$

$Eh_iI\left(\left|h_i\right|\le M_0\right)=w_1$, and $\left|h_iI\left(\left|h_i\right|\le M_0\right)\right|\le M_0$, apply the Lemma 5.6.1.A of [23],

$$E\exp\left(\frac{t}{m}h_iI\left(\left|h_i\right|\le M_0\right)-\frac{t}{m}w_1\right)\le\exp\left(\frac{t^2M_0^2}{2m^2}\right).$$

So,

$$P\left(\hat{w}_1-w_1>\epsilon\right)\le\exp\left(\frac{t^2M_0^2}{2m}-\epsilon t\right).$$

Furthermore, $P\left(\left|\hat{w}_1 - w_1\right| > \epsilon\right) \le 2\exp\left(\frac{t^2 M_0^2}{2m} - \epsilon t\right)$ by the symmetry of the U-statistic.

Now we turn to $\hat{w}_2$. $w_{k2}^2 = E^2 h_1 I\left(\left|h_1\right| > M_0\right) \le E h_1^2 P\left(\left|h_1\right| > M_0\right) \le \frac{E h_1^2 E\left|h_1\right|^{q_1}}{M_0^{q_1}}$ for any $q_1 \in N$. By the assumption that $K_Z(\cdot,\cdot)$ is bounded, there exists a positive constant M such that $\left|a_{uv}\right| < M$, and $\left|h_1\right| = \left|h\left(V_i, V_j, V_s, V_t\right)\right| \le a_{uv}\left|b_{uv} - b_{uq} - b_{vr} + b_{qr}\right| \le M\left(\left\|Y_u\right\|^2 + \left\|Y_v\right\|^2 + \left\|Y_q\right\|^2 + \left\|Y_r\right\|^2\right)$, this yields $E\left|h_1\right|^{q_1} \le 4^{q_1-1} A^{q_1} E\left\|Y\right\|^{2q_1} < \infty$ by condition (C1).

Thus, if we choose $M_0 = n^\beta$ for $0 < \beta < 1/2$, then $\left|w_2\right| \le \epsilon/2$ for sufficiently large n. Hence, $P\left(\left|\hat{w}_2 - w_2\right| \ge \epsilon\right) \le P\left(\left|\hat{w}_2\right| \ge \epsilon/2\right)$.

$$\begin{aligned} P\left(\left|\hat{w}_2\right| \ge \epsilon/2\right) &\le P\left(\bigcup_{i=1}^n \left\{\left\|Y_i\right\|^2 > \frac{M_0}{4M}\right\}\right) \\ &\le \sum_{i=1}^n P\left(\left\|Y_i\right\|^2 > \frac{M_1}{4M}\right) \\ &= nP\left(\left\|Y_i\right\|^2 > \frac{M_0}{4M}\right) \\ &\le nC\exp\left\{-r_2 n^\beta\right\}, \end{aligned}$$

for $r_2 \in \left(0, 2s_0\right]$.

$$\begin{aligned} P\left(\left|\hat{w} - w\right| > 2\epsilon\right) &\le P\left(\left|\hat{w}_1 - w_1\right| > \epsilon\right) + P\left(\left|\hat{w}_2 - w_2\right| > \epsilon\right) \\ &\le \exp\left(\frac{t^2 n^{2\beta}}{2m} - \epsilon t\right) + nC\exp\left\{-r_2 n^\beta\right\}, \end{aligned}$$

choosing $t = \epsilon m / n^{2\beta}$, $P\left(\left|\hat{w} - w\right| \ge \epsilon\right) \le 2\exp\left(-\frac{\epsilon^2 m}{2n^{2\beta}}\right) + nC\exp\left\{-r_2 n^\beta\right\}$, because $m = \lfloor n/4 \rfloor$, we have

$$P\left(\left|\hat{w}_k - w_k\right| \ge \epsilon\right) \le 2\exp\left(-r_1 n^{1-2\beta}\epsilon^2\right) + nC\exp\left\{-r_2 n^\beta\right\},$$

where the constants r_1 satisfy $nr_1 = \frac{m\eta^2}{2}$. Immediately, we have

$$P\left(\left|\hat{w}_k - w_k\right| \ge \epsilon\right) = O\left(\exp\left(-r_1 n^{1-2\beta}\epsilon^2\right) + n\exp\left\{-r_2 n^\beta\right\}\right).$$

