Journal of Machine Learning Research 23 (2022) 1-58 Submitted 5/21; Revised 3/22; Published 7/22

Kernel Partial Correlation Coefficient — a Measure of Conditional Dependence

Zhen Huang ZH2395@COLUMBIA.EDU
Nabarun Deb ND2560@COLUMBIA.EDU
Bodhisattva Sen BODHI@STAT.COLUMBIA.EDU

Department of Statistics
Columbia University
New York, NY 10027, USA

Editor: Kenji Fukumizu

Abstract

We propose and study a class of simple, nonparametric, yet interpretable measures of conditional dependence, which we call *kernel partial correlation* (KPC) coefficient, between two random variables Y and Z given a third variable X, all taking values in general topological spaces. The population KPC captures the strength of conditional dependence and it is 0 if and only if Y is conditionally independent of Z given X, and 1 if and only if Y is a measurable function of Z and X. We describe two consistent methods of estimating KPC. Our first method is based on the general framework of geometric graphs, including K-nearest neighbor graphs and minimum spanning trees. A sub-class of these estimators can be computed in near linear time and converges at a rate that adapts automatically to the intrinsic dimensionality of the underlying distributions. The second strategy involves direct estimation of conditional mean embeddings in the RKHS framework. Using these empirical measures we develop a fully model-free variable selection algorithm, and formally prove the consistency of the procedure under suitable sparsity assumptions. Extensive simulation and real-data examples illustrate the superior performance of our methods compared to existing procedures.

Keywords: Conditional mean embedding, cross-covariance operator, model-free nonlinear variable selection, nearest neighbor methods, reproducing kernel Hilbert spaces

1. Introduction

Conditional independence is an important concept in modeling causal relationships (Dawid, 1979; Pearl, 2000), in graphical models (Lauritzen, 1996; Koller and Friedman, 2009), in economics (Chiappori and Salanié, 2000), and in the literature of program evaluations (Heckman et al., 1997), among other fields. Measuring conditional dependence has many important applications in statistics such as Bayesian network learning (Pearl, 2000; Spirtes et al., 2000), variable selection (George, 2000; Azadkia and Chatterjee, 2021), dimension reduction (Cook and Li, 2002; Fukumizu et al., 2003/04; Li, 2018), and conditional independence testing (Linton and Gozalo, 1997; Bergsma, 2004; Su and White, 2007; Song, 2009; Huang,

2010; Zhang et al., 2012; Su and White, 2014; Doran et al., 2014; Wang et al., 2015; Patra et al., 2016; Runge, 2018; Ke and Yin, 2020; Shah and Peters, 2020).

Suppose that $(X, Y, Z) \sim P$ where P is supported on a subset of some topological space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ with marginal distributions P_X, P_Y and P_Z on $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ respectively. In this paper, we propose and study a class of simple, nonparametric, yet interpretable measures $\rho^2 \equiv \rho^2(Y, Z|X)$ and their empirical counterparts, that capture the strength of *conditional dependence* between Y and Z, given X. To explain our motivation, consider the case when $\mathcal{Y} = \mathcal{Z} = \mathbb{R}$ and (X, Y, Z) is jointly Gaussian, and suppose that we want to measure the strength of association between Y and Z, with the effect of X removed. In this case a well-known measure of this conditional dependence is the *partial correlation* coefficient $\rho_{YZ \cdot X}$. In particular, the partial correlation squared $\rho^2_{YZ \cdot X}$ is: (i) A deterministic number in $[0, 1]$; (ii) 0 if and only if Y is conditionally independent of Z given X (i.e., $Y \perp\!\!\!\perp Z|X$); (iii) 1 if and only if Y is a (linear) function of Z given X. Moreover, any value between 0 and 1 of $\rho^2_{YZ \cdot X}$ conveys an idea of the strength of the relationship between Y and Z given X.

In this paper we answer the following question in the affirmative: "*Is there a nonparametric generalization of $\rho^2_{YZ \cdot X}$ having the above properties that is applicable to random variables X, Y, Z taking values in general topological spaces and having any joint distribution?*".

In particular, we define a generalization of $\rho^2_{YZ \cdot X}$ — the *kernel partial correlation (KPC) coefficient* — which measures the strength of the conditional dependence between Y and Z given X, that can deal with any distribution P of (X, Y, Z) and is capable of detecting any nonlinear relationships between Y and Z (conditional on X). Moreover, given i.i.d. data from P, we develop and study two different strategies to estimate this population quantity — one based on geometric graph-based methods (Section 3) and the other based on kernel methods using cross-covariance operators (Section 4). We conduct a systematic study of the various computational and statistical properties of these two classes of estimators, including their consistency and (automatic) adaptive properties. We use these measures to develop a provably consistent model-free (high-dimensional) variable selection algorithm (Section 5).

1.1 Kernel Partial Correlation (KPC) Coefficient

Our measure of conditional dependence between Y and Z given X is defined using the framework of *reproducing kernel Hilbert spaces* (RKHSs, see Section 2.1). Let $P_{Y|xz}$ denote the regular conditional distribution of Y given $(X, Z) = (x, z)$, and $P_{Y|x}$ denote the regular conditional distribution of Y given $X = x$. We define the *kernel partial correlation* (KPC) coefficient $\rho^2(Y, Z|X)$ as:

$$\rho^2 \equiv \rho^2(Y, Z|X) := \frac{\mathbb{E}[\mathrm{MMD}^2(P_{Y|XZ}, P_{Y|X})]}{\mathbb{E}[\mathrm{MMD}^2(\delta_Y, P_{Y|X})]}, \tag{1}$$

where MMD is the *maximum mean discrepancy* — a distance metric between two probability distributions depending on a kernel $k_{\mathcal{Y}}(\cdot, \cdot)$ on $\mathcal{Y} \times \mathcal{Y}$, and δ_Y denotes the Dirac measure at Y. We show in Theorem 1 that $\rho^2(Y, Z|X)$ satisfies the following three properties for any joint distribution P of (X, Y, Z):

(i) $\rho^2 \in [0, 1]$; (ii) $\rho^2 = 0$ if and only if $Y \perp\!\!\!\perp Z|X$;

(iii) $\rho^2 = 1$ if and only if Y is a measurable function of Z and X.

Further, $\rho^2(Y, Z|X)$ monotonically increases as the 'dependence' between Y and Z given X becomes stronger: We illustrate this in Propositions 1 and 2 where we consider different kinds of dependence between Y and Z (given X). In Proposition 1 we show that ρ^2, for a large class of kernels and Y, Z being scalars, is a monotonic function of $\rho^2_{YZ\cdot X}$ (the squared partial correlation coefficient), for Gaussian data. Moreover, we show that when the linear kernel is used (i.e., $k_{\mathcal{Y}}(y, y') = yy'$ for $y, y' \in \mathbb{R}$), ρ^2 reduces exactly to $\rho^2_{YZ\cdot X}$. Thus, our proposed measure KPC is indeed a generalization of the classical partial correlation and captures the strength of conditional association.

In Azadkia and Chatterjee (2021) a measure satisfying properties (i)-(iii) was proposed and studied, when Y is a scalar (and Z and X are Euclidean). We show in Lemma 3 that this measure is a special case of our general framework by taking a specific choice of the kernel $k_{\mathcal{Y}}(\cdot, \cdot)$. The advantage of our general framework is that we no longer require Y to be a scalar. In fact, $(X, Y, Z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ can even be non-Euclidean, as long as a kernel function $k_{\mathcal{Y}}$ can be defined on $\mathcal{Y}$, and $\mathcal{X}$ and $\mathcal{Z}$ are, for example, metric spaces. Further, ρ^2, as defined in (1), provides a lot of flexibility to the practitioner as there are a number of kernels known in the literature for different spaces $\mathcal{Y}$ (Lebanon and Lafferty, 2002; Fukumizu et al., 2009b; Danafar et al., 2010; Wynne and Duncan, 2020), some of which may have better properties than others, depending on the application at hand.

In spite of all the above nice properties of ρ^2, it is not immediately clear if we can estimate ρ^2 efficiently, given i.i.d. sample $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ from P. In the following we introduce two estimators of ρ^2 that are easy to implement and enjoy many desirable statistical properties.

1.2 Estimation of KPC Using Geometric Graph-Based Methods

Our first estimation strategy is based on *geometric graphs*; e.g., K-nearest neighbor (K-NN) graphs and minimum spanning trees (MSTs). In Section 3 we describe our method and the resulting estimator $\hat{\rho^2}$ (see Equation 12). In the following we briefly summarize some of the key features of $\hat{\rho^2}$:

1. It can be computed for continuous and categorical/discrete data on Euclidean domains and also for random variables X, Y, Z taking values in general topological spaces, e.g., $\mathcal{X}$ and $\mathcal{X} \times \mathcal{Z}$ being metric spaces and $\mathcal{Y}$ being kernel-endowed would suffice. This can be particularly useful in functional regression (Morris, 2015; Wynne and Duncan, 2020), real-life machine learning and human actions recognition (Danafar et al., 2010), dynamical systems (Song et al., 2009), etc.

2. Although $\hat{\rho^2}$ has a simple interpretable form, it is fully nonparametric. No estimation of conditional densities or distributions (or characteristic functions) is involved.

3. It converges to ρ^2 under very mild assumptions, without any smoothness/continuity conditions on the conditional distributions (Theorem 3).

4. It is $O_p(n^{-1/2})$ concentrated around a population quantity (Proposition 3), under mild assumptions on the kernel. We further establish rates of convergence for $\hat{\rho^2}$ (constructed using K-NN graphs) to ρ^2 that adapt to the *intrinsic dimensionality* of X and (X, Z); see Theorem 4.

5. It can be calculated in near *linear* time (up to logarithmic factors) for a broad variety of metric spaces $\mathcal{X}$ and $\mathcal{X} \times \mathcal{Z}$ (including all Euclidean spaces); see Section 3.1.1.

6. It can be used for variable selection in regression where the response and predictors can take values in general topological spaces. In particular, it provides a stopping criterion for a forward stepwise variable selection algorithm which we call *kernel feature ordering by conditional independence* (KFOCI), inspired by the variable selection algorithm FOCI in Azadkia and Chatterjee (2021) that automatically determines the number of predictor variables to choose. We study the properties of KFOCI which is model-free and is provably consistent even in the high-dimensional regime where the number of covariates grows exponentially with the sample size, under suitable sparsity assumptions. By allowing general kernel functions and different geometric graphs, KFOCI can achieve superior performance when compared to FOCI.

As far as we are aware, our methods are the only procedures that possess all the above mentioned desirable properties.

1.3 Estimation of KPC Using RKHS-Based Methods

As the population version of the KPC coefficient ρ^2 is expressed in terms of MMD, it is natural to ask if kernel methods can be directly used to estimate ρ^2. This is precisely what we do in our second estimation strategy. Observe that the MMD between two distributions is the distance between their *kernel mean embeddings* (see Definition 1). Further, the kernel mean embedding of a conditional distribution, which is usually called the *conditional mean embedding* (CME; see Definition 6), can be expressed in terms of *cross-covariance operators* (see Definition 5) between the two RKHSs (see e.g., Baker, 1973; Fukumizu et al., 2003/04; Song et al., 2009, 2013; Klebanov et al., 2020). As cross-covariance operators can be easily estimated empirically, we can use a plug-in approach to estimate ρ^2, and denote it by $\tilde{\rho^2} \equiv \tilde{\rho^2}(Y, Z|X)$. We refer to $\tilde{\rho^2}$ as the RKHS-based estimator.

We study this estimation strategy in detail in Section 4. In particular, $\tilde{\rho^2}$ can be computed as long as $\mathcal{Y}, \mathcal{X}$ and $\mathcal{X} \times \mathcal{Z}$ are kernel-endowed, using simple matrix operations of the corresponding kernel matrices. The computation can also be accelerated using incomplete Cholesky decomposition. We derive the consistency of this estimator in Theorem 6. In the process of deriving this result we prove the consistency of the plug-in estimator of the CME; this answers an open question stated in Klebanov et al. (2020) and may be of independent interest. Furthermore, $\tilde{\rho^2}$ reduces to the empirical (classical) partial correlation squared when the linear kernel is used. Through extensive simulation studies we show that $\tilde{\rho^2}$ has good finite sample performance in a variety of tasks.

A forward stepwise variable selection algorithm, like KFOCI, can also be devised using the RKHS-based estimator $\tilde{\rho^2}$. However, unlike KFOCI, we need to prespecify the number of variables to be chosen beforehand. Both variable selection algorithms — the one based on $\tilde{\rho^2}$ and the other on $\hat{\rho^2}$ — perform very well in simulated and real data examples as illustrated in the thorough finite sample studies in Section 6.

As a consequence of our general RKHS-based estimation strategy, we can also study the problem of measuring the strength of mutual dependence between Y and Z (when there is

no X), and estimate it using $\tilde{\rho}^2(Y, Z|\emptyset)$; see Remark 10. This complements the graph-based approach of estimating $\rho^2(Y, Z|\emptyset)$ as developed in Deb et al. (2020).

Besides having superior performance on Euclidean spaces, the two proposed estimators are applicable in much more general spaces. In Section 6 we illustrate this by considering two such typical examples: One where the response Y takes values in the *special orthogonal group* SO(3), and the other where we have *compositional data* (Y taking values in the simplex $\mathcal{Y} = \{y \in \mathbb{R}^d : y_1 + \ldots + y_d = 1, y_i \geq 0\}$). In addition, $\hat{\rho}^2$ and $\tilde{\rho}^2$ can also be easily applied in the existing model-X framework (Candès et al., 2018) to yield valid tests for conditional independence and variable selection algorithms with finite sample FDR control. For model-X testing and FDR control, see our full version on arXiv (Huang et al., 2020).

1.4 Related Works

A plethora of procedures — parametric and nonparametric, applicable to discrete and continuous data — have been proposed in the literature, over the last 60 years, to detect conditional dependencies between Y and Z given X (see e.g., Cochran, 1954; Mantel and Haenszel, 1959; Linton and Gozalo, 1997; Bergsma, 2004; Su and White, 2008; Song, 2009; Doran et al., 2014; Candès et al., 2018; Neykov et al., 2021) and the references therein. However none of these methods satisfy property (ii) (as mentioned in Section 1.1). While these methods are indeed useful in practice, they have one common problem: They are all designed primarily for testing conditional independence, and not for measuring the strength of conditional dependence.

In this paper we are interested in nonparametrically measuring the strength of conditional dependence. Measures of conditional dependence such as I^{COND} (defined as the Hilbert-Schmidt norm of a normalized conditional cross-covariance operator, Fukumizu et al., 2008), HSCIC (defined as the Hilbert-Schmidt independence criterion (HSIC) between $P_{Y|X}$ and $P_{Z|X}$; see Gretton et al., 2005; Park and Muandet, 2020), conditional distance covariance CdCov (defined as the distance covariance, as in Székely et al., 2007, between $P_{Y|X}$ and $P_{Z|X}$; see Wang et al., 2015) and HS$\ddot{\text{C}}$IC (defined as the Hilbert-Schmidt norm of a conditional cross-covariance operator; see Fukumizu et al., 2003/04; Sheng and Sriperumbudur, 2019) have the property of always being nonnegative, and they attain the value 0 if and only if $Y \perp\!\!\!\perp Z|X$. However, they do not satisfy properties (i) and (iii) (mentioned in Section 1.1). The conditional distance correlation CdCor (Wang et al., 2015) is normalized between $[0, 1]$, but it is not guaranteed to satisfy property (iii). There are efforts to extend the notion of partial correlation by proposing a kernel generalization (Oh et al., 2018) and a local version (Otneim and Tjøstheim, 2021); however, Oh et al. (2018) is restricted to a regression setting, assuming independent additive noise, whereas Otneim and Tjøstheim (2021), together with HSCIC (Park and Muandet, 2020) and CdCor (Wang et al., 2015), are actually a family of measures indexed by the local variable (rather than a single number). The idea of finding a normalized measure of dependence, like $\rho^2(Y, Z|X)$, has been explored in the recent paper Ke and Yin (2020), but their estimation strategy is very different from the two class of estimators proposed here.

The most relevant work to this paper, and the main motivation behind our work, is the recent paper Azadkia and Chatterjee (2021), where a measure satisfying properties (i)-(iii) was proposed. However, their measure is only applicable to a scalar Y. Our measure KPC

provides a general framework for measuring conditional dependence that is flexible enough to allow the user to choose any kernel of their liking and can handle variables taking values in general topological spaces.

A major application of KPC is variable selection. Common variable selection methods in statistics often posit a parametric model, e.g., by assuming a linear model (Friedman, 1991; Chen and Donoho, 1994; Breiman, 1995; Tibshirani, 1996; Fan and Li, 2001; Miller, 2002; Efron et al., 2004; Zou and Hastie, 2005; Yuan and Lin, 2006; Zou, 2006; Candès and Tao, 2007; Hastie et al., 2009; Ravikumar et al., 2009; Barber and Candès, 2015). These methods are powerful when the underlying parametric assumption holds true, but could have poor performance when the data generating process is more complicated. In general nonlinear settings, although there are popular algorithms for feature selection based on machine learning methods, such as random forests and neural nets (Breiman et al., 1984; Battiti, 1994; Amit and Geman, 1997; Ho, 1998; Breiman, 2001; Hastie et al., 2009; Vergara and Estévez, 2014; Speiser et al., 2019), the performance of these algorithms could depend heavily on how well the machine learning techniques fit the data, and often their theoretical guarantees are weaker than those of model-based methods. Another class of modern nonparametric variable selection methods is by means of fitting penalized smoothing splines (Chen, 1993; Yau et al., 2003; Zhang et al., 2004; Lin and Zhang, 2006; Huang et al., 2010), but these methods still assume certain additive structure of the model. There are also model-free methods in the literature of *sufficient dimension reduction* (SDR) (Cook, 2004; Li et al., 2005; Fukumizu et al., 2009a; Bondell and Li, 2009). However, SDR based methods are not without limitations: (i) some of these methods require strong assumptions such as the *linearity condition* and *constant variance condition*, which may not hold if the predictor is not elliptically distributed; (ii) the dimension of the central subspace are often assumed known.

In this paper, we propose an easily implementable, yet fully model-free variable selection method that is able to automatically determine the number of variables to select, extending the idea of the recent paper Azadkia and Chatterjee (2021). We also formally prove the consistency of our procedure, under suitable assumptions.

1.5 Organization

In Section 2, we introduce and study the population version of the KPC coefficient $\rho^2(Y, Z|X)$. In Section 3, we describe our first method of estimation, based on geometric graphs such as K-NNs and MSTs. Section 4 describes our second estimation strategy, using a RKHS-based method. Applications to variable selection, including a consistency theorem for variable selection, are provided in Section 5. A detailed simulation study and real data analyses are presented in Section 6. Appendix A contains some general discussions that were deferred from the main text of the paper. In Appendix B we provide the proofs of our results.

2. Kernel Partial Correlation (KPC)

Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be topological spaces equipped with Borel probability measures and let $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ be the product space. Suppose that $(X, Y, Z) \sim P$ is a random element on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ with marginal distributions P_X, P_Y and P_Z, on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, respectively. Let P_{XZ} denote the

joint distribution of (X, Z). Recall the notation $P_{Y|xz}$, $P_{Y|x}$ from the Introduction; we will assume the existence of these regular conditional distributions.

2.1 Preliminaries

Let $\mathcal{H}_\mathcal{Y}$ be an RKHS with kernel $k_\mathcal{Y}(\cdot,\cdot)$ on the space $\mathcal{Y}$. By a kernel function $k_\mathcal{Y} : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ we mean a symmetric and nonnegative definite function such that $k_\mathcal{Y}(y,\cdot)$ is a (real-valued) measurable function on $\mathcal{Y}$, for all $y \in \mathcal{Y}$. $\mathcal{H}_\mathcal{Y}$ is a Hilbert space of real-valued functions on $\mathcal{Y}$ such that, for any $f \in \mathcal{H}_\mathcal{Y}$, we have $f(y) = \langle f, k_\mathcal{Y}(y,\cdot)\rangle_{\mathcal{H}_\mathcal{Y}}$, for all $y \in \mathcal{Y}$; this is usually referred to as the *reproducing property* of the kernel $k_\mathcal{Y}(\cdot,\cdot)$. Let us denote the inner product and norm on $\mathcal{H}_\mathcal{Y}$ by $\langle\cdot,\cdot\rangle_{\mathcal{H}_\mathcal{Y}}$ and $\|\cdot\|_{\mathcal{H}_\mathcal{Y}}$ respectively. For an introduction to the theory of RKHS and its applications in statistics we refer the reader to Berlinet and Thomas-Agnan (2004); Steinwart and Christmann (2008). In the following we define two concepts that will be crucial in defining the KPC coefficient.

Definition 1 (Kernel mean embedding) *Suppose that W has a probability distribution Q on $\mathcal{Y}$ such that $\mathbb{E}_Q[\sqrt{k_\mathcal{Y}(W,W)}] < \infty$. There exists a unique $\mu_Q \in \mathcal{H}_\mathcal{Y}$ satisfying*

$$\langle \mu_Q, f\rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_Q[f(W)], \qquad \text{for all } f \in \mathcal{H}_\mathcal{Y}, \tag{2}$$

which is called the (kernel) mean embedding of the distribution Q into $\mathcal{H}_\mathcal{Y}$.

Definition 2 (Maximum mean discrepancy) *The difference between two probability distributions Q_1 and Q_2 on $\mathcal{Y}$ can then be conveniently measured by $\mathrm{MMD}(Q_1,Q_2) := \|\mu_{Q_1} - \mu_{Q_2}\|_{\mathcal{H}_\mathcal{Y}}$ (here μ_{Q_i} is the mean embedding of Q_i, for $i = 1, 2$) which is called the maximum mean discrepancy (MMD) between Q_1 and Q_2. The following alternative representation of the squared MMD is also known (Gretton et al., 2012, Lemma 6):*

$$\mathrm{MMD}^2(Q_1,Q_2) = \mathbb{E}[k_\mathcal{Y}(S,S')] + \mathbb{E}[k_\mathcal{Y}(T,T')] - 2\mathbb{E}[k_\mathcal{Y}(S,T)], \tag{3}$$

where $S, S' \overset{i.i.d.}{\sim} Q_1$ and $T, T' \overset{i.i.d.}{\sim} Q_2$.

Let $\mathcal{P}$ be the class of all Borel probability distributions on $\mathcal{Y}$. The kernel mean embedding defines a map from $\mathcal{P}$ to $\mathcal{H}_\mathcal{Y}$ such that $Q \mapsto \mu_Q$, where $Q \in \mathcal{P}$.

Definition 3 (Characteristic kernel) *The kernel $k_\mathcal{Y}(\cdot,\cdot)$ is said to be characteristic if and only if the kernel mean embedding is injective, i.e., $\mu_{Q_1} = \mu_{Q_2} \implies Q_1 = Q_2$. Note that the last condition is equivalent to $\mathbb{E}_{S\sim Q_1}[f(S)] = \mathbb{E}_{T\sim Q_2}[f(T)]$, for all $f \in \mathcal{H}_\mathcal{Y} \implies Q_1 = Q_2$; this implicitly assumes that the associated RKHS is rich enough.*

Remark 1 (Examples of characteristic kernels) *A number of popular characteristic kernels have been studied in the literature. Some popular ones in $\mathbb{R}^d$ include the Gaussian kernel (i.e., $k(u,v) := \exp(-\sigma\|u-v\|^2)$ with $\sigma > 0$) (Sriperumbudur et al., 2011), the Laplace kernel (i.e., $k(u,v) = \exp(-\sigma\|u-v\|_1)$ with $\sigma > 0$ where $\|\cdot\|_1$ denotes the L_1-norm) (Sriperumbudur et al., 2011) and the distance kernel (Sejdinovic et al., 2013)*

$$k(u,v) := 2^{-1}\Big(\|u\|^\alpha + \|v\|^\alpha - \|u-v\|^\alpha\Big), \qquad \text{for all } u, v \in \mathbb{R}^d, \tag{4}$$

for $\alpha \in (0, 2)$. See Fukumizu et al. (2008, 2009b); Danafar et al. (2010); Wynne and Duncan (2020) for other examples of characteristic kernels on more general topological spaces. Sufficient conditions for a kernel to be characteristic are discussed in Fukumizu et al. (2008); Sriperumbudur et al. (2008); Fukumizu et al. (2009b); Sriperumbudur et al. (2010, 2011); Szabó and Sriperumbudur (2017).

2.2 KPC: The Population Version

We are now ready to formally define and study our measure *kernel partial correlation* (KPC) coefficient $\rho^2 \equiv \rho^2(Y, Z|X)$. Let $\mathcal{H}_\mathcal{Y}$ be an RKHS on $\mathcal{Y}$ with kernel $k_\mathcal{Y}(\cdot, \cdot)$. Recall the definition of ρ^2 from (1):

$$\rho^2(Y, Z|X) := \frac{\mathbb{E}[\mathrm{MMD}^2(P_{Y|XZ}, P_{Y|X})]}{\mathbb{E}[\mathrm{MMD}^2(\delta_Y, P_{Y|X})]},$$

where δ_Y denotes the Dirac measure at Y. To study the various properties of ρ^2 defined above we will assume the following regularity conditions:

Assumption 1 *$k_\mathcal{Y}(\cdot, \cdot)$ is characteristic and $\mathbb{E}[k_\mathcal{Y}(Y, Y)] < \infty$.*

Assumption 2 *$\mathcal{H}_\mathcal{Y}$ is separable.*

Assumption 3 *Y is not a measurable function of X; equivalently, $Y|X = x$ is not degenerate for almost every (a.e.) x.*

Remark 2 (On Assumptions 1–3) *Note that $\mathbb{E}[k_\mathcal{Y}(Y, Y)] < \infty$ in Assumption 1 is very common in the kernel literature (see e.g., Baker, 1973; Fukumizu et al., 2007, 2008, 2009a; Park and Muandet, 2020). See Remark 3 below as to why a characteristic kernel is necessary; see Remark 1 for some examples of characteristic kernels. Assumption 2 is needed for technical reasons and can be ensured under mild conditions[1]; see Remark 15 for a detailed discussion. Assumption 3 just ensures that Y is not degenerate given X, so that the denominator of $\rho^2(Y, Z|X)$ is not 0; see Remark 16 for a proof of the equivalence.*

We will assume that Assumptions 1–3 hold throughout the paper unless otherwise specified. The following lemma (proved in Appendix B.1) shows that $\rho^2(Y, Z|X)$, as defined in (1), is well-defined.

Lemma 1 *Under Assumptions 1–3, $\rho^2(Y, Z|X)$ in (1) is well-defined.*

The following lemma (proved in Appendix B.2) gives another alternate form for ρ^2 which will be especially useful to us while constructing estimators of ρ^2.

Lemma 2 *Suppose that Assumptions 1–3 hold. Then, we have*

$$\rho^2(Y, Z|X) = \frac{\mathbb{E}\left[\mathbb{E}[k_\mathcal{Y}(Y_2, Y_2')|X, Z]\right] - \mathbb{E}\left[\mathbb{E}[k_\mathcal{Y}(Y_1, Y_1')|X]\right]}{\mathbb{E}[k_\mathcal{Y}(Y, Y)] - \mathbb{E}[\mathbb{E}[k_\mathcal{Y}(Y_1, Y_1')|X]]}, \tag{5}$$

1. For example, if $\mathcal{Y}$ is a separable space and $k_\mathcal{Y}(\cdot, \cdot)$ is continuous, then $\mathcal{H}_\mathcal{Y}$ is separable; see e.g., Hein and Bousquet (2004, Theorem 7), Steinwart and Christmann (2008, Lemma 4.33).

where in the denominator (X, Y_1, Y_1') *has the following joint distribution:*

$$X \sim P_X, \qquad Y_1|X \sim P_{Y|X}, \qquad Y_1'|X \sim P_{Y|X}, \quad \textit{and} \quad Y_1 \perp\!\!\!\perp Y_1'|X; \tag{6}$$

and in the numerator (X, Y_2, Y_2', Z) *has the following joint distribution:*

$$(X, Z) \sim P_{XZ}, \quad Y_2|X, Z \sim P_{Y|XZ}, \quad Y_2'|X, Z \sim P_{Y|XZ}, \quad \textit{and} \quad Y_2 \perp\!\!\!\perp Y_2'|X, Z. \tag{7}$$

Our first main result, Theorem 1 (proved in Appendix B.3), shows that indeed, under the above assumptions (i.e., Assumptions 1–3), our measure of conditional dependence satisfies the three desirable properties (i)-(iii) mentioned in the Introduction.

Theorem 1 *Under Assumptions 1–3,* $\rho^2(Y, Z|X)$ *in* (1) *satisfies:*

(i) $\rho^2(Y, Z|X) \in [0, 1]$*;*

(ii) $\rho^2(Y, Z|X) = 0$ *if and only if* Y *is conditionally independent of* Z *given* X *(i.e.,* $Y \perp\!\!\!\perp Z|X$*); equivalently,* $\rho^2 = 0$ *if and only if* $P_{Y|XZ} = P_{Y|X}$ *almost surely (a.s.);*

(iii) $\rho^2(Y, Z|X) = 1$ *if and only if* Y *is a measurable function of* Z *and* X *a.s. (equivalently,* Y *given* X *and* Z *is a degenerate random variable a.s.).*

Remark 3 (Characteristic kernel) *A close examination of the proof of Theorem 1 (in Section B.3) reveals that* $k_{\mathcal{Y}}(\cdot, \cdot)$ *being characteristic is used for: (a) Proving* $\rho^2 = 0$ *implies* $Y \perp\!\!\!\perp Z|X$*; (b) proving that* $\rho^2 = 1$ *implies* Y *is a function of* X *and* Z *(here actually the weaker assumption that the feature map* $y \mapsto k_{\mathcal{Y}}(y, \cdot)$ *is injective would have sufficed); (c) the denominator of* ρ^2 *is non-zero.*

In the following we provide two results that go beyond Theorem 1 and illustrate that KPC indeed measures conditional association—any value of KPC between 0 and 1 conveys an idea of the strength of the association between Y and Z, given X. In Proposition 1 below (proved in Appendix B.4) we show that when the underlying distribution is multivariate Gaussian, ρ^2 is a strictly monotonic function of the classical partial correlation coefficient squared $\rho^2_{YZ \cdot X}$ (for a large class of kernels), and equals $\rho^2_{YZ \cdot X}$ if the linear kernel (i.e., $k_{\mathcal{Y}}(u, v) = u^\top v$ for $u, v \in \mathbb{R}^d, d \geq 1$) is used. In the following we will restrict attention to kernels having the following form:

$$k_{\mathcal{Y}}(u, v) = h_1(u) + h_2(v) + h_3(\|u - v\|), \qquad \text{for } u, v \in \mathbb{R}^d \tag{8}$$

where h_i $(i = 1, 2, 3)$ are arbitrary real-valued functions, $\|\cdot\|$ is the usual Euclidean norm in $\mathbb{R}^d$, and h_3 is nonincreasing. Note that the Gaussian and distance kernels in Remark 1 and the linear kernel (and the Laplace kernel when $d = 1$) are all of this form.

Proposition 1 (Connection to classical partial correlation) *Suppose that* (Y, Z, X) *are jointly normal with* Y *and* Z *(given* X*) having (classical) partial correlation* $\rho_{YZ \cdot X}$*. Suppose the kernel* $k_{\mathcal{Y}}(\cdot, \cdot)$ *has the form in* (8) *where* h_3 *is assumed to be strictly decreasing. Then:*

(a) $\rho^2(Y, Z|X)$ *is a strictly increasing function of* $\rho^2_{YZ \cdot X}$*, provided* $\mathrm{Var}(Y|X) > 0$ *is held fixed (as we change the joint distribution of* (Y, Z, X)*).*

(b) $\rho^2(Y, Z|X) = 0$ *(resp. 1) if and only if* $\rho^2_{YZ \cdot X} = 0$ *(resp. 1).*

(c) If the distance kernel is used (see Equation 4), $\rho^2(Y, Z|X)$ is a strictly increasing function of $\rho^2_{YZ\cdot X}$, irrespective of the value of $\mathrm{Var}(Y|X)$.
(d) If the linear kernel is used, then $\rho^2(Y, Z|X) = \rho^2_{YZ\cdot X}$.

Next, we consider the two extreme cases: (i) If $Y = g(X)$, then $Y \perp\!\!\!\perp Z|X$; (ii) if $Y = f(X, Z)$, then Y is a function of Z given X and consequently $\rho^2 = 1$. The following proposition (proved in Appendix B.5) states that if Y follows a regression model with noise, and we are somewhere in between the above two extreme cases (i) and (ii), expressed as a convex combination, then $\rho^2(Y, Z|X)$ is monotonic in the weight used in the convex combination, regardless of how complicated the dependencies $g(X)$ and $f(X, Z)$ are.

Proposition 2 (Monotonicity) *Consider the following regression model with $\mathcal{Y} = \mathbb{R}$, and arbitrary measurable functions $g : \mathcal{X} \to \mathcal{Y}$ and $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$, and $\lambda \in [0, 1]$:*

$$Y = (1 - \lambda)g(X) + \lambda f(X, Z) + \epsilon,$$

where ϵ is the noise variable (independent of X and Z) such that for another independent copy ϵ', $\epsilon - \epsilon'$ is unimodal[2]*. Then $\rho^2(Y, Z|X)$ is monotone nondecreasing in λ, when a kernel of the form* (8) *is used to define ρ^2.*

In addition to the properties already described above, ρ^2 also possesses important *invariance* and *continuity* properties; see Appendix A.3 for a detailed discussion on this.

When Y is a scalar, in Azadkia and Chatterjee (2021) a measure $T(Y, Z|X)$, satisfying properties (i)-(iii) of Theorem 1 was proposed and studied. More specifically,

$$T(Y, Z|X) := \frac{\int \mathbb{E}(\mathrm{Var}(\mathbb{P}(Y \geq t|Z, X)|X))\mathrm{d}P_Y(t)}{\int \mathbb{E}(\mathrm{Var}(1_{Y \geq t}|X))\mathrm{d}P_Y(t)}.$$

Indeed, as we see in the following result (see Appendix B.7 for a proof), $T(Y, Z|X)$ can actually be seen as a special case of ρ^2, for a suitable choice of the kernel $k_{\mathcal{Y}}(\cdot, \cdot)$.

Lemma 3 *$T(Y, Z|X) = \rho^2(Y, Z|X)$ when we use the kernel $k_{\mathcal{Y}}(y_1, y_2) := \int 1_{y_1 \geq t} 1_{y_2 \geq t} \mathrm{d}P_Y(t)$, for $y_1, y_2 \in \mathbb{R}$.*[3] *Further, if Y has a continuous cumulative distribution function F_Y, then $T(Y, Z|X) = \rho^2(F_Y(Y), Z|X)$ where we consider the distance kernel, as in* (4) *with $\alpha = 1$.*

Thus, our measure ρ^2 can be thought of as a generalization of $T(Y, Z|X)$ in Azadkia and Chatterjee (2021), allowing X, Y, Z to take values in more general spaces. In fact, our framework is more general, as we allow for 'any' choice of the kernel $k_{\mathcal{Y}}(\cdot, \cdot)$.

Remark 4 (Measuring association between Y and Z) *Consider the special case when there is no X, i.e.,*

$$\rho^2(Y, Z|X) = \rho^2(Y, Z|\emptyset) = \frac{\mathbb{E}[\mathrm{MMD}^2(P_{Y|Z}, P_Y)]}{\mathbb{E}[\mathrm{MMD}^2(P_{\delta_Y}, P_Y)]}.$$

2. A random variable with distribution function F is unimodal about v if $F(x) = p\delta_v(x) + (1 - p)F_1(x)$ for some $p \in [0, 1]$, where δ_v is the distribution function of the Dirac measure at v, and F_1 is an absolutely continuous distribution function with density nondecreasing on $(-\infty, v]$ and nonincreasing on $[v, \infty)$ (Purkayastha, 1998). If ϵ is unimodal, then $\epsilon - \epsilon'$ is symmetric and unimodal about 0 (Purkayastha, 1998, Theorem 2.2).
3. Note that this kernel is not characteristic, but the distance kernel mentioned later is characteristic.

Now, $\rho^2(Y, Z|\emptyset)$ can be used to measure the unconditional dependence between Y and Z. This measure $\rho^2(Y, Z|\emptyset)$ has been proposed and studied in detail in the recent paper Deb et al. (2020); also see Ke and Yin (2020, Section 2.4). In particular, as illustrated in Deb et al. (2020), $\rho^2(Y, Z|\emptyset)$ can be effectively estimated using graph-based methods (in near linear time) and can also be readily used to test the hypothesis of mutual independence between Y and Z. In Section 4 we also develop an RKHS-based estimator of $\rho^2(Y, Z|\emptyset)$.

Let us now discuss some properties of ρ^2 when we use the linear kernel (i.e., $k_{\mathcal{Y}}(u, v) = u^\top v$, for $u, v \in \mathcal{Y} = \mathbb{R}^d$, with $d \geq 1$). Suppose $Y = (Y^{(1)}, \ldots, Y^{(d)})^\top \in \mathcal{Y} = \mathbb{R}^d$. Then, from (5), we have the following expression for $\rho^2(Y, Z|X)$:

$$\rho^2(Y, Z|X) = \frac{\sum_{i=1}^d \mathbb{E}\left(\mathrm{Var}\left[\mathbb{E}(Y^{(i)}|X, Z)|X\right]\right)}{\sum_{i=1}^d \mathbb{E}[\mathrm{Var}(Y^{(i)}|X)]}. \tag{9}$$

Remark 5 (Connection to Zhang and Janson, 2020) *When $d = 1$, the numerator of $\rho^2(Y, Z|X)$ is equal to the minimum mean squared error gap (mMSE gap): $\mathbb{E}\big[(Y - \mathbb{E}[Y|X])^2\big] - \mathbb{E}\left[(Y - \mathbb{E}[Y|X, Z])^2\right]$, which has been used to quantify the conditional relationship between Y and Z given X in the recent paper Zhang and Janson (2020). Note that mMSE gap is not invariant under arbitrary scalings of Y, but $\rho^2(Y, Z|X)$ (which is equal to the squared partial correlation $\rho^2_{YZ \cdot X}$; see Proposition 1) is. Thus, $\rho^2(Y, Z|X)$ can be viewed as a normalized version of the mMSE gap.*

Remark 6 (Linear kernel and Theorem 1) *As the linear kernel is not characteristic, ρ^2, defined with the linear kernel, does not satisfy all the three properties in Theorem 1. It satisfies (i) and (iii): If Y is a function of X and Z, then $\rho^2 = 1$. Conversely, if $\rho^2 = 1$, then $\mathbb{E}[\mathrm{Var}(Y^{(i)}|X, Z)] = 0$ for all i in (9), which implies that $Y^{(i)}|X, Z$ is degenerate for all i, i.e., Y is a function of X and Z. However, it is not guaranteed to satisfy (ii): If $Y \perp\!\!\!\perp Z|X$, then indeed $\rho^2 = 0$; but $\rho^2 = 0$ does not necessarily imply that $Y \perp\!\!\!\perp Z|X$.[4] This is because the linear kernel is not characteristic. However, if (Y, X, Z) is jointly normal, then $\rho^2 = 0$ does imply $Y \perp\!\!\!\perp Z|X$ (see Remark 18) and ρ^2 satisfies all three properties.*

3. Estimating KPC with Geometric Graph-Based Methods

Suppose that $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ are i.i.d. observations from P on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Here we assume that $\mathcal{X}$ and $\mathcal{X} \times \mathcal{Z}$ are metric spaces and we have a kernel function $k_{\mathcal{Y}}(\cdot, \cdot)$ defined on $\mathcal{Y}$. In fact, $\mathcal{X}$ and $\mathcal{X} \times \mathcal{Z}$ can be even more general spaces—with metrics being replaced by certain semimetrics, "similarity" functions or "divergence" measures (see e.g., Boytsov and Naidan, 2013; Athitsos et al., 2004; Miranda et al., 2013; Jacobs et al., 2000; Gottlieb et al., 2017). In this section we propose and study a general framework—using *geometric graphs*—to estimate $\rho^2(Y, Z|X)$ (as in Equation 1). We will estimate each of the terms in (1) separately to obtain our final estimator $\hat{\rho^2}$ of $\rho^2(Y, Z|X)$. Note that $\mathbb{E}[k_{\mathcal{Y}}(Y, Y)]$ in (1) can be easily estimated by $\frac{1}{n}\sum_{i=1}^n k_{\mathcal{Y}}(Y_i, Y_i)$. So we will focus on estimating the two

4. A counter example: Let X, Z, U be i.i.d. having a continuous distribution with mean 0 and let $Y := X + ZU$. Then $\rho^2 = 0$ but Y is not independent of Z given X.

other terms—$\mathbb{E}[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]]$ and $\mathbb{E}[\mathbb{E}[k_{\mathcal{Y}}(Y_2, Y_2')|X, Z]]$; recall the joint distributions of (X, Y_1, Y_1') and (X, Z, Y_2, Y_2') as mentioned in (6), (7). As our estimation strategy for both the above terms is similar, let us first focus on estimating $T := \mathbb{E}[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]]$. To motivate our estimator of T, let us consider a simple case, where X_i's are categorical, i.e., take values in a finite set. A natural estimator for T in that case would be

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\#\{j : X_j = X_i\}} \sum_{j:X_j=X_i} k_{\mathcal{Y}}(Y_i, Y_j), \tag{10}$$

as in Ke and Yin (2020, Section 3.1). In this section, instead of assuming X_i's are categorical, we will focus on general distributions for X_i's, typically continuous.

We will use the notion of *geometric graph functionals* on $\mathcal{X}$ (see Deb et al., 2020; Bhattacharya, 2019). $\mathcal{G}$ is said to be a geometric graph functional on $\mathcal{X}$ if, given any finite subset S of $\mathcal{X}$, $\mathcal{G}(S)$ defines a graph with vertex set S and corresponding edge set, say $\mathcal{E}(\mathcal{G}(S))$, which is invariant under any permutation of the elements in S. The graph can be both directed or undirected, and we will restrict ourselves to simple graphs, i.e., graphs without multiple edges and self loops. Examples of such functionals include minimum spanning trees (MSTs) and K-nearest neighbor (K-NN) graphs, as described below. Define $\mathcal{G}_n := \mathcal{G}(X_1, \ldots, X_n)$ where $\mathcal{G}$ is some graph functional on $\mathcal{X}$.