Thus, we obtain

$$P\left(\left|\left(\tilde{C}\cdot\tilde{B}\right) - \mathrm{KCMD}\left(Y \mid W\right)\right| \ge \epsilon\right) \le O\left(\exp\left(-r_1 n^{1-2\beta}\epsilon^2\right) + n\exp\left\{-r_2 n^\beta\right\}\right),$$

$$P\left(\left|\left(\tilde{A}\cdot\tilde{B}\right) - \mathrm{KCMD}\left(Y \mid Z\right)\right| \ge \epsilon\right) \le O\left(\exp\left(-r_1 n^{1-2\beta}\epsilon^2\right) + n\exp\left\{-r_2 n^\beta\right\}\right).$$

Because these kernels are bounded, according Theorem 3 in [24], we have

$$P\left(\left|\left(\tilde{A}\cdot\tilde{C}\right) - \mathrm{HSIC}\left(Z, W\right)\right| \ge \epsilon\right) \le O\left(\exp\left(-r_3 n\epsilon^2\right)\right),$$

$$P\left(\left|\left(\tilde{A}\cdot\tilde{A}\right) - \mathrm{HSIC}\left(Z, Z\right)\right| \ge \epsilon\right) \le O\left(\exp\left(-r_3 n\epsilon^2\right)\right).$$

$$
\begin{aligned}
&P\left(\left|\mathrm{pKCMD}_n\left(Y \mid X; Z\right)-\mathrm{pKCMD}\left(Y \mid X; Z\right)\right|\right) \geq \epsilon \\
&\leq P\left(\left|\left(\tilde{C} \cdot \tilde{B}\right)-\frac{\left(\tilde{A} \cdot \tilde{B}\right)\left(\tilde{A} \cdot \tilde{C}\right)}{\left(\tilde{A} \cdot \tilde{A}\right)}-\mathrm{KCMD}\left(Y \mid W\right)+\frac{\mathrm{KCMD}\left(Y \mid Z\right)\mathrm{HSIC}\left(Z, W\right)}{\mathrm{HSIC}\left(Z, Z\right)}\right| \geq \epsilon\right) \\
&\leq P\left(\left|\left(\tilde{C} \cdot \tilde{B}\right)-\mathrm{KCMD}\left(Y \mid W\right)\right| \geq \epsilon / 2\right)+P\left(\left|\left(\tilde{A} \cdot \tilde{B}\right)-\mathrm{KCMD}\left(Y \mid Z\right)\right| \geq C_1 \epsilon\right) \\
&\quad+P\left(\left|\left(\tilde{A} \cdot \tilde{C}\right)-\mathrm{HSIC}\left(Z, W\right)\right| \geq C_2 \epsilon\right)+P\left(\left|\left(\tilde{C} \cdot \tilde{C}\right)-\mathrm{HSIC}\left(Z, Z\right)\right| \geq C_3 \epsilon\right) \\
&\leq O\left(\exp\left(-r_1 n^{1-2\beta} \epsilon^2\right)+n \exp\left\{-r_2 n^{\beta}\right\}\right)+O\left(\exp\left(-r_3 n \epsilon^2\right)\right) \\
&\leq O\left(\exp\left(-r_1 n^{1-2\beta} \epsilon^2\right)+n \exp\left\{-r_2 n^{\beta}\right\}\right).
\end{aligned}
$$

This completes the proof of theorem. □