1. **K-NN graph**: The directed K-NN graph $\mathcal{G}_n$ puts an edge from each node X_i to its K-NNs among $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ (so X_i is excluded from the set of its K-NNs). Ties will be broken at random if they occur, to ensure the out-degree is always K. The undirected K-NN graph is obtained by ignoring the edge direction in the directed K-NN graph and removing multiple edges if they exist.

2. **MST**: An MST is a subset of edges of an edge-weighted undirected graph connecting all the vertices with the least possible sum of edge weights and contains no cycles. For instance, given a set of points $X_1, \ldots, X_n$ one can construct an MST for the complete graph with vertices as X_i's and edge weights being the distances among them.

In order to estimate $\mathbb{E}[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]]$, ideally we would like to have multiple Y_i's (say Y_i''s) from the conditional distribution $P_{Y|X_i}$, so as to average over all such $k_{\mathcal{Y}}(Y_i, Y_i')$. However, this is rarely possible in real data (if X_1 is continuous, for example). As a result, our strategy is to find X_j's that are "close" to X_i and average over all such $k_{\mathcal{Y}}(Y_i, Y_j)$. The notion of geometric graph functionals comes in rather handy in formalizing this notion. The key intuition is to define a graph functional $\mathcal{G}$ where X_i and X_j are connected (via an edge) in $\mathcal{G}_n := \mathcal{G}(X_1, \ldots, X_n)$ provided they are "close". Towards this direction, let us define the following statistic (as in Deb et al., 2020):

$$T_n(Y, X) \equiv T_n := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{d_i} \sum_{j:(i,j)\in\mathcal{E}(\mathcal{G}_n)} k_{\mathcal{Y}}(Y_i, Y_j), \tag{11}$$

where $\mathcal{G}_n = \mathcal{G}(X_1, \ldots, X_n)$ for some graph functional $\mathcal{G}$ on $\mathcal{X}$, and $\mathcal{E}(\mathcal{G}_n)$ denotes the set of (directed/undirected) edges of $\mathcal{G}_n$, i.e., $(i, j) \in \mathcal{E}(\mathcal{G}_n)$ if and only if there is an edge from $i \to j$ if $\mathcal{G}_n$ is a directed graph, or an edge between i and j if $\mathcal{G}_n$ is an undirected graph.

Here $d_i := |\{j : (i,j) \in \mathcal{E}(\mathcal{G}_n)\}|$ denotes the degree (or out-degree in a directed graph) of X_i in $\mathcal{G}_n$. Note that when $\mathcal{G}_n$ is undirected, $(i,j) \in \mathcal{E}(\mathcal{G}_n)$ if and only if $(j,i) \in \mathcal{E}(\mathcal{G}_n)$.

The next natural question is: "Does T_n, as defined in (11), consistently estimate T?". The following result in Deb et al. (2020, Theorem 3.1) answers this question in the affirmative, under appropriate assumptions on the graph functional.

Theorem 2 (Deb et al., 2020, Theorem 3.1) *Suppose $\mathcal{G}_n$ satisfies Assumptions 10–12 (detailed in Appendix A.5) and $\mathcal{H}_\mathcal{Y}$ is separable. For $\theta > 0$, let $\mathcal{M}^\theta_{k_\mathcal{Y}}(\mathcal{Y})$ be the collection of all Borel probability measures Q over $\mathcal{Y}$ such that $\mathbb{E}_Q\left[k^\theta_\mathcal{Y}(Y,Y)\right] < \infty$. If $P_Y \in \mathcal{M}^{2+\epsilon}_{k_\mathcal{Y}}(\mathcal{Y})$ for some fixed $\epsilon > 0$, then $T_n \xrightarrow{p} T$. If $P_Y \in \mathcal{M}^{4+\epsilon}_{k_\mathcal{Y}}(\mathcal{Y})$ for some fixed $\epsilon > 0$, then $T_n \xrightarrow{a.s.} T$.*

Note that Assumptions 10–12 required on the graph functional $\mathcal{G}_n$ for the above result were made in Deb et al. (2020, Theorem 3.1); see Deb et al. (2020, Section 3) for a detailed discussion on these assumptions. It can be further shown that for the K-NN graph and the MST, these assumptions are satisfied under mild assumptions. For example, in the Euclidean space, they hold for K-NN graph when $\|X_1 - X_2\|$ has a continuous distribution and $K = o(n/\log n)$. For the MST these assumptions are satisfied when X_1 has an absolutely continuous distribution (Deb et al., 2020, Proposition 3.2).

Remark 7 (K-NN graph versus MST) *In practice, a K-NN graph is recommended as a primary choice of the geometric graph over MST. This is because of the following reasons. Firstly, in order to ensure connectivity, MST sometimes has to place an edge between two distant points, which is not desirable for an accurate estimation of ρ^2. Secondly, the MST in practice often achieves similar performance as a 2-NN graph (note that the average degree of an MST is $2 - \frac{2}{n}$), and K-NN graphs offer more flexibility. When using $\hat{\rho^2}$ to estimate ρ^2, we recommend using a small K since a K-NN, for a large K, can be far from the point under consideration, especially in high dimensions. However, for testing and variable selection, a larger K is seen to be beneficial; see Section 6 for more simulation evidence. Thirdly, MST has higher computational complexity when compared to a K-NN graph.*

3.1 Estimation of $\hat{\rho^2}$

In the previous subsection we constructed a very general geometric graph-based consistent estimator T for $\mathbb{E}[\mathbb{E}[k_\mathcal{Y}(Y_1, Y_1')|X]]$. We will use a similar strategy to estimate the other term $\mathbb{E}[\mathbb{E}[k_\mathcal{Y}(Y_2, Y_2')|X, Z]]$, i.e., we define a geometric graph functional on the space $\mathcal{X} \times \mathcal{Z}$ and construct an estimator like T_n in (11), but now the geometric graph is defined on $\mathcal{X} \times \mathcal{Z}$ with the data points $\{(X_i, Z_i)\}_{i=1}^n$.

For simplicity of notation, we let $\ddot{X} := (X, Z)$ and $\ddot{\mathcal{X}} := \mathcal{X} \times \mathcal{Z}$. Let $\mathcal{G}_n^X$ (resp. $\mathcal{G}_n^{\ddot{X}}$) be the graph constructed based on $\{X_i\}_{i=1}^n$ (resp. $\{\ddot{X}_i \equiv (X_i, Z_i)\}_{i=1}^n$). Let d_i^X (resp. $d_i^{\ddot{X}}$) be the degree of X_i (resp. $\ddot{X}_i$) in $\mathcal{G}_n^X$ (resp. $\mathcal{G}_n^{\ddot{X}}$), for $i = 1, \ldots, n$. We are now ready to define our graph-based estimator of ρ^2:

$$\hat{\rho^2} \equiv \hat{\rho^2}(Y, Z|X) := \frac{\frac{1}{n}\sum_{i=1}^n \frac{1}{d_i^{\ddot{X}}} \sum_{j:(i,j)\in\mathcal{E}(\mathcal{G}_n^{\ddot{X}})} k_\mathcal{Y}(Y_i, Y_j) - \frac{1}{n}\sum_{i=1}^n \frac{1}{d_i^X} \sum_{j:(i,j)\in\mathcal{E}(\mathcal{G}_n^X)} k_\mathcal{Y}(Y_i, Y_j)}{\frac{1}{n}\sum_{i=1}^n k_\mathcal{Y}(Y_i, Y_i) - \frac{1}{n}\sum_{i=1}^n \frac{1}{d_i^X} \sum_{j:(i,j)\in\mathcal{E}(\mathcal{G}_n^X)} k_\mathcal{Y}(Y_i, Y_j)}. \quad (12)$$

The estimator $\hat{\rho^2}$ is consistent for ρ^2; this follows easily[5] from Theorem 2. We formalize this in the next result.

Theorem 3 (Consistency) *Suppose that Assumptions 10–12 (see Appendix A.5) hold for both $\mathcal{G}_n^X$ and $\mathcal{G}_n^{\ddot{X}}$. If $P_Y \in \mathcal{M}_{k_\mathcal{Y}}^{2+\varepsilon}(\mathcal{Y})$ (as in Theorem 2) for some $\varepsilon > 0$, then $\hat{\rho^2}(Y, Z|X) \xrightarrow{p} \rho^2(Y, Z|X)$. If $P_Y \in \mathcal{M}_{k_\mathcal{Y}}^{4+\varepsilon}(\mathcal{Y})$ for some $\varepsilon > 0$, then $\hat{\rho^2}(Y, Z|X) \xrightarrow{a.s.} \rho^2(Y, Z|X)$.*

A salient aspect of Theorem 3 is that consistency of $\hat{\rho^2}(Y, Z|X)$ does not need any continuity assumptions of the conditional distributions $P_{Y|x}$ and $P_{Y|xz}$, as $x \in \mathcal{X}$ and $z \in \mathcal{Z}$ vary. Our approach leverages Lusin's Theorem (Lusin, 1912), which states that any measurable function agrees with a continuous function on a "large" set. This is a generalization of the technique used in Azadkia and Chatterjee (2021) and Chatterjee (2021).

The following result (see Appendix B.8 for a proof) provides a concentration bound for T_n and states that $\hat{\rho^2}$ is $O_p(n^{-1/2})$-concentrated around a population quantity, if the underlying kernel is bounded.

Proposition 3 (Concentration) *Under the same assumptions as in Theorem 3 (except Assumption 10) on the two graphs $\mathcal{G}_n^X$ and $\mathcal{G}_n^{\ddot{X}}$, and provided $\sup_{y\in\mathcal{Y}} k_\mathcal{Y}(y, y) \leq M$ for some $M > 0$, there exists a fixed positive constant C^* (free of n and t), such that for any $t > 0$, the following holds:*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i^X}\sum_{j:(i,j)\in\mathcal{E}(\mathcal{G}_n^X)} k_\mathcal{Y}(Y_i, Y_j) - \mathbb{E}[k_\mathcal{Y}(Y_1, Y_{N(1)})]\right| \geq t\right] \leq 2\exp(-C^* n t^2), \tag{13}$$

where $N(1)$ is a uniformly sampled index from the neighbors (or out-neighbors in a directed graph) of X_1 in $\mathcal{G}_n^X$. A similar sub-Gaussian concentration bound also holds for the term $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i^{\ddot{X}}}\sum_{j:(i,j)\in\mathcal{E}(\mathcal{G}_n^{\ddot{X}})} k_\mathcal{Y}(Y_i, Y_j)$, when centered around $\mathbb{E}[k_\mathcal{Y}(Y_1, Y_{\ddot{N}(1)})]$; here $\ddot{N}(1)$ is a uniformly sampled neighbor of (X_1, Z_1) in $\mathcal{G}_n^{\ddot{X}}$. Consequently,

$$\sqrt{n}\left(\hat{\rho^2} - \frac{\mathbb{E}[k_\mathcal{Y}(Y_1, Y_{\ddot{N}(1)})] - \mathbb{E}[k_\mathcal{Y}(Y_1, Y_{N(1)})]}{\mathbb{E}[k_\mathcal{Y}(Y_1, Y_1)] - \mathbb{E}[k_\mathcal{Y}(Y_1, Y_{N(1)})]}\right) = O_p(1).$$

The above result shows that $\hat{\rho^2}$ has a rate of convergence $n^{-1/2}$ around a limit which is not necessarily ρ^2. Further, by Proposition 3, it is clear that the rate of convergence of $\hat{\rho^2}$ to ρ^2 will be chiefly governed by the rates at which $\mathbb{E}[k_\mathcal{Y}(Y_1, Y_{N(1)})]$ and $\mathbb{E}[k_\mathcal{Y}(Y_1, Y_{\ddot{N}(1)})]$ converge to $\mathbb{E}[k_\mathcal{Y}(Y_1, Y_1')]$ and $\mathbb{E}[k_\mathcal{Y}(Y_2, Y_2')]$ respectively. As it turns out this rate of convergence is heavily dependent on the underlying graph functional $\mathcal{G}$. In Section 3.2 we will focus on K-NN graphs and provide an upper bound on this rate of convergence.

5. By the strong law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} k_\mathcal{Y}(Y_i, Y_i) \xrightarrow{a.s.} \mathbb{E}[k_\mathcal{Y}(Y, Y)]$. The result now follows from Theorem 2 and the continuous mapping theorem.

3.1.1 Computational Complexity of $\hat{\rho^2}$

Observe that, when using the Euclidean K-NN graph, the computation of $\hat{\rho^2}(Y, Z|X)$ takes $O(Kn \log n)$ time. This is because the K-NN graph can be found in $O(Kn \log n)$ time (e.g., using the k-d tree; see Bentley, 1975) and the K-NN graph has $O(Kn)$ edges[6] (thus, we just have to sum over $O(Kn)$ terms in computing each of the two main quantities in Equation 12). The computational complexity of computing Euclidean MSTs is $O(n \log n)$ in $\mathbb{R}^d$ when $d = 1$ or 2 (Shamos and Hoey, 1975; Buchin and Mulzer, 2011), and $O(n^{2-2^{-(d+1)}}(\log n)^{1-2^{-(d+1)}})$ when $d \geq 3$ (Yao, 1982). As an MST has just $n - 1$ edges, the computational complexity for computing $\hat{\rho^2}$, using the MST, is of the same order as that of finding the MST. Thus, in Euclidean settings, with $K = O(\log^\gamma n)$, $\gamma \geq 0$, for K-NN graph (or $d = 1, 2$ for MST), $\hat{\rho^2}$ can be computed in near linear time (up to logarithmic factors).

Several authors have proposed tree-based data structures to speed up K-NN graph construction. Examples include ball-trees (Omohundro, 1989) and cover-trees (Beygelzimer et al., 2006). In Beygelzimer et al. (2006) the authors study K-NN graphs in general metric spaces and show that if the data set has a bounded expansion constant (which is a measure of its intrinsic dimensionality) the cover-tree data structure can be constructed in $O(n \log n)$ time for bounded K.

3.2 Rate of Convergence of $\hat{\rho^2}$

In this subsection, we will assume that the geometric graph functionals $\mathcal{G}_n^X$ and $\mathcal{G}_n^{\ddot{X}}$ belong to the family of K-NN graphs (directed or undirected) on the spaces $\mathcal{X}$ and $\ddot{\mathcal{X}}$ (assumed to be general metric spaces equipped with metrics $\rho_{\mathcal{X}}$ and $\rho_{\ddot{\mathcal{X}}}$) respectively. Our main result in this subsection, Theorem 4, shows that $\hat{\rho^2}$ converges to ρ^2, at a rate that depends on the *intrinsic dimensions* of X and $\ddot{X}$, as opposed to the ambient dimensions of $\mathcal{X}$ and $\ddot{\mathcal{X}}$. This highlights the *adaptive* nature of the estimator $\hat{\rho^2}$. Let us first define the intrinsic dimensionality of a random variable, which is a relaxation of the *Assouad dimension* (Robinson, 2011, Section 9). Recall that by a *cover* of a subset $A \subset (\mathcal{X}, \rho_{\mathcal{X}})$, we mean a collection of subsets of $\mathcal{X}$ whose union contains A. Denote by $B(x^*, r) \subset \mathcal{X}$ the closed ball centered at $x^* \in \mathcal{X}$ with radius $r > 0$, and by $\text{supp}(X)$ the support of the random variable X.

Definition 4 (Intrinsic dimension of X) *Let X be a random variable taking values in a metric space $\mathcal{X}$. X is said to have intrinsic dimension at most d, with constant $C > 0$ at $x^* \in \mathcal{X}$, if for any $t > 0$, $B(x^*, t) \cap \text{supp}(X)$ can be covered with at most $C(t/\varepsilon)^d$ closed balls of radius ε in $\mathcal{X}$, for any $\varepsilon \in (0, t]$.*

This notion of the intrinsic dimensionality of X extends the usual notion of dimension of a Euclidean set; see e.g., Chen and Shah (2018, Definition 3.3.1) and Deb et al. (2020, Definition 5.1). For example, any probability measure on $\mathbb{R}^d$ has intrinsic dimension at most d. Moreover, if X is supported on a d_0-dimensional hyperplane (where $d_0 \leq d$), then X has intrinsic dimension at most d_0. Further, if X is supported on a d_0-dimensional manifold which is bi-Lipschitz[7] to some $\Omega \subset \mathbb{R}^{d_0}$, then X has intrinsic dimension at most

6. For a directed graph, there are exactly Kn edges; for an undirected graph, there are no more than Kn edges, but each edge will be used twice in the summation.
7. A manifold $\mathcal{M}$ with metric $\rho_{\mathcal{M}}$ is said to be *bi-Lipschitz* to $\Omega \subset \mathbb{R}^{d_0}$ if there exists $L > 0$ and a bijection $\varphi : \Omega \to (\mathcal{M}, \rho_{\mathcal{M}})$ satisfying $L^{-1}\|x_1 - x_2\| \leq \rho_{\mathcal{M}}(\varphi(x_1), \varphi(x_2)) \leq L\|x_1 - x_2\|$, for all $x_1, x_2 \in \Omega$.

d_0 (Robinson, 2011, Lemma 9.3). Note that the intrinsic dimensions in the above examples are valid at every $x^* \in \text{supp}(X)$, but Theorem 4 only requires its validity at one x^*. The intrinsic dimension need not be an integer, and can be defined on any metric space.

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_X$, and let $\mathcal{G}_n^X$ be the K_n-NN (here $\{K_n\}_{n\geq 1}$ is assumed to be a fixed sequence). We will assume the following conditions:

Assumption 4 *$\rho_{\mathcal{X}}(X_1, X_2)$ has a continuous distribution.*

Assumption 5 *X has intrinsic dimension at most d with constant C_1 at $x^* \in \mathcal{X}$.*

Assumption 6 *Let s_n be the number of points having X_1 as a K_n-NN. Suppose that $\frac{s_n}{K_n} \leq C_2$ a.s., for some constant $C_2 > 0$ and all $n \geq 1$.*

Assumption 7 *There exist $\alpha, C_3, C_4 > 0$ such that $\mathbb{P}\left(\rho_{\mathcal{X}}(X_1, x^*) \geq t\right) \leq C_3 \exp(-C_4 t^\alpha)$ for all $t > 0$, where $x^* \in \mathcal{X}$ is defined in Assumption 5.*

Assumption 8 *Set $g(x) := \mathbb{E}[k_{\mathcal{Y}}(Y, \cdot)|X = x]$ for $x \in \mathcal{X}$. There exist $\beta_1 \geq 0$, $\beta_2 \in (0, 1]$, $C_5 > 0$ such that for any $x_1, x_2, \tilde{x}_1, \tilde{x}_2 \in \mathcal{X}$,*

$$\begin{aligned}\left|\langle g(x_1), g(x_2)\rangle_{\mathcal{H}_{\mathcal{Y}}} - \langle g(\tilde{x}_1), g(\tilde{x}_2)\rangle_{\mathcal{H}_{\mathcal{Y}}}\right| \leq C_5\big(1 + \rho_{\mathcal{X}}(x^*, x_1)^{\beta_1} + \rho_{\mathcal{X}}(x^*, x_2)^{\beta_1} \\ + \rho_{\mathcal{X}}(x^*, \tilde{x}_1)^{\beta_1} + \rho_{\mathcal{X}}(x^*, \tilde{x}_2)^{\beta_1}\big)\left(\rho_{\mathcal{X}}(x_1, \tilde{x}_1)^{\beta_2} + \rho_{\mathcal{X}}(x_2, \tilde{x}_2)^{\beta_2}\right),\end{aligned}$$

where $x^* \in \mathcal{X}$ is defined in Assumption 5.

Remark 8 (On the assumptions) *Assumption 4 guarantees that the K_n-NN graph is uniquely defined. If $\mathcal{X}$ is a Euclidean space, Assumption 6 is satisfied because the number of points having X_1 as a K-NN is bounded by $KC(d)$, where $C(d)$ is a constant depending only on the dimension d of $\mathcal{X}$ (Yukich, 1998, Lemma 8.4). Assumption 7 just says that X_1 satisfies a tail decay condition that can be even slower than sub-exponential. Assumption 8 is a technical condition on the smoothness of conditional expectation. Without such an assumption, the rate of convergence of $\hat{\rho^2}$ to ρ^2 may be arbitrarily slow (Azadkia and Chatterjee, 2021). See Deb et al. (2020, Proposition 5.1) for sufficient conditions under which Assumption 8 holds. Note that similar assumptions were also made in Azadkia and Chatterjee (2021); in fact our assumptions are less stringent in the sense that they allow for: (a) Any general metric space $\mathcal{X}$, (b) tail decay rates of X slower than sub-exponential, and (c) β_2 to vary in $(0, 1]$. $g(\cdot)$ is also called the conditional mean embedding (see Definition 6).*

The following result (see Appendix B.9 for a proof) gives an upper bound on the rate of convergence of $\hat{\rho^2}$. In particular, it shows that, if $d > 2$ is an upper bound on the intrinsic dimensions of X and (X, Z) then $\hat{\rho^2}$ converges to ρ^2 at the rate $O_p\left(n^{-\beta_2/d}\right)$, up to a logarithmic factor (provided K_n grows no faster than a power of $\log n$). Note that in certain situations, while the actual dimension of $\mathcal{X}$ (resp. $\ddot{\mathcal{X}}$) may be large, the intrinsic dimensionality of X (resp. $\ddot{X}$) may be much smaller—the rate of convergence of $\hat{\rho^2}$ automatically adapts to the unknown intrinsic dimensions of X and $\ddot{X}$.

Theorem 4 (Adaptive rate of convergence) *Suppose $k_{\mathcal{Y}}(Y_1, Y_1)$ has sub-exponential tail*[8]*. Let $K_n = o(n(\log n)^{-1})$. Suppose that Assumptions 4–8 hold for (Y, X) and also for $(Y, \ddot{X})$ (i.e., by replacing X with $\ddot{X}$ in each of the Assumptions 4–8) with the same constants α and β_2 (in Assumptions 7 and 8). Define*

$$\nu_n := \begin{cases} \frac{(\log n)^2}{n} + \frac{K_n}{n}(\log n)^{2\beta_2/d + 2\beta_2/\alpha} & \text{if } d < 2\beta_2, \\ \frac{(\log n)^2}{n} + \frac{K_n}{n}(\log n)^{2+d/\alpha} & \text{if } d = 2\beta_2, \\ \frac{(\log n)^2}{n} + \left(\frac{K_n}{n}\right)^{2\beta_2/d}(\log n)^{2\beta_2/d + 2\beta_2/\alpha} & \text{if } d > 2\beta_2, \end{cases}$$

where d is the maximum of the intrinsic dimensions of X and $\ddot{X}$ (in Assumption 5). Then

$$\hat{\rho^2}(Y, Z|X) = \rho^2(Y, Z|X) + O_p\left(\sqrt{\nu_n}\right).$$

Although Theorem 4 has many similarities with Azadkia and Chatterjee (2021, Theorem 4.1), our result is more general on various fronts: (a) We can handle $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ taking values in general metric spaces; (b) our upper bound ν_n depends on the intrinsic dimensions of X and $\ddot{X}$ as opposed to their ambient dimensions; (c) our assumptions are less stringent (as discussed in Remark 8); (d) our upper bound ν_n is also sharper in the logarithmic factor. A similar result, as Theorem 4, can be found in Deb et al. (2020) for estimating $\rho^2(Y, Z|\emptyset)$.

4. Estimating KPC Using RKHS Methods

As the population version of KPC ρ^2 is expressed in terms of MMD, it is natural to ask if the RKHS framework can be directly used to estimate ρ^2. This is precisely what we do in our second estimation strategy. In this section, we further assume that $\mathcal{X}$ and $\ddot{\mathcal{X}} = \mathcal{X} \times \mathcal{Z}$ are equipped with separable RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\ddot{\mathcal{X}}}$ respectively, with kernels $k_{\mathcal{X}}$ and $k_{\ddot{\mathcal{X}}}$. Let $\ddot{X} = (X, Z) \sim P_{XZ}$. We also assume that $\mathbb{E}[k_{\mathcal{X}}(X, X)] < \infty$ and $\mathbb{E}[k_{\ddot{\mathcal{X}}}(\ddot{X}, \ddot{X})] < \infty$.

In the following we define two concepts—the *(cross)-covariance operator* and *conditional mean embedding*—that will be fundamental in the developments of this section.

Definition 5 (Cross-covariance operator) *The cross-covariance operator $C_{XY} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{X}}$ is the unique bounded linear operator that satisfies $\langle g, C_{XY} f\rangle_{\mathcal{H}_{\mathcal{X}}} = \mathrm{Cov}(f(Y), g(X))$ for all $f \in \mathcal{H}_{\mathcal{Y}}$, $g \in \mathcal{H}_{\mathcal{X}}$.*

The existence of the cross-covariance operator follows from the Riesz representation theorem (Fukumizu et al., 2003/04, Theorem 1). The *covariance operator* of X, denoted by C_X, is obtained when the two RKHSs in Definition 5 are the same, namely $\mathcal{H}_{\mathcal{X}}$. Note that the covariance operator C_X is bounded, nonnegative, self-adjoint, and trace-class if $\mathcal{H}_{\mathcal{X}}$ is separable and $\mathbb{E}[k_{\mathcal{X}}(X, X)] < \infty$. We direct readers to Rynne and Youngson (2008) and Aubin (2011) for basic concepts from functional analysis.

The following explicit representation of the cross-covariance operator C_{XY} will be useful:

$$C_{XY} = \mathbb{E}[k_{\mathcal{X}}(\cdot, X) \otimes k_{\mathcal{Y}}(\cdot, Y)] - \mathbb{E}[k_{\mathcal{X}}(\cdot, X)] \otimes \mathbb{E}[k_{\mathcal{Y}}(\cdot, Y)] \tag{14}$$

8. It means $\mathbb{P}(k_{\mathcal{Y}}(Y_1, Y_1) \geq t) \leq L_1 e^{-L_2 t}$ for all $t > 0$, for some $L_1, L_2 > 0$.

where we have identified the tensor product space $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ with the space of Hilbert-Schmidt operators from $\mathcal{H}_{\mathcal{Y}}$ to $\mathcal{H}_{\mathcal{X}}$, such that $k_{\mathcal{X}}(\cdot, u) \otimes k_{\mathcal{Y}}(\cdot, v)(h) := h(v)k_{\mathcal{X}}(\cdot, u)$, for all $u \in \mathcal{X}, v \in \mathcal{Y}$ and $h \in \mathcal{H}_{\mathcal{Y}}$. See Remark 17 for more details on how (14) is derived.

Definition 6 (Conditional mean embedding, CME) *The CME $\mu_{Y|x} \in \mathcal{H}_{\mathcal{Y}}$, for $x \in \mathcal{X}$, is defined as the kernel mean embedding of the conditional distribution of Y given $X = x$, i.e., $\mu_{Y|x} = \mathbb{E}_{Y \sim P_{Y|x}}[k_{\mathcal{Y}}(Y, \cdot)]$.*

CMEs have proven to be a powerful tool in various machine learning applications, such as dimension reduction (Fukumizu et al., 2003/04), dynamic systems (Song et al., 2009), hidden Markov models (Song et al., 2010a), and Bayesian inference (Fukumizu et al., 2013); see the recent paper Klebanov et al. (2020) for a rigorous treatment. Under certain assumptions, cross-covariance operators can be used to provide simpler expressions of CMEs; see e.g., Klebanov et al. (2020). The following assumption is crucial for this purpose.

Assumption 9 *For any $g \in \mathcal{H}_{\mathcal{Y}}$, there exists $h \in \mathcal{H}_{\mathcal{X}}$ such that $\mathbb{E}[g(Y)|X = \cdot] - h(\cdot)$ is constant P_X-a.e.*

Lemma 4 (Klebanov et al., 2020, Theorem 4.3) *Suppose $\mathbb{E}[k_{\mathcal{X}}(X, X)]$ and $\mathbb{E}[k_{\mathcal{Y}}(Y, Y)]$ are finite, and both $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}$ are separable. Let C_{XY} and C_X be the usual cross-covariance and covariance operators respectively. Further let $\operatorname{ran} C_X$ denote the range of C_X and let $C_X^\dagger$ denote the Moore-Penrose inverse (see e.g., Engl et al., 1996, Definition 2.2) of C_X. Suppose further that Assumption 9 holds. Then $\operatorname{ran} C_{XY} \subset \operatorname{ran} C_X$, $C_X^\dagger C_{XY}$ is a bounded linear operator, and for P_X-a.e. $x \in \mathcal{X}$,*

$$\mu_{Y|x} = \mu_Y + \left(C_X^\dagger C_{XY}\right)^* (k_{\mathcal{X}}(x, \cdot) - \mu_X) \tag{15}$$

where for a bounded operator A, A^* is the adjoint of A, and μ_X and μ_Y are the kernel mean embeddings of P_X and P_Y respectively (i.e., $\mu_X = \mathbb{E}[k_{\mathcal{X}}(X, \cdot)] \in \mathcal{H}_{\mathcal{X}}$).

Remark 9 *Note that we do not require $k_{\mathcal{X}}(\cdot, \cdot)$ and $k_{\mathcal{Y}}(\cdot, \cdot)$ to be characteristic for* (15) *to be valid. See Klebanov et al. (2020) for other sufficient conditions, different from Assumption 9, that guarantee* (15)*. The CME formula given in* (15) *is the centered version which uses the centered (cross)-covariance operators C_X and C_{XY}. Uncentered covariance operators have also been used to define CMEs in the existing literature (see e.g., Song et al., 2010a,b; Fukumizu et al., 2013); see Appendix A.2 for a discussion. But it is known that the centered CME formula* (15) *requires less restrictive assumptions and hence is preferable. Our simulation results also validated this observation, and hence in this paper we advocate the use of the centered CME formula (as in Equation 15). However in practice, sufficient conditions for explicit expressions of CMEs in terms of centered and uncentered (cross)-covariance operators (as Assumption 9) are usually hard to verify (Klebanov et al., 2020).*

The following result (which follows from Lemma 4) shows that $\rho^2(Y, Z|X)$ can be expressed in terms of CMEs, which in turn, can be explicitly simplified in terms of (cross)-covariance operators under appropriate assumptions (as in Equation 15). This will form

our basis for estimation of $\rho^2(Y, Z|X)$ using the RKHS framework—we will replace each of the terms in (16) below with their sample counterparts to obtain the estimator $\tilde{\rho^2}$ (see Equation 18 below).

Proposition 4 *Suppose that the assumptions in Lemma 4 hold for (Y, X) and $(Y, \ddot{X})$ (i.e. replacing X with $\ddot{X}$). Then $\rho^2(Y, Z|X)$ in (1) can be simplified as*

$$\begin{aligned}\rho^2 &= \frac{\mathbb{E}\big[\|\mu_{Y|XZ} - \mu_{Y|X}\|^2_{\mathcal{H}_\mathcal{Y}}\big]}{\mathbb{E}[\|k_\mathcal{Y}(Y,\cdot) - \mu_{Y|X}\|^2_{\mathcal{H}_\mathcal{Y}}]} \\ &= \frac{\mathbb{E}\Big[\big\|(C^\dagger_{\ddot{X}} C_{\ddot{X}Y})^*\big(k_{\ddot{\mathcal{X}}}(\ddot{X},\cdot) - \mu_{\ddot{X}}\big) - (C^\dagger_X C_{XY})^*\left(k_\mathcal{X}(X,\cdot) - \mu_X\right)\big\|^2_{\mathcal{H}_\mathcal{Y}}\Big]}{\mathbb{E}\Big[\big\|k_\mathcal{Y}(Y,\cdot) - \mu_Y - (C^\dagger_X C_{XY})^*\left(k_\mathcal{X}(X,\cdot) - \mu_X\right)\big\|^2_{\mathcal{H}_\mathcal{Y}}\Big]}. \end{aligned} \tag{16}$$

Here, for $x \in \mathcal{X}$ and $z \in \mathcal{Z}$, $\mu_{Y|xz}$ is the CME of Y given $X = x$ and $Z = z$.

4.1 Estimation of ρ^2 by $\tilde{\rho^2}$

Suppose that we have i.i.d. data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ from P on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Let us first consider the estimation of the covariance operator C_X. The empirical covariance operator $\hat{C}_X$ is easily estimated by the sample analogue of (14), i.e., by replacing the expectations in (14) by their empirical counterparts:

$$\hat{C}_X := \frac{1}{n}\sum_{i=1}^n k_\mathcal{X}(X_i,\cdot) \otimes k_\mathcal{X}(X_i,\cdot) - \hat{\mu}_X \otimes \hat{\mu}_X \tag{17}$$

where $\hat{\mu}_X := \frac{1}{n}\sum_{i=1}^n k_\mathcal{X}(X_i,\cdot)$ is the estimator of the kernel mean embedding $\mu_X = \mathbb{E}[k_\mathcal{X}(X,\cdot)]$. Similarly, the cross-covariance operator C_{YX} can be estimated by $\hat{C}_{YX} := \frac{1}{n}\sum_{i=1}^n k_\mathcal{Y}(Y_i,\cdot) \otimes k_\mathcal{X}(X_i,\cdot) - \hat{\mu}_Y \otimes \hat{\mu}_X$, where $\hat{\mu}_Y := \frac{1}{n}\sum_{i=1}^n k_\mathcal{Y}(Y_i,\cdot)$ is the estimator of the kernel mean embedding $\mu_Y = \mathbb{E}[k_\mathcal{Y}(Y,\cdot)]$. Further note that $\operatorname{ran}\hat{C}_X$ is spanned by the set $\{k_\mathcal{X}(X_i,\cdot) : i = 1,\ldots,n\}$, which implies that $\hat{C}_X$ is not invertible in general, since $\mathcal{H}_\mathcal{X}$ is typically infinite-dimensional. In fact, estimating the inverse of the compact operator C_X is in general an ill-posed inverse problem (Manton and Amblard, 2014, Section 8.6). Hence the Tikhonov approach is often used for regularization which estimates $C^\dagger_X$ by $(\hat{C}_X + \varepsilon I)^{-1}$, for a tuning parameter $\varepsilon > 0$ (e.g., Song et al., 2010a,b; Fukumizu et al., 2013). Thus, $(C^\dagger_X C_{XY})^*$ can be estimated by $\big((\hat{C}_X + \varepsilon I)^{-1}\hat{C}_{XY}\big)^* = \hat{C}_{YX}(\hat{C}_X + \varepsilon I)^{-1}$. ρ^2 in (16) can therefore be naturally estimated empirically by

$$\tilde{\rho^2} := \frac{\frac{1}{n}\sum_{i=1}^n \|\hat{\mu}_{Y|\ddot{X}_i} - \hat{\mu}_{Y|X_i}\|^2_{\mathcal{H}_\mathcal{Y}}}{\frac{1}{n}\sum_{i=1}^n \|k_\mathcal{Y}(Y_i,\cdot) - \hat{\mu}_{Y|X_i}\|^2_{\mathcal{H}_\mathcal{Y}}}, \tag{18}$$

where, for $x \in \mathcal{X}$, and $\ddot{x} \in \ddot{\mathcal{X}}$,

$$\begin{aligned}\hat{\mu}_{Y|\ddot{x}} &:= \hat{\mu}_Y + \hat{C}_{Y\ddot{X}}(\hat{C}_{\ddot{X}} + \varepsilon I)^{-1}\big(k_{\ddot{\mathcal{X}}}(\ddot{x},\cdot) - \hat{\mu}_{\ddot{X}}\big), \\ \hat{\mu}_{Y|x} &:= \hat{\mu}_Y + \hat{C}_{YX}(\hat{C}_X + \varepsilon I)^{-1}\big(k_\mathcal{X}(x,\cdot) - \hat{\mu}_X\big)\end{aligned} \tag{19}$$

and $\hat{\mu}_{\ddot{X}} := \frac{1}{n}\sum_{i=1}^{n} k_{\ddot{\mathcal{X}}}(\ddot{X}_i, \cdot)$. Here $\ddot{X}_i \equiv (X_i, Z_i)$. Note that $\tilde{\rho^2}$ is always nonnegative, but it is not guaranteed to be less than or equal to 1. In practice, since we know that $\rho^2 \in [0,1]$, we can always truncate $\tilde{\rho^2}$ at 1 when it exceeds 1. Note that as opposed to the graph-based estimator $\hat{\rho^2}$, $\tilde{\rho^2}$ is always nonrandom. For example, it does not involve tie-breakings for the K-NN graphs. Although the expression for $\tilde{\rho^2}$ in (18) looks complicated, it can be simplified considerably; see Proposition 5 below (and Appendix B.10 for a proof). Before we describe the result, let us introduce some notation. We denote by K_X, K_Y and $K_{\ddot{X}}$ the $n \times n$ kernel matrices, where for $i, j \in \{1, \ldots, n\}$, $(K_X)_{ij} = k_{\mathcal{X}}(X_i, X_j)$, $(K_Y)_{ij} = k_{\mathcal{Y}}(Y_i, Y_j)$, $(K_{\ddot{X}})_{ij} = k_{\ddot{\mathcal{X}}}(\ddot{X}_i, \ddot{X}_j)$. Let $H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ be the centering matrix (here $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$ and $I \equiv I_n$ denotes the $n \times n$ identity matrix). Then, $\tilde{K}_X := HK_XH$, $\tilde{K}_Y := HK_YH$, $\tilde{K}_{\ddot{X}} := HK_{\ddot{X}}H$ are the corresponding centered kernel matrices.

Proposition 5 *Fix $\varepsilon > 0$. Let*

$$M := \tilde{K}_X(\tilde{K}_X + n\varepsilon I)^{-1} - \tilde{K}_{\ddot{X}}(\tilde{K}_{\ddot{X}} + n\varepsilon I)^{-1} = n\varepsilon\left((\tilde{K}_{\ddot{X}} + n\varepsilon I)^{-1} - (\tilde{K}_X + n\varepsilon I)^{-1}\right),$$
$$N := I - \tilde{K}_X(\tilde{K}_X + n\varepsilon I)^{-1} = n\varepsilon(\tilde{K}_X + n\varepsilon I)^{-1}.$$

Then, $\tilde{\rho^2}$ in (18) can be expressed as

$$\tilde{\rho^2} = \frac{\mathrm{Tr}(M^\top \tilde{K}_Y M)}{\mathrm{Tr}(N^\top \tilde{K}_Y N)}, \tag{20}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix.

A few remarks are now in order.

Remark 10 (Kernel measure of association) *If X is not present, then $\rho^2(Y, Z|\emptyset)$ yields a measure of association between the two variables Y and Z. Our estimation strategy readily yields an empirical estimator of $\rho^2(Y, Z|\emptyset)$, namely,*

$$\tilde{\rho^2}(Y, Z|\emptyset) = \frac{\mathrm{Tr}(M^\top \tilde{K}_Y M)}{\mathrm{Tr}(N^\top \tilde{K}_Y N)}, \tag{21}$$

where $M = \tilde{K}_Z\left(\tilde{K}_Z + n\varepsilon I\right)^{-1}$, $N = I$. This estimator can be viewed as the kernel analogue to the graph-based estimator employed in Deb et al. (2020) to approximate $\rho^2(Y, Z|\emptyset)$.

Remark 11 (Uncentered estimates of CMEs) *Instead of using the centered estimates of CMEs, as in (15), to approximate $\tilde{\rho^2}$, as in (18), one could use their uncentered analogues; see Appendix A.2 for a discussion on this where an explicit expression for the corresponding 'uncentered' estimator $\tilde{\rho^2_u}$ of ρ^2 is derived. In Proposition 7 (in Appendix A.2) we further show that $\tilde{\rho^2_u}$ has an interesting connection to kernel ridge regression.*

Remark 12 (Approximate computation of $\tilde{\rho^2}$) *The exact computation of $\tilde{\rho^2}$ costs $O(n^3)$ time as we will have to invert $n \times n$ matrices (see Equation 20). A fast approximation of $\tilde{\rho^2}$ can be done using the method of incomplete Cholesky decomposition (Bach and Jordan, 2003). In particular, if we use incomplete Cholesky decomposition of all the three kernel matrices—K_X, K_Y and $K_{\ddot{X}}$—with ranks less than (or equal to) r, then the desired approximation of $\tilde{\rho^2}$ can be computed in time $O(nr^2)$; see Appendix A.4 for the details.*

An interesting property of $\tilde{\rho^2}$ is that it reduces to the empirical classical partial correlation squared when linear kernels are used and $\varepsilon \to 0$; this is stated in Proposition 6 (see Appendix B.12 for a proof). Thus, $\tilde{\rho^2}$ can indeed be seen as a natural generalization of squared partial correlation.

Proposition 6 *Suppose $\mathcal{Y} = \mathcal{Z} = \mathbb{R}$, $\mathcal{X} = \mathbb{R}^d$, with linear kernels used for all the three spaces $\mathcal{Y}, \mathcal{X}$, and $\ddot{\mathcal{X}}$. If X and $\ddot{X}$ have nonsingular sample covariance matrices, then $\tilde{\rho^2}$ reduces to the classical empirical partial correlation squared as $\varepsilon \to 0$, i.e.,*

$$\lim_{\varepsilon \to 0} \tilde{\rho^2}(Y, Z|X) = \hat{\rho^2}_{YZ \cdot X}.$$

4.2 Consistency Results

We first state a result that shows the consistency of the CME estimator $\hat{\mu}_{Y|\cdot}$ in (19). In particular, we show in Theorem 5 (see Appendix B.13 for a proof) that $\hat{\mu}_{Y|\cdot}$ is consistent in estimating $\mu_{Y|\cdot}$ in the averaged $\|\cdot\|^2_{\mathcal{H}_\mathcal{Y}}$-loss. This answers an open question mentioned in Klebanov et al. (2020, Section 8) and may be of independent interest.

Theorem 5 *Suppose the CME formula* (15) *holds, and the regularization parameter $\varepsilon \equiv \varepsilon_n \to 0^+$ (as $n \to \infty$) at a rate slower than $n^{-1/2}$ (i.e., $\varepsilon_n n^{1/2} \to \infty$). Then*

$$\frac{1}{n} \sum_{i=1}^{n} \|\hat{\mu}_{Y|X_i} - \mu_{Y|X_i}\|^2_{\mathcal{H}_\mathcal{Y}} \xrightarrow{p} 0.$$

As a consequence of Theorem 5, our RKHS-based estimator $\tilde{\rho^2}$ (see Equation 18) is consistent for estimating ρ^2 (as in Equation 1). This result is formally stated below in Theorem 6 and proved in Appendix B.13.

Theorem 6 *Suppose the CME formula* (15) *holds for both (Y, X) and $(Y, \ddot{X})$. Let the regularization parameter $\varepsilon \equiv \varepsilon_n \to 0^+$ (as $n \to \infty$) at a rate slower than $n^{-1/2}$. Then*

$$\tilde{\rho^2}(Y, Z|X) \xrightarrow{p} \rho^2(Y, Z|X).$$

Remark 13 *From the forms of $\mu_{Y|x}$ (see Equation 15) and $\hat{\mu}_{Y|x}$ (see Equation 19), one might conjecture whether $\hat{C}_{YX}(\hat{C}_X + \varepsilon_n I)^{-1}$ could converge to $(C_X^\dagger C_{XY})^*$ in some sense (e.g., in the Hilbert-Schmidt norm or operator norm), as was explored in Song et al. (2010b). However, such a convergence is rarely possible. In particular, it does not hold when $\mathcal{H}_\mathcal{X}$ is infinite-dimensional*[9].

9. Consider $Y = X$. Then $\|(C_X^\dagger C_{XY})^* - \hat{C}_{YX}(\hat{C}_X + \varepsilon_n I)^{-1}\|_{\text{op}} = \|I - \hat{C}_X(\hat{C}_X + \varepsilon_n I)^{-1}\|_{\text{op}} = \|\varepsilon_n(\hat{C}_X + \varepsilon_n I)^{-1}\|_{\text{op}} \equiv 1$. The last equality follows as: (i) $\hat{C}_X$ is finite-rank (recall that a finite-rank operator is a bounded linear operator between Banach spaces whose range is finite-dimensional) having at least one zero eigenvalue; (ii) any eigenvalue of $\varepsilon_n(\hat{C}_X + \varepsilon_n I)^{-1}$ has the form $\frac{\varepsilon_n}{\lambda + \varepsilon_n}$, where λ is an eigenvalue of $\hat{C}_X$; and (iii) taking $\lambda = 0$ yields the desired conclusion. Consequently the convergence of Hilbert-Schmidt norm is also impossible as $\|\cdot\|_{\text{op}} \leq \|\cdot\|_{\text{HS}}$.

5. Variable Selection Using KPC

Suppose that we have a regression problem with p predictor variables $X_1, \ldots, X_p$ and a response variable Y. Here the response $Y \in \mathcal{Y}$ is allowed to be continuous/categorical, multivariate and even non-Euclidean (Ramsay and Silverman, 2002; Fukumizu et al., 2008, 2009b; Danafar et al., 2010; Tsagris, 2015; Hron et al., 2016; Petersen and Müller, 2016, 2019), as long as a kernel function can be defined on $\mathcal{Y} \times \mathcal{Y}$. Similarly, the predictors $X_1, \ldots, X_p$ could also be non-Euclidean; we just want each X_i to take values in some metric space $\mathcal{X}_i$. In regression the goal is to study the effect of the predictors on the response Y. We can postulate the following general model:

$$Y = f(X_1, \ldots, X_p, \epsilon) \tag{22}$$

where ϵ (the unobserved error) is independent of $(X_1, \ldots, X_p)$ and f is an unknown function.

The problem of *variable selection* is to select a subset of predictive variables from $X_1, \ldots, X_p$ to explain the response Y in the simplest possible way. For $S \subset \{1, \ldots, p\}$, let us write $X_S := (X_j)_{j \in S}$. Our goal is to find an $S \subset \{1, \ldots, p\}$ such that

$$Y \perp\!\!\!\perp X_{S^c} | X_S. \tag{23}$$

Such an S (satisfying 23) is called a *sufficient subset* (Vergara and Estévez, 2014; Azadkia and Chatterjee, 2021). Ideally, we would want to select a sufficient subset S that has the smallest cardinality, so that we can write $Y = \mathbb{E}[f(X_1, \ldots, X_p, \epsilon) | X_S, \epsilon] =: g(X_S, \epsilon)$ in (22).

The main idea is to use our proposed KPC ρ^2 to detect conditional independence in (23) (note that ρ^2 is 0 if and only if conditional independence holds). In the following two subsections we propose two model-free variable selection algorithms—one based on our graph-based estimator $\hat{\rho^2}$ and the other on the RKHS-based framework $\tilde{\rho^2}$. Our procedures do not make any parametric model assumptions, are easily implementable and have strong theoretical guarantees. They provide a more general framework for feature selection (when compared to Azadkia and Chatterjee, 2021) that can handle any kernel function $k_{\mathcal{Y}}(\cdot, \cdot)$ any geometric graph (including K-NN graphs for any $K \geq 1$). Further, we only require $\mathcal{Y}$ to be kernel-endowed, and the predictor variables X_i's can take values in metric spaces (for our graph-based estimators) and general kernel-endowed spaces (for our RKHS-based estimators). These flexibilities indeed yield more powerful variable selection algorithms, having better finite sample performance (see Section 6.1.2).

5.1 Variable Selection with Graph-Based Estimator (KFOCI)

We introduce below the algorithm *Kernel Feature Ordering by Conditional Independence* (KFOCI) which is a model-free forward stepwise variable selection algorithm. The proposed algorithm has an automatic stopping criterion and yields a provably consistently variable selection method (i.e., it selects a sufficient subset of predictors with high probability) even in the high-dimensional regime under suitable assumptions; see Theorem 7 below.

Let us describe the algorithm KFOCI. Suppose that predictors $X_{j_1}, \ldots, X_{j_k}$, for $k \geq 0$, have already been selected by KFOCI. Quite naturally, we would like to find $X_{j_{k+1}}$ that maximizes $\rho^2(Y, X_{j_{k+1}} | X_{j_1}, \ldots, X_{j_k})$. To this end define

$$T(S) := \mathbb{E}[\mathbb{E}[k_{\mathcal{Y}}(Y, Y') | X_S]], \tag{24}$$

where we first draw X_S, and then draw Y, Y' i.i.d. from the conditional distribution of Y given X_S. A closer look of the expression of ρ^2 in (5) reveals that finding $X_{j_{k+1}}$ that maximizes $\rho^2(Y, X_{j_{k+1}}|X_{j_1}, \ldots, X_{j_k})$ is equivalent to finding $X_{j_{k+1}}$ that maximizes $T(\{X_1, \ldots, X_{j_{k+1}}\})$ in (24), for $j_{k+1} \in \{1, \ldots, p\} \setminus \{j_1, \ldots, j_k\}$. Note that $T(S)$ satisfies

$$T(S') \geq T(S) \qquad \text{whenever} \qquad S' \supset S,$$

since the numerator of $\rho^2(Y, X_{S' \setminus S}|X_S)$ in (5) is always greater than (or equal to) 0. If S_0 is a sufficient subset, then $T(S_0) = T(\{1, \ldots, p\}) \geq T(S)$, for all $S \subset \{1, \ldots, p\}$. Therefore, $T(S)$ can be viewed as measuring the importance of S in predicting Y.

For our implementation, we propose the use of the estimator T_n (in 11) instead of the unknown T (in 24). Note that Theorem 2 shows that $T_n(S) \equiv T_n(Y, X_S)$ is a consistent estimator of $T(S)$, for every $S \subset \{1, \ldots, p\}$. What is even more interesting is that the use of $T_n(\cdot)$ automatically yields a stopping rule—we stop our algorithm when adding any variable does not increase our objective function, i.e., $T_n(\hat{S} \cup \{\ell\}) < T_n(\hat{S})$, for any $\ell \in \{1, \ldots, p\} \setminus \hat{S}$ where $\hat{S}$ is the current sufficient subset. Algorithm 1 gives the pseudocode.

Algorithm 1: KFOCI: a forward stepwise variable selection algorithm

Data: $(Y_i, X_{1i}, \ldots, X_{pi})$, for $i = 1, \ldots, n$

Initialization: $k = -1$, $\hat{S} \leftarrow \emptyset$, $\{j_0\} = \emptyset$, $T_n(\emptyset) = -\infty$;

do

1. $k \leftarrow k + 1$;

2. $\hat{S} \leftarrow \hat{S} \cup \{j_k\}$;

3. Choose $j_{k+1} \in \{1, \ldots, p\} \setminus \hat{S}$ such that $T_n(\hat{S} \cup \{j_{k+1}\})$ is *maximized*, i.e.,

$$j_{k+1} := \arg \max_{\ell \in \{1, \ldots, p\} \setminus \hat{S}} T_n(\{j_1, \ldots, j_k, \ell\});$$

while $T_n(\hat{S} \cup \{j_{k+1}\}) \geq T_n(\hat{S})$ and $k < p$;

Output: $\hat{S}$

At each step, we are actually selecting $X_{j_{k+1}}$ to maximize $\hat{\rho}^2(Y, X_{j_{k+1}}|X_{j_1}, \ldots, X_{j_k})$, and the stopping criterion corresponds to the case when $\hat{\rho}^2(Y, X_{j_{k+1}}|X_{j_1}, \ldots, X_{j_k}) < 0$ for all $j_{k+1} \in \{1, \ldots, p\} \setminus \{j_1, \ldots, j_k\}$. Our method therefore has substantial difference to marginal screening methods. $X_{j_{k+1}}$ will be selected only if it contains information that has not been explained by $X_1, \ldots, X_{j_k}$, instead of being marginally strongly correlated to the response. The following result shows the variable selection consistency of KFOCI.

Theorem 7 *Suppose the following assumptions hold:*

(a) There exists $\delta > 0$ such that for any insufficient subset $S \subset \{1, \ldots, p\}$, there is some j such that $T(S \cup \{j\}) \geq T(S) + \delta$.

(b) Suppose that kernel satisfies $\sup_{y \in \mathcal{Y}} k_{\mathcal{Y}}(y, y) \leq M < \infty$. Let $\kappa := \lfloor \frac{M}{\delta} + 1 \rfloor$.

(c) Suppose that the K_n-NN graph is used as the geometric graph in (11) (with $K_n \leq C_6(\log n)^\gamma$ for some $C_6 > 0$, $\gamma \geq 0$). For every $S \subset \{1, \ldots, p\}$ of size less than or equal to κ, we suppose that Assumptions 4-8 hold with X replaced by X_S, with the same constants $d, \{C_i\}_{i=1}^5, \alpha, \beta_1, \beta_2$ (uniformly over all subsets S) in Assumptions 5-8.

Then there exist $L_1, L_2 > 0$ *depending only on* $\alpha, \beta_1, \beta_2, \gamma, \{C_i\}_{i=1}^6, d, M, \delta$ *such that*

$$\mathbb{P}(\hat{S} \text{ is sufficient}) \geq 1 - L_1 p^{\kappa} e^{-L_2 n}.$$

See Section B.14 for a proof. Suppose the above algorithm selects $\hat{S}$. Then Theorem 7 shows that if $n \gg \log p$, then the algorithm selects a sufficient subset with high probability. In particular, in the low dimensional setting (where p is fixed), the algorithm selects a sufficient subset with probability that goes to 1 exponentially fast. Theorem 7 is in the same spirit as Azadkia and Chatterjee (2021, Theorem 6.1), but allows for the predictors and response variable to be metric-space valued, and offers the flexibility of using any kernel and a general K-NN graph functional (note that the FOCI algorithm in Azadkia and Chatterjee, 2021 used the 1-NN graph). This flexibility leads to better finite sample performance for KFOCI, when compared with FOCI (Azadkia and Chatterjee, 2021). Even in parametric settings, our performance is comparable to (and sometimes even better than) classical methods such as the Lasso (Tibshirani, 1996) and the Dantzig selector (Candès and Tao, 2007); see Section 6.1.2 for the detailed simulation studies.

Algorithm 2: Forward stepwise variable selection algorithm using $\tilde{\rho}^2$

Input: The number of variables $p_0 \leq p$ to select; a kernel function $k_{\mathcal{Y}}(\cdot,\cdot)$ on $\mathcal{Y} \times \mathcal{Y}$; a kernel k_S for each X_S, $S \subset \{1,\ldots,p\}$ with $|S| \leq p_0$, and regularization parameter $\varepsilon > 0$ for computing $\tilde{\rho}^2$.

Data: $(Y_i, X_{1i}, \ldots, X_{pi})$, for $i = 1, \ldots, n$

Initialization: $k = -1$, $\tilde{S} \leftarrow \emptyset$, $\{j_0\} = \emptyset$;

do

1. $k \leftarrow k + 1$;

2. $\tilde{S} \leftarrow \tilde{S} \cup \{j_k\}$;

3. Choose the next $j_{k+1} \in \{1,\ldots,p\} \setminus \tilde{S}$ such that $\tilde{\rho}^2(Y, X_{j_{k+1}} | \tilde{S})$ is *maximized*, i.e.,

$$j_{k+1} := \arg \max_{\ell \in \{1,\ldots,p\} \setminus \tilde{S}} \tilde{\rho}^2(Y, X_{j_{k+1}} | X_{j_1}, \ldots, X_{j_k});$$

while $k < p_0$;

Output: $\tilde{S}$

Remark 14 (On our assumptions) *(a) is essentially a sparsity assumption, which is also assumed in Azadkia and Chatterjee (2021, Theorem 6.1). Note that as the kernel* $k_{\mathcal{Y}}$ *is bounded by* M *(by Assumption b),* $T(S) = \mathbb{E}[\mathbb{E}[k_{\mathcal{Y}}(Y, Y')|X_S]]$ *is also bounded by* M*. Thus (a) implies that there exists a sufficient subset of size less than* $\lfloor \frac{M}{\delta} + 1 \rfloor$*. Another implication of (a) is a lower bound assumption on the signal strength of important predictors, indicating that we can expect an improvement of* $\delta > 0$ *in terms of* $T(\cdot)$ *for each iteration of KFOCI. Condition (c) can be viewed as a uniform version of Assumptions 4-8, in the sense that those assumptions need to hold with the same constants uniformly over all subsets of* $\{1,\ldots,p\}$ *with cardinality no larger than* κ*. Similar to Remark 8 Assumption (c) is less stringent when compared to the assumptions made in Azadkia and Chatterjee (2021, Theorem 6.1)*

in the sense that it allows for: (i) any general metric space $\mathcal{X}_S$ (recall X_S takes values in $\mathcal{X}_S$), (ii) tail decay rates of X_S slower than sub-exponential, and (iii) β_2 to vary in $(0, 1]$.

5.2 Variable Selection Using the RKHS-Based Estimator

As in Section 5.1, we can also develop a similar model-free forward stepwise variable selection algorithm using the RKHS-based estimator $\tilde{\rho^2}$ (instead of $\hat{\rho^2}$). Algorithm 2 gives the pseudocode of the proposed procedure. As $\tilde{\rho^2}$ is always nonnegative (see Equation 18 and 20) one can no longer specify an automatic stopping criterion as in Algorithm 1. Thus, in Algorithm 2 we have to prespecify the number of variables $p_0 \leq p$ to be chosen a priori. Note that to use Algorithm 2 one must also specify a kernel function k_S, for every $S \subset \{1, \ldots, p\}$ with cardinality $|S| \leq p_0$. For the most common case where all the X_i's are real-valued and are suitably normalized (e.g., all X_i's have mean 0 and variance 1), an automatic choice of k_S can be the Gaussian kernel with empirically chosen bandwidth[10]. In our numerical studies we see that Algorithm 2 can also detect complex nonlinear conditional dependencies and has good finite sample performance; see Section 6.1.2 for the details.

6. Finite Sample Performance of Our Methods

In this section, we report the finite sample performance of $\hat{\rho^2}$, $\tilde{\rho^2}$ and the related variable selection algorithms, on both simulated and real data examples. We consider both Euclidean and non-Euclidean responses Y in our examples. Even when restricted to Euclidean settings, our algorithms achieve superior performance compared to existing methods. All results are reproducible using our `R` package `KPC` (Huang, 2021) available on CRAN.

6.1 Examples with Simulated Data

6.1.1 Consistency of $\hat{\rho^2}$ and $\tilde{\rho^2}$

Here we examine the consistency of our two empirical estimators $\hat{\rho^2}$ and $\tilde{\rho^2}$. As has been mentioned earlier, the consistency of $\hat{\rho^2}$ requires only very weak moment assumptions on the kernel (see Theorem 3), whereas the consistency of $\tilde{\rho^2}$ depends on the validity of the CME formula (in Equation 15) which in turn depends on the hard to verify Assumption 9. We first restrict ourselves to the Euclidean setting and consider the following models:

- **Model I**: X, Z *i.i.d.* $N(0,1)$, $Y = X + Z + N(1,1)$.
- **Model II**: X, Z *i.i.d.* $N(0,1)$, $Y \sim \text{Bernoulli}(e^{-Z^2/2})$.
- **Model III**: X, Z *i.i.d.* $\text{Uniform}[0,1]$, $Y = X + Z \pmod 1$.

Model I: Here we let $k, k_{\mathcal{X}}$ and $k_{\ddot{\mathcal{X}}}$ be linear kernels. As we are in a Gaussian setting, both our estimators $\hat{\rho^2}$ and $\tilde{\rho^2}$ are consistent[11]. One can check that $\rho^2(Y, Z|X) = 0.5$. For the graph-based estimator $\hat{\rho^2}$, we use directed 1-NN, 2-NN graphs and MST. For $\tilde{\rho^2}$, we set

10. For example, for $x, x' \in \mathbb{R}^{|S|}$, $k_S(x, x') := \exp\left(-\|x - x'\|^2/(2s^2)\right)$, where s is the median of pairwise distances $\{\|(X_S)_i - (X_S)_j\|\}_{i<j}$.

11. Note that the assumptions in Theorem 3 hold in this setting and thus the graph-based estimator $\hat{\rho^2}$ is consistent. Further, Assumption 9 holds which implies that $\tilde{\rho^2}$ is also consistent (by Theorem 6). To check that Assumption 9 holds, we notice that the RKHS associated with the linear kernel $k_{\mathcal{X}}$ is $\mathcal{H}_{\mathcal{X}} = \{f(x) = a^\top x | a \in \mathcal{X}\}$—the space of all linear functions on $\mathcal{X} = \mathbb{R}$. So $\mathbb{E}[a^\top Y|X] = a^\top(X + 1) \in \mathcal{H}_{\mathcal{X}} + \mathbb{R}$. Note that Assumption 9 also holds for $(Y, \ddot{X})$ by the same argument. For the RKHS-based estimator

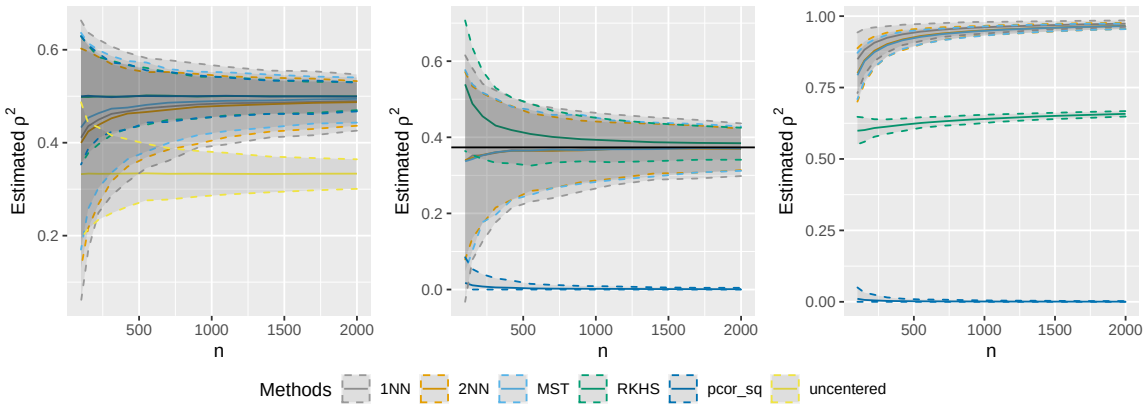

Figure 1: The performance of different estimators of ρ^2 as the sample increases. The solid lines show the mean for each estimator; the dashed lines are the corresponding 2.5% and 97.5% quantiles. Here 'MST', '1NN', '2NN' denote the graph-based estimators constructed using MST, 1-NN, 2-NN graphs respectively, 'RKHS' denotes $\tilde{\rho^2}$, 'pcor_sq' denotes the squared partial correlation, and 'uncentered' denotes the RKHS-based estimator constructed using the uncentered CME. The left, middle and right panels correspond to models I, II, and III respectively.

$\varepsilon_n = 10^{-3} \cdot n^{-0.4}$ for all the three models considered here (which satisfies the condition for consistency in Theorem 6). Although the linear kernel is not characteristic, as (Y, Z, X) is jointly normal, all the desired properties of $\rho^2(Y, Z|X)$ in Theorem 1 hold (see Remark 6); and ρ^2 is equal to the squared partial correlation coefficient (see Proposition 1-d). The left panel of Figure 1 shows, for different sample sizes n, the mean and 2.5%, 97.5% quantiles of the various estimators of ρ^2 (obtained from 1000 replications). It can be seen that all the estimators except the RKHS-based estimator using the uncentered CME formula (see Remark 11; also see Appendix A.2) are consistent, converging to the true value $\rho^2 = 0.5$. As expected, $\hat{\rho^2}$ constructed from the 1-NN graph has less bias but higher variance compared to $\hat{\rho^2}$ constructed using the 2-NN graph. Note that $\tilde{\rho^2}$ achieves almost the same statistical performance as the squared partial correlation coefficient; a consequence of Proposition 6. As model I describes a Gaussian setting, the classical partial correlation coefficient (and $\tilde{\rho^2}$) has the best performance. Notice that $\hat{\rho^2}$, in spite of being fully nonparametric, also achieves good performance.

Model II: We let $k_{\mathcal{Y}}(y_1, y_2) := \mathbf{1}\{y_1 = y_2\}$ be the discrete kernel. In this case, $\rho^2 = \frac{2\sqrt{6}+2\sqrt{3}-3\sqrt{2}-3}{3} \approx 0.37$. As the kernel $k_{\mathcal{Y}}(\cdot, \cdot)$ is bounded, $\hat{\rho^2}$ is automatically consistent (see Theorem 3). To compute the RKHS-based estimator $\tilde{\rho^2}$ we take $k_{\mathcal{X}}(x_1, x_2) = e^{-|x_1-x_2|^2/2}$ (a Gaussian kernel) and $k_{\ddot{\mathcal{X}}}((x_1, z_1), (x_2, z_2)) = \left(e^{-|x_1-x_2|^2/2} + 1\right) e^{-|z_1-z_2|^2/2}$ (a product kernel). One can also check that Assumption 9 holds in this case, and thus $\tilde{\rho^2}$ is consistent (by Theorem 6). The behavior of all the estimators, as the sample size in-

using the uncentered CME (see Remark 11), the analogous sufficient condition (like Assumption 9; see Remark 19) does not hold, since for $a \neq 0$, $a^\top x + a \notin \mathcal{H}_{\mathcal{X}}$.

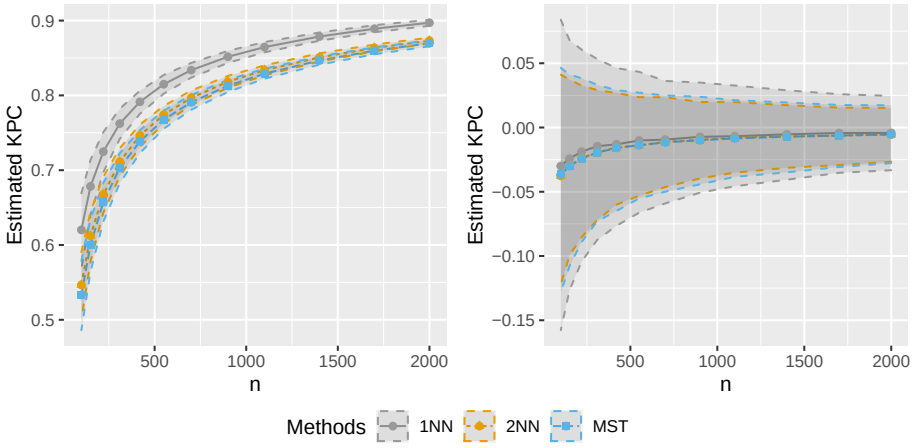

Figure 2: The performance of $\hat{\rho}^2$ (constructed using 1-NN, 2-NN graphs and MST) as sample increases. The solid lines show the mean and the confidence bands show the corresponding 2.5% and 97.5% quantiles as a function of the sample size n. The left (resp. right) panel corresponds to model IV (resp. V).

creases, is shown in the middle panel of Figure 1. Note that the classical partial correlation coefficient fails to capture the conditional dependence and is almost 0.

Model III: We let k, $k_{\mathcal{X}}$, and $k_{\ddot{\mathcal{X}}}$ be Gaussian kernels (see Equation 1) with different bandwidths[12]. This model has also been considered in Azadkia and Chatterjee (2021). Note that here $\rho^2(Y, Z|X) = 1$ since Y is a measurable function of X and Z. Here $\hat{\rho}^2$ (constructed using 1-NN, 2-NN graphs and MST) is consistent as the Gaussian kernel is bounded (which is a sufficient condition for Theorem 3 to hold). From the right panel of Figure 1 we see that both of the graph-based estimators are very close to 1. Assumption 9 for the CME formula holds with (Y, X) (since Y and X are independent) but it does not hold for $(Y, \ddot{X})$[13]. Therefore, Theorem 6 does not hold and $\tilde{\rho}^2$ cannot be guaranteed to be consistent. As can be seen from the right panel of Figure 1, $\tilde{\rho}^2$ does not seem to converge to 1. However, compared to the classical partial correlation, which is almost 0, both $\hat{\rho}^2$ and $\tilde{\rho}^2$ provide evidence that Y is highly dependent on Z, conditional on X.

A non-Euclidean example: Next, we consider the case where $\mathcal{Y}$ is the *special orthogonal group* SO(3), the space consisting of 3×3 orthogonal matrices with determinant 1. SO(3) has been used to characterize the rotation of tectonic plates in geophysics (Hanna and Chang, 2000) as well as in the studies of human kinematics and robotics (Stavdahl et al., 2005). We use the following characteristic kernel on SO(3) (Fukumizu et al., 2009b):

$$k_{\mathcal{Y}}(A, B) := \frac{\pi\theta(\pi - \theta)}{8\sin(\theta)}, \tag{25}$$

12. Here $k_{\mathcal{Y}}(\cdot,\cdot) = k_{\mathcal{X}}(\cdot,\cdot) = \exp(-5|\cdot - \cdot|^2)$ and $k_{\ddot{\mathcal{X}}}(\cdot,\cdot) = \exp(-2\|\cdot - \cdot\|^2)$. These bandwidths are just arbitrary choices that approximately fit to the scale of the data.

13. Note that as $k_{\ddot{\mathcal{X}}}$ is continuous on $\ddot{\mathcal{X}} \times \ddot{\mathcal{X}}$ with $\ddot{\mathcal{X}}$ compact, all the functions in $\mathcal{H}_{\ddot{\mathcal{X}}}$ are continuous (Cucker and Smale, 2002, Chapter 3, Theorem 2). But for any $g \in \mathcal{H}_{\mathcal{Y}}$, and $h \in \mathcal{H}_{\ddot{\mathcal{X}}}$ (which is continuous), we have $\mathbb{E}[g(Y)|\ddot{X} = \ddot{x}] - h(\ddot{x}) = g(Y(\ddot{x})) - h(\ddot{x})$ which is discontinuous and cannot be a constant $P_{\ddot{X}}$-a.e.

Table 1: $\tilde{\rho^2}$ for $n = 1000$ observations with different ε's.

Estimands	Estimators	$\varepsilon = 10^{-3}$	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
$\rho^2(Y_1, Z\|X) = 1$	$\tilde{\rho^2}(Y_1, Z\|X)$	0.427	0.548	0.620	0.665	0.697	0.724	0.747
$\rho^2(Y_2, Z\|X) = 0$	$\tilde{\rho^2}(Y_2, Z\|X)$	0.027	0.038	0.052	0.067	0.080	0.093	0.106

where $e^{\pm\sqrt{-1}\theta}$ $(0 \le \theta \le \pi)$ are the eigenvalues of $B^{-1}A$, i.e., $\cos\theta = \frac{\mathrm{Tr}(B^{-1}A)-1}{2}$. Define the rotation around x- and z-axis as $R_1, R_3 : \mathbb{R} \to \mathrm{SO}(3)$, defined by (for $x, z \in \mathbb{R}$)

$$R_1(x) := \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(x) & -\sin(x) \\ 0 & \sin(x) & \cos(x) \end{pmatrix}, \qquad R_3(z) := \begin{pmatrix} \cos(z) & -\sin(z) & 0 \\ \sin(z) & \cos(z) & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Let $X, Z \overset{i.i.d.}{\sim} N(0,1)$. Consider the two models:

- **Model IV**: $Y_1 = R_1(X)R_3(Z)$. Here Y_1 is a function of X and Z. $\rho^2(Y_1, Z|X) = 1$.
- **Model V**: $Y_2 = R_1(X)R_3(\varepsilon)$, for an independent $\varepsilon \sim N(0,1)$. In this model, $Y_2 \perp\!\!\!\perp Z|X$, and thus $\rho^2(Y_2, Z|X) = 0$.

Figure 2 shows the means and 95% confidence bands for $\hat{\rho^2}(Y_1, Z|X)$ and $\hat{\rho^2}(Y_2, Z|X)$ constructed using 1-NN, 2-NN graphs and MST from 5000 replications. It can be seen that as n increases, $\hat{\rho^2}(Y_1, Z|X)$ gets very close to 1 (which provides evidence that Y_1 is a function of Z given X), and $\hat{\rho^2}(Y_2, Z|X)$ comes very close to 0 (suggesting that Y_2 is conditionally independent of Z given X).

We did not plot $\tilde{\rho^2}$ as it is not consistent and also quite sensitive to the choice of the regularization parameter ε; see Table 1 where we report $\tilde{\rho^2}$ for models IV and V when $n = 1000$, $k_{\mathcal{X}}(x, x') = \exp(-|x - x'|^2)$, and $k_{\ddot{\mathcal{X}}}(x, x') = \exp(-\|x - x'\|^2/2)$. However, $\tilde{\rho^2}(Y_1, Z|X)$ is always much larger than $\tilde{\rho^2}(Y_2, Z|X)$, indicating the conditional association between Y_1 and Z is much stronger than that between Y_2 and Z when controlling for X.

Summary: In all the simulation examples we see that the graph-based estimator $\hat{\rho^2}$ has very good performance. It is able to capture different kinds of nonlinear conditional dependencies, under minimal assumptions. The RKHS-based estimator $\tilde{\rho^2}$ also performs quite well, although it need not be consistent always (as Assumption 9 may not hold in certain applications). Further, as both $\hat{\rho^2}$ and $\tilde{\rho^2}$ can handle non-Euclidean responses, we believe that they are useful tools for detecting conditional dependencies in any application.

6.1.2 Variable Selection

In this subsection we examine the performance of our proposed variable selection procedures—KFOCI (Algorithm 1) and Algorithm 2—in a variety of settings. Our examples include both low-dimensional and high-dimensional models. Note that KFOCI can automatically determine the number of variables to select, while Algorithm 2 requires prespecifying the number of variables to be chosen.

We consider the following models with $X = (X_1, \ldots, X_p) \sim N(0, I_p) \in \mathbb{R}^p$:

- LM (linear model): $Y = 3X_1 + 2X_2 - X_3 + N(0,1)$.
- GAM (generalized additive model): $Y = \sin(X_1) + 2\cos(X_2) + e^{X_3} + N(0,1)$.
- Nonlin1 (nonlinear model in Azadkia and Chatterjee, 2021): $Y = X_1X_2 + \sin(X_1X_3)$.

- Nonlin2 (heavy-tailed): $Y = \frac{2\log(X_1^2+X_2^4)}{\cos(X_1)+\sin(X_3)} + \epsilon$ where $\epsilon \sim t_1$, the t-distribution with 1 degree of freedom.
- Nonlin3 (non-additive noise): $Y = |X_1 + U|^{\sin(X_2 - X_3)}$, where $U \sim \text{Uniform}[0, 1]$.
- SO(3) (non-Euclidean response): $Y = R_1(X_1)R_3(X_2X_3) \in \text{SO}(3)$.

In all the examples, the noise variable is assumed to be independent of X. The above models cover different kinds of linear and nonlinear relationships between Y and X.

Low dimensional setting: We first consider the low-dimensional setting $n = 200$, $p = 10$ and compare the performance of KFOCI with other competing methods which themselves determine the number of variables to select. We implement the KFOCI algorithm with the directed 1-NN, 2-NN, 10-NN graphs and MST using Algorithm 1 (denoted by '1-NN', '2-NN', '10-NN', 'MST' in Table 2). For all the models except when the response takes values in SO(3), we use the Gaussian kernel with empirically chosen bandwidth, i.e., $k_{\mathcal{Y}}(y, y') = \exp\left(\frac{-|y-y'|^2}{2s^2}\right)$, where s is the median of pairwise distances $\{|y_i - y_j|\}_{i<j}$. When $\mathcal{Y} = \text{SO}(3)$, we use the kernel in (25). A natural competitor of KFOCI is FOCI (Azadkia and Chatterjee, 2021), implemented in the `R` package `FOCI` (Azadkia et al., 2020). We also compare KFOCI with 'ols' which is the forward stepwise variable selection algorithm in linear regression (where a variable with the smallest p-value less than 0.01 enters the model at every stage), implemented using the function `ols_step_forward_p` in the `R` package `olsrr` (Hebbali, 2018). We also consider 'VSURF', variable selection using random forests, implemented using the `R` package `VSURF`[14] (Genuer et al., 2019). Note that for all the considered models, X_1, X_2, X_3 are the "correct" variables. In the first tabula of Table 2 we report: (i) The proportion of times $\{X_1, X_2, X_3\}$ is exactly selected, (ii) the proportion of times $\{X_1, X_2, X_3\}$ is selected with possibly other variables, and (iii) the average number of variables selected, by the different methods in 100 replications. It can be seen from Table 2 that KFOCI achieves the best performance in all the nonlinear settings considered; in particular, KFOCI with the 10-NN graph selects exactly $\{X_1, X_2, X_3\}$ more than 90% of the times in all the nonlinear examples and also has good performance in the linear setting. Although FOCI uses the 1-NN graph in its algorithm, it has inferior performance compared to KFOCI with 1-NN; this indicates that the Gaussian kernel may be better at detecting various conditional associations than the special kernel used in FOCI (see Lemma 3). MST achieves similar (often slightly worse) performance as 2-NN, and their performance is between 1-NN and 10-NN. Note that the choice of the number of neighbors $K = 10$ has not been optimized, which can be seen from Figure 3, where the proportion of exact selection and the proportion of selecting $\{X_1, X_2, X_3\}$ are plotted against different K and n under the above setting. We believe that the optimal value of K should grow sub-linearly with n. But for samples with up to a few hundred observations, $K \approx 5\% \cdot n$ may be a good heuristic choice, as marked by the grey lines in Figure 3.

Next, we consider the case where the number of variables to select is set by the oracle as 3. For KFOCI, we still use 1,2,10-NN graphs and MST as before, but without imposing the automatic stopping criterion. For Algorithm 2 denoted by 'KPC (RKHS)', we set the kernel on $\mathcal{Y}$ as the same kernel for the methods '1-NN'/'10-NN'; the kernel on $\mathcal{X}_S$ is taken as $k_{\mathcal{X}_S}(x, x') = \exp\left(-\|x - x'\|^2_{\mathbb{R}^{|S|}}/|S|\right)$, and $\varepsilon = 10^{-3}$. We compare our methods with

14. For `VSURF`, we take the variables obtained at the "interpretation step", which aims to select all variables related to the response for interpretation purpose.

Table 2: Performance of the various variable selection algorithms in low-dimensional ($n = 200$, $p = 10$) and high-dimensional ($n = 200$, $p = 1000$) settings. In the case where the number of variable to select is unspecified, the reported numbers are: The proportion of times $\{X_1, X_2, X_3\}$ is exactly selected / the proportion of times $\{X_1, X_2, X_3\}$ is selected possibly with other variables / the average number of variables selected. In the case where the number of variables to be selected is set by the oracle as 3, the reported numbers are: The proportion of times $\{X_1, X_2, X_3\}$ is exactly selected / the average number of correct variables among the 3 selected variables (i.e. $|\hat{S} \cap \{X_1, X_2, X_3\}|$). The method with the best performance is highlighted in bold. '—' means the method cannot deal with responses in SO(3).

Low dimension, not specifying the number of variables to select

Models	LM	GAM	Nonlin1	Nonlin2	Nonlin3	SO3
1-NN	0.87/0.98/3.09	0.39/0.74/3.28	0.88/0.95/2.98	0.41/0.79/3.36	0.53/0.82/3.16	**1.00/1.00**/3.00
MST	0.96/0.97/2.99	0.43/0.79/2.99	0.96/0.99/3.01	0.52/0.84/3.23	0.70/0.89/3.11	0.99/0.99/2.98
2-NN	**0.99**/0.99/2.99	0.63/0.90/3.21	0.96/**1.00**/3.05	0.65/0.89/3.18	0.84/0.97/3.10	**1.00/1.00**/3.00
10-NN	0.81/0.81/2.81	**0.92/0.93**/2.94	**1.00/1.00**/3.00	**0.93/0.97**/3.01	**1.00/1.00**/3.00	0.97/0.97/2.94
FOCI	0.57/0.87/3.25	0.15/0.64/3.62	0.56/0.72/2.87	0.22/0.53/2.93	0.28/0.52/2.74	—
ols	0.95/**1.00**/3.05	0.03/0.04/2.03	0.00/0.00/1.06	0.00/0.00/1.04	0.00/0.00/1.07	—
VSURF	0.82/0.82/2.82	0.71/0.71/2.68	0.21/0.21/2.21	0.00/0.03/2.67	0.06/0.23/2.90	—

Low dimension, specifying 3 variables to select

Models	LM	GAM	Nonlin1	Nonlin2	Nonlin3	SO(3)
1-NN	**1.00/3.00**	0.78/2.78	0.88/2.86	0.64/2.54	0.68/2.57	**1.00/3.00**
MST	0.99/2.99	0.84/2.84	0.96/2.95	0.79/2.72	0.80/2.75	**1.00/3.00**
2-NN	**1.00/3.00**	0.92/2.92	0.96/2.95	0.78/2.75	0.90/2.89	**1.00/3.00**
10-NN	**1.00/3.00**	0.99/2.99	**1.00/3.00**	**1.00/3.00**	**1.00/3.00**	**1.00/3.00**
KPC (RKHS)	**1.00/3.00**	**1.00/3.00**	**1.00/3.00**	0.99/2.99	**1.00/3.00**	**1.00/3.00**
FOCI	0.93/2.93	0.54/2.52	0.60/2.24	0.40/2.18	0.46/1.93	—
FWDselect	**1.00/3.00**	**1.00/3.00**	0.05/1.67	0.02/1.15	0.10/1.69	—
varimp	**1.00/3.00**	0.97/2.97	0.66/2.66	0.00/1.11	0.47/2.33	—

High dimension, not specifying the number of variables to select

Models	LM	GAM	Nonlin1	Nonlin2	Nonlin3	SO(3)
1-NN	0.23/0.90/3.72	0.01/0.21/3.83	0.21/0.49/3.77	0.00/0.03/3.48	0.00/0.03/3.21	0.82/0.83/2.84
MST	0.41/0.93/3.51	0.01/0.47/4.00	0.55/0.73/3.57	0.02/0.15/4.65	0.03/0.20/4.05	**0.99**/**0.99**/2.99
2-NN	0.70/0.98/3.28	0.03/0.60/4.13	0.49/0.77/3.32	0.02/0.23/4.64	0.08/0.28/3.58	0.97/0.97/2.97
10-NN	**0.82**/0.82/2.82	**0.76**/**0.92**/3.14	**0.99**/**1.00**/3.01	**0.38**/**0.78**/8.82	**0.87**/**0.94**/3.50	0.95/0.95/2.90
FOCI	0.02/0.53/3.99	0.00/0.05/4.14	0.01/0.03/3.43	0.00/0.00/3.31	0.00/0.01/3.30	—
Lasso	0.66/**1.00**/4.19	0.00/0.00/1.17	0.00/0.00/0.02	0.00/0.00/0.00	0.00/0.00/0.00	—
Dantzig	0.01/**1.00**/29.63	0.00/0.00/2.39	0.00/0.00/0.00	0.00/0.00/0.00	0.00/0.00/0.00	—

High dimension, specifying 3 variables to select

Models	LM	GAM	Nonlin1	Nonlin2	Nonlin3	SO(3)
1-NN	0.88/2.88	0.14/1.81	0.24/1.18	0.02/0.26	0.02/0.28	0.83/2.75
MST	0.93/2.93	0.37/2.28	0.57/2.04	0.13/0.77	0.15/0.88	0.99/2.99
2-NN	0.97/2.97	0.53/2.49	0.49/2.04	0.15/0.79	0.16/0.99	0.97/2.97
10-NN	**1.00/3.00**	0.96/2.96	0.99/2.99	0.72/2.42	0.89/2.79	**1.00/3.00**
KPC (RKHS)	**1.00/3.00**	**0.98/2.98**	**1.00/3.00**	**0.92/2.88**	**1.00/3.00**	**1.00/3.00**
FOCI	0.42/2.42	0.02/1.09	0.02/0.11	0.00/0.10	0.01/0.07	—
LARS	**1.00/3.00**	0.01/1.82	0.00/0.11	0.00/0.02	0.00/0.24	—

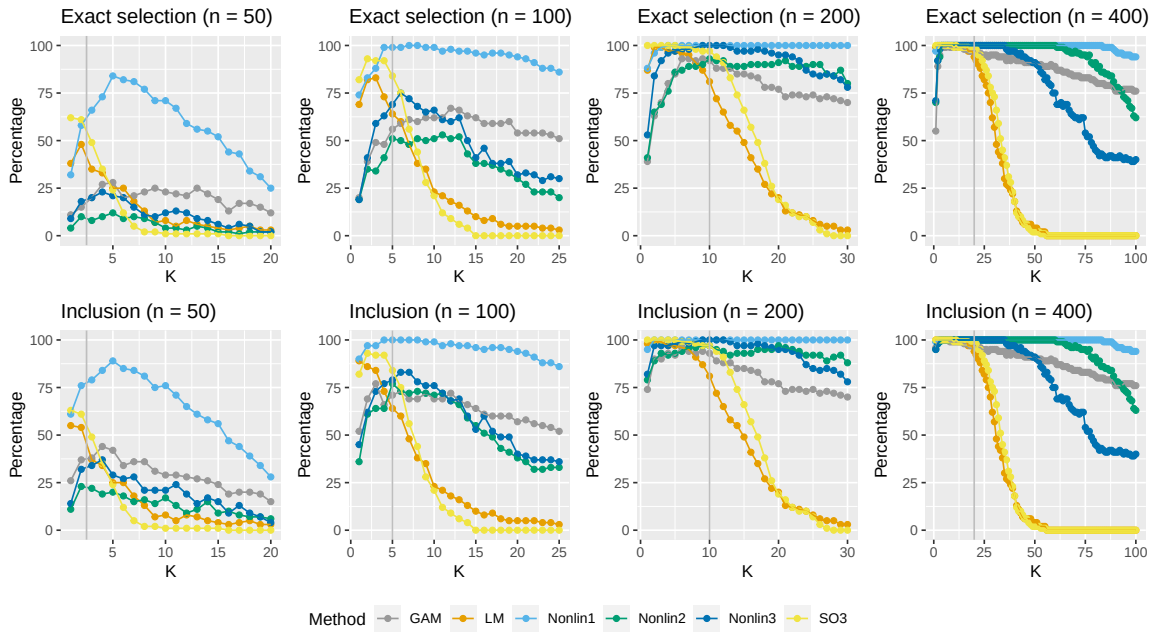

Figure 3: The proportion of times $\{X_1, X_2, X_3\}$ is exactly selected and the proportion of times $\{X_1, X_2, X_3\}$ is selected (possibly with other variables) with different K-NN graphs and the number of samples n.

'FWDselect (GAM)', the forward stepwise variable selection algorithm for general additive models, using the function `selection` in the `R` package `FWDselect` (Sestelo et al., 2015). We also compare with `varimp`, which selects three variables with the highest importance scores in the random forest model, implemented by the function `varimp` in the `R` package `party` (Hothorn et al., 2015) (using default settings). In the second tabula of Table 2 we report: (i) The proportion of times $\{X_1, X_2, X_3\}$ is exactly selected, and (ii) the average number of correct variables among the 3 selected variables $\hat{S}$ (i.e., $|\hat{S} \cap \{X_1, X_2, X_3\}|$), by the different methods (in 100 replications). It can be seen that our methods achieve superior performance compared to the other algorithms. In particular, '10-NN' and 'KPC (RKHS)' select exactly X_1, X_2, X_3 more than 99% of the times, in all the models.

High-dimensional setting: To study the performance of the methods in the high-dimensional setting we increase p from 10 to 1000 keeping $n = 200$ and consider the same models as in the beginning of Section 6.1.2. We first compare our method KFOCI with popular high-dimensional variable selection algorithms which themselves determine the number of variables to select, such as Lasso (Tibshirani, 1996) and Dantzig selector (Candès and Tao, 2007). Lasso was implemented using the `R` package `glmnet` (Friedman et al., 2010) and the Dantzig selector was implemented using the package `hdme` (Sorensen, 2019); in both cases the tuning parameters were chosen as "lambda.1se"[15] obtained from 5-fold cross-validation. Their performances are reported in the third tabula of Table 2 as in the low-dimensional

15. "lambda.1se", as proposed in Hastie et al. (2009), is the largest value of lambda such that the mean cross-validated error is within 1 standard error of the minimum.

setting. It can be seen that our method KFOCI ('10-NN') still selects the correct variables most of the times in the high-dimensional setting and outperforms all other competitors in all models except in the linear model (LM). Even in the linear model, if the goal is to exactly select $\{X_1, X_2, X_3\}$, then KFOCI ('10-NN') also outperforms Lasso and the Dantzig selector that are designed specifically for this setting. In the nonlinear models, all the other methods are essentially never able to select the correct set of predictors.

We now fix the number of covariates to be selected to 3 to examine how well Algorithm 2 performs in the high-dimensional setting. We compare our methods with *least angle regression* (LARS) (Efron et al., 2004) implemented in the `R` package `lars` (Hastie and Efron, 2013). We use the first 3 variables selected by LARS; and in our examples it almost always selected the first 3 variables entering the Lasso path before any of the variables left the active set. With the same choices of the kernel and ε as before, in the fourth tabula of Table 2 we report the same numbers as in the low dimensional setting. Besides '10-NN', KPC (RKHS) performs very well in the high-dimensional regime, exactly selecting $\{X_1, X_2, X_3\}$ more than 90% of the times and achieving the best performance among all the methods in all the considered models. It can also be seen that the performance (exact selecting proportion) of KFOCI improves for all graphs when compared to the case when the number of predictors was not prespecified.

Summary: The comparison with the various variable selection methods reveal that KFOCI and Algorithm 2 (with the Gaussian kernel) have excellent performance, both in the low- and high-dimensional regimes, in linear and nonlinear models. Further, KFOCI has a stopping criterion that can automatically determine the number of variables to select. Even in the high-dimensional linear regression setting, our model-free approach KFOCI yields comparable, and sometimes better results than the Lasso and the Dantzig selector.

6.2 Real Data Examples

In this subsection we study the performance of our proposed methods on real data. The real data sets considered involve continuous, discrete, and non-Euclidean response variables. We also compare and contrast the performance of our methods with a number of existing and useful alternatives. In the following, unless otherwise specified, for the kernels $k, k_{\mathcal{X}}, k_{\ddot{\mathcal{X}}}$, we use the Gaussian kernel with empirically chosen bandwidth, i.e., $k_{\mathcal{Y}}(x, x') = \exp\left(\frac{-\|x-x'\|^2}{2s^2}\right)$, where s is the median (or the mean if the median is 0) of the pairwise distances $\{\|x_i - x_j\|\}_{i<j}$. We also normalize each real-valued variable to have mean 0 and variance 1. The MST on real data is often not unique, and we will first randomly permute the data before applying KFOCI() in `KPC`.

Surgical data: The *surgical* data, available in the `R` package `olsrr` (Hebbali, 2018), consists of survival data for $n = 54$ patients undergoing liver operation along with 8 covariates[16]. The response Y is the survival time. Investigators have found that a linear model taking $\log(Y)$ as response and 4 out of the 8 covariates as predictors describe the data well (Kutner et al., 2004, Section 9.4, Section 10.6). In particular, this linear model

16. The 8 covariates are: bcs (blood clotting score), pindex (prognostic index), enzyme_test (enzyme function test score), liver_test (liver function test score), age, gender, indicator variable for gender, alc_mod (indicator variable for history of alcohol use), and alc_heavy (indicator for history of alcohol use).

Table 3: The variables selected by different methods for the surgical data set.

Methods	# variables selected	Selected variables
Stepwise forward regression	4/8	enzyme_test, pindex, alc_heavy, bcs
Best subset (BIC, PRESS_p)	4/8	enzyme_test, pindex, alc_heavy, bcs
KFOCI (1,2,3-NN, MST)	4/8	enzyme_test, pindex, liver_test, alc_heavy
KPC (RKHS, first 3 variables)	3/8	enzyme_test, pindex, bcs
VSURF (interpretation step)	4/8	enzyme_test, liver_test, pindex, alc_heavy
VSURF (prediction step)	3/8	enzyme_test, liver_test, pindex
FOCI	3/8	enzyme_test, liver_test, alc_heavy
npvarselec	4/8	pindex, enzyme_test, liver_test, alc_heavy
MMPC	4/8	pindex, enzyme_test, liver_test, alc_heavy

can be obtained by stepwise forward regression with any *penter* $\in [0.001, 0.1]$[17], and it is also the best submodel in terms of BIC and PRESS_p (Kutner et al., 2004, Section 9.4). We also work with the transformed response $\log(Y)$. We compare this linear model with our KFOCI and other variable selection methods such as FOCI (Azadkia and Chatterjee, 2021), `VSURF` (Genuer et al., 2019), `npvarselec` (Zambom and Akritas, 2017), and `MMPC` (Lagani et al., 2017) (all using their default settings). The selected variables, obtained by the different methods, are shown in Table 3. It can be seen the variables selected by KFOCI (with MST, 1-NN, 2-NN, 3-NN graphs) and Algorithm 2 are very similar to those selected by the carefully analyzed linear regression approach[18]. Further, KFOCI selects the same set of variables as many well-implemented variable selection algorithms such as `VSURF` (Genuer et al., 2019), `npvarselec` (Zambom and Akritas, 2017), and `MMPC` (Lagani et al., 2017).

Spambase data: This data set is available from the UCI Machine Learning Repository (Dua and Graff, 2017), consisting of a response denoting whether an e-mail is spam or not, along with 57 covariates (this example has also been studied in Azadkia and Chatterjee, 2021). The number of instances is $n = 4601$. We first use KFOCI to select a subset of variables. Directed 1,2,10-NN graphs and MST are considered. Since there is randomness in breaking ties for K-NN graphs, and in permuting the data for MST, we show in the histogram in Figure 4 the number of variables selected by different methods in 200 repetitions. We then assign each data point with probability 0.8 (0.2) to training (test) set, and fit a random forest model (implemented in the `R` package `randomForest` (Liaw and Wiener, 2002) using default settings) with only the selected covariates. The mean squared error (MSE) on the test set is reported on the right panel of Figure 4. Note that as there is randomness in breaking ties, splitting the training/test sets, as well as fitting the random forest, we repeat the whole procedure 200 times and present the box-plot for the MSEs. It can be seen that in terms of MSE, '1-NN' is better than FOCI, and '2-NN' and '10-NN' are better than '1-NN'. The random forest fit on the whole set of covariates achieves the best performance in prediction as expected. However, our methods use only about 1/3 of the 57 covariates to achieve a prediction performance which is not much worse than that of the random forest fit with all the covariates.

17. The variable with the smallest p-value less than 'penter' enters the model at every stage.
18. The 'bcs' selected by the final linear model in Kutner et al. (2004) is replaced by 'liver_test' by KFOCI (MST,1,2,3-NN). The first 3 variables selected by Algorithm 2 are also among the 4 predictors of the final linear model.

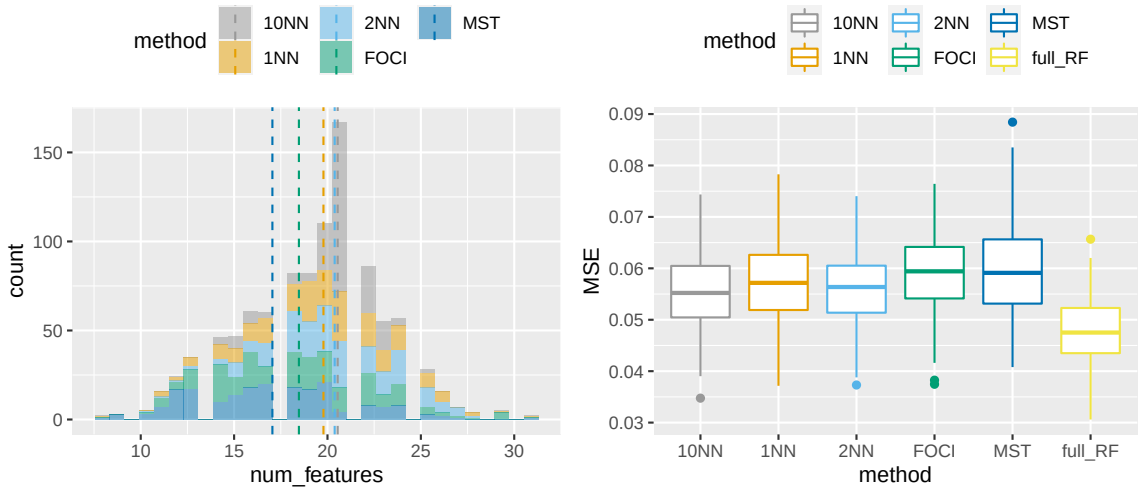

Figure 4: Left: Histogram of the number of selected variables (from 200 replications). The dashed lines show the mean for each group. Right: MSE after fitting random forest on test set for different methods (using 200 replications). For FOCI and KFOCI ('1-NN','2-NN','10-NN','MST'), the random forests were fit on the selected covariates. 'full_RF' fits random forest with the entire set of covariates.

Election data (histogram-valued response): Consider the 2017 Korean presidential election data collected by `https://github.com/OhmyNews/2017-Election`, which has been analyzed in the recent paper Jeon and Park (2020). The data consist of the voting results earned by the top five candidates from 250 electoral districts in Korea. Since the top three candidates from three major parties representing progressivism, conservatism and centrism earned most of the votes, we will focus on the proportion of votes earned by each of these three candidates among them, i.e., $Y = (Y_1, Y_2, Y_3)$ with $Y_1, Y_2, Y_3 \geq 0$ and $Y_1 + Y_2 + Y_3 = 1$. The demographic information *average age* (X_1), *average years of education* (X_2), *average housing price per square meter* (X_3) and *average paid national health insurance premium* (X_4) are available for each electoral district. Note that Y can be viewed as a histogram-valued variable for which the following two characteristic kernels on $[0, \infty)^3$ are available (Fukumizu et al., 2009b): $k_1(a, b) = \prod_{i=1}^{3}(a_i + b_i + 1)^{-1}$, $k_2(a, b) = e^{-\sum_{i=1}^{3}\sqrt{a_i + b_i}}$, where $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$. Since $Y \in \mathbb{R}^3$, we can also use a Gaussian kernel. Note that Y, taking values in the probability simplex, is an example of compositional data. Besides political science, such data are prevalent in many other fields such as sedimentology (Hijazi and Jernigan, 2009; Aitchison, 1986), hydrochemistry (Otero et al., 2005), economics (Morais et al., 2018), and bioinformatics (Xia et al., 2013; Chen and Li, 2016; Shi et al., 2016). Regression methods have been developed with compositional data as covariates (Shi et al., 2016; Lin et al., 2014; Susin et al., 2020), as response (Aitchison, 1986; Iyengar and Dey, 2002; Hijazi and Jernigan, 2009; Glahn and Hron, 2012; Tsagris, 2015; Tsagris et al., 2020), or as both covariates and response (Chen et al., 2017). But few variable selection methods have been proposed for data with compositional response and Euclidean covariates,

Table 4: The selected variables with different kernels and different graphs.

Kernels	1-NN	2-NN	MST	3-NN	4-NN	5-NN
Gaussian	1 2 3 4	1 2 4 3	1 2 4 3	1 2 4 3	1 2 4	1 2 4
k_1	1 2	1 2	1 2 4 3	1 2 4 3	1 2 4 3	1 2 4
k_2	1 2	1 2	1 2 4 3	2 1 4 3	1 2 4 3	1 2 4

Table 5: The estimates of $\rho^2(Y, X_i|X_{-i})$ with different graphs and regularizers.

Estimators	1-NN	2-NN	MST	3-NN	4-NN	5-NN
$\hat{\rho}^2(Y, X_1\|X_{-1})$	0.07	0.15	0.13	0.17	0.14	0.13
$\hat{\rho}^2(Y, X_2\|X_{-2})$	0.19	0.15	0.16	0.09	0.05	0.04
$\hat{\rho}^2(Y, X_3\|X_{-3})$	0.10	0.06	0.02	0.01	-0.02	-0.02
$\hat{\rho}^2(Y, X_4\|X_{-4})$	0.07	0.07	0.06	0.04	0.03	0.01
Estimators	$\varepsilon = 10^{-3}$	$\varepsilon = 10^{-3.5}$	$\varepsilon = 10^{-4}$	$\varepsilon = 10^{-4.5}$	$\varepsilon = 10^{-5}$	$\varepsilon = 10^{-5.5}$
$\tilde{\rho}^2(Y, X_1\|X_{-1})$	0.08	0.12	0.15	0.17	0.19	0.21
$\tilde{\rho}^2(Y, X_2\|X_{-2})$	0.03	0.06	0.07	0.09	0.11	0.14
$\tilde{\rho}^2(Y, X_3\|X_{-3})$	0.04	0.05	0.06	0.08	0.10	0.12
$\tilde{\rho}^2(Y, X_4\|X_{-4})$	0.03	0.04	0.05	0.07	0.08	0.10

as in our case here. Since our method tackles random variables taking values in general topological spaces, it readily yields a variable selection method for this case.

The variables selected by KFOCI with different kernels and graphs are given in Table 4. It can be seen that in all cases, X_1 and X_2 are selected, indicating that they may be more relevant for predicting Y than X_3, X_4. This agrees with Jeon and Park (2020, Figure 3), where an additive regression was fit to the election data. It was also found that as people are more educated, from 'low' to 'medium' educational level, their political orientation becomes more conservative, but interestingly it is reversed for people as they move from 'medium' to 'high' educational level (Jeon and Park, 2020). This nonlinear relationship is also captured by KPC, as X_2 is always selected. Note that X_3 and X_4 are both measures of wealth. It is natural to conjecture if one of them, say X_4 alone, would be enough in predicting Y. We can answer this by examining how large $\rho^2(Y, X_3|X_{-3})$ is: If the estimate of $\rho^2(Y, X_3|X_{-3})$ is close to 0 it provides evidence that Y is conditionally independent of X_3 given $X_{-3} = \{X_1, X_2, X_4\}$. The estimates of $\rho^2(Y, X_i|X_{-i})$ are given in Table 5. It can be seen that for $i = 3, 4$, the conditional association of Y and X_i given all other variables is weaker than that of $i = 1, 2$; this agrees with the intuition that X_3 and X_4 are both capturing the same latent factor—wealth. In Table 5 we also give values of $\tilde{\rho}^2$ as we change ε. For a suitably chosen ε, say $\varepsilon = 10^{-4}$, $\tilde{\rho}^2$ is similar to $\hat{\rho}^2$.

Appendix A. Some General Discussions

In this section we elaborate on some parts of the main text that were initially deferred so as not to impede the flow of the paper.

A.1 Some Remarks

Remark 15 (About Assumption 2) *Assumption 2 is needed to ensure following: (a) The feature map $y \mapsto k_{\mathcal{Y}}(y, \cdot)$ is measurable*[19]*; (b) any Borel measurable function $g : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ is strongly measurable*[20] *and the Bochner integral (see e.g., Cohn, 2013, Appendix E for its formal definition) $\mathbb{E}[g(Y)]$ is well-defined whenever $\mathbb{E}\|g(Y)\|_{\mathcal{H}_{\mathcal{Y}}} < \infty$*[21]*; (c) the relevant cross-covariance operators are Hilbert-Schmidt. Thus, for simplicity, we assume the more natural Assumption 2 rather than the conditions (a)-(c) above.*

Remark 16 *Observe that if for almost every x, $Y|X = x$ is degenerate, then Y is almost surely a measurable function of X. To see this, let $Q(x, \cdot)$ be a regular conditional distribution of Y given $X = x$. Then there exists $A \subset \mathcal{X}$ of probability 1 such that $Q(x, \cdot)$—the transition kernel—is degenerate for all $x \in A$. Let $f(x) \in \mathcal{Y}$ be the support (which is a single element in $\mathcal{Y}$) of $Q(x, \cdot)$ if $x \in A$, and some fixed $y_0 \in \mathcal{Y}$ if $x \in A^c$. Then $f : \mathcal{X} \to \mathcal{Y}$ is a measurable function because for any measurable $B \subset \mathcal{Y}$, $f^{-1}(B) = \{x \in \mathcal{X} : Q(x, B) = 1\} \cap A$ if $y_0 \notin B$ and $f^{-1}(B) = (\{x \in \mathcal{X} : Q(x, B) = 1\} \cap A) \cup A^c$ if $y_0 \in B$. In either case, $f^{-1}(B)$ is a measurable set. Note that $Y = f(X)$ a.s. so Y is a measurable function of X.*

Remark 17 (Cross-covariance operator in tensor-product space) *Here we show that* (14) *holds. We will use the isometric isomorphism $\Phi : \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}} \to \mathrm{HS}(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{X}})$ which satisfies $\langle f, \Phi(a)g\rangle_{\mathcal{H}_{\mathcal{X}}} = \langle a, f \otimes g\rangle_{\mathcal{H}_{\mathcal{X}} \times \mathcal{H}_{\mathcal{Y}}}$ for $a \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$, $f \in \mathcal{H}_{\mathcal{X}}$, $g \in \mathcal{H}_{\mathcal{Y}}$. Denote the right-hand side of* (14) *by C. Then*

$$\begin{aligned}
\langle f, Cg\rangle_{\mathcal{H}_{\mathcal{X}}} &= \langle C, f \otimes g\rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} \\
&= \langle \mathbb{E}\left[k_{\mathcal{X}}(X, \cdot) \otimes k_{\mathcal{Y}}(Y, \cdot)\right], f \otimes g\rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} - \langle \mathbb{E}\left[k_{\mathcal{X}}(X, \cdot)\right], f\rangle_{\mathcal{H}_{\mathcal{X}}} \langle \mathbb{E}\left[k_{\mathcal{Y}}(Y, \cdot)\right], g\rangle_{\mathcal{H}_{\mathcal{Y}}} \\
&= \mathbb{E}\left[\langle k_{\mathcal{X}}(X, \cdot) \otimes k_{\mathcal{Y}}(Y, \cdot), f \otimes g\rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}}\right] - \mathbb{E}\left[\langle k_{\mathcal{X}}(X, \cdot), f\rangle_{\mathcal{H}_{\mathcal{X}}}\right] \cdot \mathbb{E}\left[\langle k_{\mathcal{Y}}(Y, \cdot), g\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&= \mathbb{E}\left[f(X)g(Y)\right] - \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)].
\end{aligned}$$

In the third line we used the fact that if $X \in \mathcal{X}$ is Bochner integrable, then $\mathbb{E}\varphi(X) = \varphi(\mathbb{E}X)$ for any bounded linear functional $\varphi \in \mathcal{X}^*$ (see e.g., Cohn, 2013, Appendix E).

Remark 18 *Here we show that suppose (Y, X, Z) is jointly Gaussian. Then, when using the linear kernel, $\rho^2(Y, Z|X) = 0$ implies $Y \perp\!\!\!\perp Z|X$.*

For Gaussian distribution, the conditional distribution $(Y, Z)|X$ is Gaussian with conditional variance and covariance not depending on the value of X. Suppose:

$$(Y^{(i)}, Z)^\top | X = x \sim N\left(\begin{pmatrix}\mu_i(x)\\ \mu_z(x)\end{pmatrix}, \begin{pmatrix}\sigma_i^2 & \beta_i^\top\\ \beta_i & \Sigma_z\end{pmatrix}\right)$$

19. Note that the canonical feature map $y \mapsto k_{\mathcal{Y}}(y, \cdot)$ is measurable if $\mathcal{H}_{\mathcal{Y}}$ is separable (Steinwart and Christmann, 2008, Lemma 4.25).
20. A function is strongly measurable if it is Borel measurable and has a separable range.
21. Thus, if $\mathbb{E}\sqrt{k_{\mathcal{Y}}(Y, Y)} < \infty$, $\mu_Y := \mathbb{E}k_{\mathcal{Y}}(Y, \cdot)$ is well-defined as a Bochner integral. As for any bounded linear functional $T : \mathcal{H}_{\mathcal{Y}} \to \mathbb{R}$, we have $\mathbb{E}\left[T(g(Y))\right] = T\left(\mathbb{E}[g(Y)]\right)$, we can write $\langle \mu_Y, f\rangle_{\mathcal{H}_{\mathcal{Y}}} = \langle \mathbb{E}[k_{\mathcal{Y}}(Y, \cdot)], f\rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}[\langle k_{\mathcal{Y}}(Y, \cdot), f\rangle_{\mathcal{H}_{\mathcal{Y}}}] = \mathbb{E}f(Y)$.

where $\mu_i(\cdot)$ and $\mu_z(\cdot)$ are linear functions. By (9)*,*

$$\rho^2(Y,Z|X) = \frac{\sum_{i=1}^d \mathbb{E}(\mathrm{Var}[\mathbb{E}(Y^{(i)}|X,Z)|X])}{\mathbb{E}\left[\sum_{i=1}^d \mathrm{Var}(Y^{(i)}|X)\right]} = \frac{\sum_{i=1}^d \mathbb{E}(\mathrm{Var}[\mu_i(X) + \beta_i^\top \Sigma_z^{-1}(Z-\mu_z(X))|X])}{\mathbb{E}[\sum_{i=1}^d \sigma_i^2]}$$
$$= \frac{\sum_{i=1}^d \beta_i^\top \Sigma_z^{-1}\Sigma_z\Sigma_z^{-1}\beta_i}{\sum_{i=1}^d \sigma_i^2} = \frac{\sum_{i=1}^d \beta_i^\top \Sigma_z^{-1}\beta_i}{\sum_{i=1}^d \sigma_i^2}.$$

Here Σ_z^{-1} is regarded as the generalized inverse or Moore-Penrose inverse of Σ_z. We can suppose Z is non-degenerate, otherwise $Y \perp\!\!\!\perp Z|X$ is trivial. Without loss of generality, we may further suppose Σ_z to be invertible, since otherwise we can perform a bijective linear transformation to reduce the dimensionality of Z. Now Σ_z^{-1} is positive-definite and $\rho^2(Y,Z|X) = 0$ implies $\beta_i = 0$ for all i, which further indicates $Y \perp\!\!\!\perp Z|X$ by the property of multivariate Gaussian distribution.

A.2 Uncentered CME Estimators

We used the 'centered' CMEs (15) to derive an expression for the estimator of ρ^2 since that requires less restrictive assumptions (Klebanov et al., 2020). But using 'uncentered CMEs', i.e., using uncentered (cross)-covariance operators to construct an expression of the CME is also possible. Under appropriate assumptions, the uncentered (cross)-covariance operators yield the following uncentered CME formula (Klebanov et al., 2020, Theorem 5.3):

$$\mu_{Y|X=x} = \left({}^uC_X^\dagger \, {}^uC_{XY}\right)^* k_{\mathcal{X}}(x,\cdot), \tag{26}$$

where ${}^uC_{XY}$ (resp. ${}^uC_X$) is the uncentered cross-covariance (resp. covariance) operator defined as $\langle f, {}^uC_{XY} g\rangle = \mathbb{E}\left[f(X)g(Y)\right]$. The same methodology as in Section 4.1 shows that one can simply replace every occurence $\tilde{K}$ by the corresponding K (e.g., $\tilde{K}_X$ by K_X) to obtain the following simplified formula: $\tilde{\rho}_u^2 = \frac{\mathrm{Tr}(M^\top K_Y M)}{\mathrm{Tr}(N^\top K_Y N)}$, where $M = K_X\left(K_X + \varepsilon n I\right)^{-1} - K_{\ddot{X}}\left(K_{\ddot{X}} + \varepsilon n I\right)^{-1}$ and $N = I - K_X\left(K_X + \varepsilon n I\right)^{-1}$.

We see that the centered estimator $\tilde{\rho}^2$ (in Equation 18) can be obtained by putting a 'tilde' on all the kernel matrices (i.e., $K_X, K_Y, K_{\ddot{X}}$) in $\tilde{\rho}_u^2$, and this can be viewed as performing a 'centralization' of the feature maps of the corresponding kernel, i.e., if $K = \Phi\Phi^\top$, then $\tilde{K} = HKH = H\Phi\Phi^\top H = \Phi_c\Phi_c^\top$, with $\Phi_c := H\Phi$ being the centralized feature.

Remark 19 (Sufficient conditions for uncentered CME) *A sufficient condition for* (26) *to be valid is: For any $g \in \mathcal{H}_{\mathcal{Y}}$, a version of $\mathbb{E}[g(Y)|X=\cdot] \in \mathcal{H}_{\mathcal{X}}$. In the case when X and Y are independent, $\mathbb{E}[g(Y)|X=\cdot]$ is a constant (P_X-a.e.). But it is well-known that the RKHS with the Gaussian kernel does not contain nonzero constant functions whenever $\mathcal{X}$ has non-empty interior (Minh, 2010). This provides an example where the CME is not guaranteed to be expressed using uncentered (cross)-covariance operators as in* (26)*. Note that Assumption 9 always holds when X and Y are independent. However, in general, sufficient conditions for explicit expressions of CMEs in terms of centered and uncentered (cross)-covariance operators are usually restrictive and hard to verify (Klebanov et al., 2020).*

The following result (see Section B.11 for a proof) shows that $\tilde{\rho^2_u}$ has an interesting connection to kernel ridge regression.

Proposition 7 *Suppose $\mathcal{Y} = \mathbb{R}$ is equipped with the linear kernel $k_{\mathcal{Y}}(u, v) = uv$. Then, $\tilde{\rho^2_u} = \frac{\|\hat{\mathbf{Y}}_x - \hat{\mathbf{Y}}_{xz}\|^2}{\|\hat{\mathbf{Y}}_x - \mathbf{Y}\|^2}$, where $\mathbf{Y} = (Y_1, \ldots, Y_n)$, $\hat{\mathbf{Y}}_x$ is the kernel ridge regression estimator when regressing Y on X and $\hat{\mathbf{Y}}_{xz}$ is the kernel ridge regression estimator when regressing Y on $\ddot{X}$, both with regularization parameter $n\varepsilon$.*

A.3 Invariance and Continuity

Informally, *invariance* means that the measure should be unaffected under a "suitable" class of transformations and *continuity* means that whenever a sequence of measures P_n converges to P (in an "appropriate" sense), the sequence of values of the dependence measure for P_n's should converge to that of P. We show, in the following result (proved in Appendix B.6) that ρ^2 satisfies these properties for a class of kernels.

Proposition 8 *The following two properties hold for $\rho^2(Y, Z|X)$:*

1. *Invariance. $\rho^2(Y, Z|X)$ is invariant to any bijective transformation of X, and any bijective transformation of Z. If $\mathcal{Y} = \mathbb{R}^d$ and the kernel is of the form (8) then ρ^2 is also invariant to orthogonal transformations and translations of Y.*
2. *Continuity. Let $\mathcal{Y}$ be a separable metric space and let $(X_n, Y_n, Z_n) \sim P_n$. Let $(Y_{n,1}, Y'_{n,1})$, $(Y_{n,2}, Y'_{n,2})$ be generated from the distribution of (X_n, Y_n, Z_n) as described in (6) and (7). If $(X_n, Y_n, Z_n, Y_{n,1}, Y'_{n,1}, Y_{n,2}, Y'_{n,2}) \xrightarrow{d} (X, Y, Z, Y_1, Y'_1, Y_2, Y'_2)$ where $(X, Y, Z) \sim P$, and $\limsup_{n\to\infty} \mathbb{E}[k_{\mathcal{Y}}^{1+\varepsilon}(Y_{n,i}, Y'_{n,i})] < \infty$ for $i = 1, 2$, and for some $\varepsilon > 0$, $\limsup_{n\to\infty} \mathbb{E}[k_{\mathcal{Y}}^{1+\varepsilon}(Y_n, Y_n)] < \infty$, then $\rho^2(Y_n, Z_n|X_n) \to \rho^2(Y, Z|X)$ as $n \to \infty$.*

For the estimators $\hat{\rho^2}, \tilde{\rho^2}$, it is also easy to see that if $\mathcal{X}, \mathcal{Y}, \mathcal{X} \times \mathcal{Z}$ are Euclidean, and both the geometric graph and kernels $k_{\mathcal{Y}}, k_{\mathcal{X}}, k_{\ddot{\mathcal{X}}}$ only depend on the inter-point distance, then $\hat{\rho}^2(Y, Z|X)$ and $\tilde{\rho}^2(Y, Z|X)$ are invariant to any orthogonal transformation of X or Y or Z.

A.4 Approximate Computation of $\tilde{\rho^2}$ Using Incomplete Cholesky Decomposition

For fast computation, we can approximate each of the kernel matrices by *incomplete Cholesky decomposition*. For example, we can approximate $K_X \approx L_1 L_1^\top$ where $L_1 \in \mathbb{R}^{n\times d_1}$ is the incomplete Cholesky decomposition of the kernel matrix[22] K_X which can be computed in $O(nd_1^2)$ time (here $d_1 \le n$) without the need to compute and store the full gram matrix K_X (Bach and Jordan, 2003). Let $\tilde{L}_1 := HL_1$ denote the centralized feature matrix (recall that $H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$). Then $\tilde{K}_X \approx \tilde{L}_1 \tilde{L}_1^\top$. Similarly, we can approximate each of the centralized kernel matrices $\tilde{K}_{\ddot{X}} \approx \tilde{L}_2 \tilde{L}_2^\top$, $\tilde{K}_Y \approx \tilde{L}_3 \tilde{L}_3^\top$, where $L_i \in \mathbb{R}^{n\times d_i}$ for $i = 2, 3$.

To compute the numerator of $\tilde{\rho^2}$ (see Equation 20), by the Woodbury matrix identity, we can approximate $(\tilde{K}_X + n\varepsilon I)^{-1}$ by $(\tilde{L}_1\tilde{L}_1^\top + n\varepsilon I_n)^{-1} = \frac{1}{n\varepsilon} I_n - \frac{1}{n\varepsilon}\tilde{L}_1(n\varepsilon I_{d_1} + \tilde{L}_1^\top \tilde{L}_1)^{-1}\tilde{L}_1^\top$. The same strategy applied to $(\tilde{K}_{\ddot{X}} + n\varepsilon I)^{-1}$ shows M can be approximated by $\breve{M} := \tilde{L}_1(n\varepsilon I_{d_1} +$

22. Note that if the linear kernel is used and $X_i \in \mathbb{R}^{d_1}$, then the decomposition $K_X = \mathbf{X}\mathbf{X}^\top$ (here $\mathbf{X} := [X_1, \ldots, X_n]^\top$ is the data matrix of the X_i's) is straight-forward and exact.

$\tilde{L}_1^\top \tilde{L}_1)^{-1}\tilde{L}_1^\top - \tilde{L}_2(n\varepsilon I_{d_2} + \tilde{L}_2^\top \tilde{L}_2)^{-1}\tilde{L}_2^\top$. Thus $\mathrm{Tr}(M^\top \tilde{K}_Y M) \approx \mathrm{Tr}(\check{M}^\top \tilde{L}_3 \tilde{L}_3^\top \check{M}) = \|\tilde{L}_3^\top \check{M}\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm.

For the denominator of $\tilde{\rho^2}$, note that $N = n\varepsilon(\tilde{K}_X + n\varepsilon I)^{-1} \approx I_n - \tilde{L}_1(n\varepsilon I_{d_1} + \tilde{L}_1^\top \tilde{L}_1)^{-1}\tilde{L}_1^\top =: \check{N}$. Thus, $\mathrm{Tr}(N^\top \tilde{K}_Y N) \approx \|\tilde{L}_3^\top \check{N}\|_F^2$. Combining the calculations above, we see that the approximate version of $\tilde{\rho^2}$ can be computed in $O\left(n \max_{1\le i\le 3} d_i^2\right)$ time.

A.5 Assumptions on the Geometric Graph

We will use the following assumptions on the graph functional $\mathcal{G}$; these assumptions were also made in Deb et al. (2020, Section 3).

Assumption 10 *Given the graph $\mathcal{G}_n$, let $N(i)$ be a uniformly sampled index from among the (out-)neighbors of X_i in $\mathcal{G}_n$, i.e., $\{j : (X_i, X_j) \in \mathcal{E}(\mathcal{G}_n)\}$. Assume that $\mathcal{X}$ is a metric space with metric $\rho_{\mathcal{X}}$, and that $\rho_{\mathcal{X}}(X_1, X_{N(1)}) \xrightarrow{p} 0, \quad \text{as } n \to \infty$.*

Assumption 11 *Assume that there exists a deterministic positive sequence $r_n \ge 1$ (may or may not be bounded), such that $\min_{1\le i\le n} d_i \ge r_n$ almost surely. Let $\mathcal{G}_{n,i}$ denote the graph obtained from $\mathcal{G}_n$ by replacing X_i with an i.i.d. random element X_i'. Assume that there exists a deterministic positive sequence q_n (may or may not be bounded), such that*

$$\max_{1\le i\le n} \max\{|\mathcal{E}(\mathcal{G}_n) \setminus \mathcal{E}(\mathcal{G}_{n,i})|, |\mathcal{E}(\mathcal{G}_{n,i}) \setminus \mathcal{E}(\mathcal{G}_n)|\} \le q_n \quad a.s. \qquad \text{and} \qquad \frac{q_n}{r_n} = O(1).$$

Assumption 12 *There exists a deterministic sequence $\{t_n\}_{n\ge 1}$ (may or may not be bounded) such that the vertex degree (including both in- and out-degrees for directed graphs) of every point X_i (for $i = 1, \ldots, n$) is bounded by t_n, and $\frac{t_n}{r_n} = O(1)$.*

Appendix B. Proofs

B.1 Proof of Lemma 1

First observe that by Assumption 1, $\mathbb{E}\left[\|k_{\mathcal{Y}}(Y_1, \cdot)\|_{\mathcal{H}_{\mathcal{Y}}}^2\right] = \mathbb{E}\left[k_{\mathcal{Y}}(Y, Y)\right] < \infty$. By the Cauchy-Schwarz's inequality, $\mathbb{E}[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]]$ is upper bounded by

$$\mathbb{E}\Big[\mathbb{E}[\|k_{\mathcal{Y}}(Y_1, \cdot)\|_{\mathcal{H}_{\mathcal{Y}}} \cdot \|k_{\mathcal{Y}}(Y_1', \cdot)\|_{\mathcal{H}_Y} \ |X]\Big] \le \mathbb{E}\left[\mathbb{E}\Big[\frac{\|k_{\mathcal{Y}}(Y_1, \cdot)\|_{\mathcal{H}_{\mathcal{Y}}}^2 + \|k_{\mathcal{Y}}(Y_1', \cdot)\|_{\mathcal{H}_{\mathcal{Y}}}^2}{2}|X\Big]\right] < \infty.$$

Similarly, $\mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_2, Y_2')|X, Z]\right] < \infty$. Hence the numerator of $\rho^2(Y, Z|X)$ in (5) is finite. The denominator is also finite by Assumption 1.

Next we show that $\mathbb{E}[\mathrm{MMD}^2(\delta_Y, P_{Y|X})] \ne 0$. If $\mathbb{E}[\mathrm{MMD}^2(\delta_Y, P_{Y|X})] = 0$, then equivalently $\mathbb{E}\|k_{\mathcal{Y}}(Y, \cdot) - \mathbb{E}[k_{\mathcal{Y}}(Y, \cdot)|X]\|_{\mathcal{H}_{\mathcal{Y}}}^2 = 0$. So, conditional on $X = x$, $k_{\mathcal{Y}}(Y, \cdot) = \mathbb{E}[k_{\mathcal{Y}}(Y, \cdot)|X = x]$ for P_X-a.e. x. Since $k_{\mathcal{Y}}$ is characteristic, $y \mapsto k_{\mathcal{Y}}(y, \cdot)$ is injective, which implies that $Y|X = x$ is degenerate for P_X-a.e. x. This is a contradiction to Assumption 3. ■

B.2 Proof of Lemma 2

Let us first try to explain the notation in definition (1). From the definition of the MMD, ρ^2 in (1) can be re-expressed as: $\rho^2 = \frac{\mathbb{E}\left[\|\mu_{P_{Y|XZ}} - \mu_{P_{Y|X}}\|_{\mathcal{H}_{\mathcal{Y}}}^2\right]}{\mathbb{E}\left[\|\mu_{\delta_Y} - \mu_{P_{Y|X}}\|_{\mathcal{H}_{\mathcal{Y}}}^2\right]}$ where the mean embeddings

above have the following expressions: $\mu_{\delta_Y}(\cdot) = k_{\mathcal{Y}}(\cdot, Y)$, $\mu_{P_{Y|X}}(\cdot) = \mathbb{E}_{Y_1 \sim Y|X}[k_{\mathcal{Y}}(\cdot, Y_1)] = \mathbb{E}[k_{\mathcal{Y}}(\cdot, Y)|X]$, $\mu_{P_{Y|XZ}}(\cdot) = \mathbb{E}_{Y_2 \sim Y|XZ}[k_{\mathcal{Y}}(\cdot, Y_2)] = \mathbb{E}[k_{\mathcal{Y}}(\cdot, Y)|X, Z]$, where the expectations should be understood as Bochner integrals (see e.g., Diestel and Faires, 1974; Dinculeanu, 2011). Using the notation in the statement of the lemma, we have

$$\begin{aligned}
\mathbb{E}\left[\|\mu_{\delta_Y} - \mu_{P_{Y|X}}\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] &= \mathbb{E}\left[\|\mu_{\delta_Y}\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] + \mathbb{E}\left[\|\mu_{P_{Y|X}}\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] - 2\mathbb{E}\left[\langle \mu_{\delta_Y}, \mu_{P_{Y|X}} \rangle_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&= \mathbb{E}[k_{\mathcal{Y}}(Y, Y)] + \mathbb{E}_X\left[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]\right] - 2\mathbb{E}_{Y,X}[\mathbb{E}[k_{\mathcal{Y}}(Y, Y_1)|X]] \\
&= \mathbb{E}[k_{\mathcal{Y}}(Y, Y)] - \mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]\right],
\end{aligned}$$

where $Y_1, Y_1'|X = x \overset{i.i.d.}{\sim} P_{Y|x}$, Y_1, Y are conditionally independent given X, and the second equality follows from the observations: $\|\mu_{P_{Y|X}}\|^2_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]$, and $\langle \mu_{\delta_Y}, \mu_{P_{Y|X}} \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}[k_{\mathcal{Y}}(Y, Y_1)|X]$. Similarly, we can show that $\mathrm{MMD}^2(P_{Y|XZ}, P_{Y|X})$ equals

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbb{E}[k_{\mathcal{Y}}(Y, \cdot)|X, Z] - \mathbb{E}[k_{\mathcal{Y}}(Y, \cdot)|X]\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] \\
= \;& \mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_2, Y_2')|X, Z]\right] + \mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]\right] - 2\mathbb{E}\left[\mathbb{E}_{Y_1 \sim Y|X, Y_2 \sim Y|X, Z} k_{\mathcal{Y}}(Y_1, Y_2)\right] \\
= \;& \mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_2, Y_2')|X, Z]\right] - \mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]\right].
\end{aligned}$$

This proves the result. ■

B.3 Proof of Theorem 1

The proof is divided into three parts.

Step 1. We will first show that $\rho^2(Y, Z|X) \in [0, 1]$. Observe that $\rho^2 \geq 0$ is clear. To show that $\rho^2 \leq 1$, we will use the following result—a version of Jensen's inequality (Perlman, 1974, Theorems 3.6 and 3.8).

Lemma 5 (Jensen's inequality) *Let $\mathcal{W}$ be a real Banach space, W be a Bochner integrable random variable taking value in $\mathcal{W}$, and $g : \mathcal{W} \to \mathbb{R}$ be a lower-semicontinuous convex function such that $g(W)$ is integrable. Then $g(\mathbb{E}W) \leq \mathbb{E}g(W)$. If g is strictly convex*[23] *and $\mathbb{P}(W = \mathbb{E}W) < 1$, then $g(\mathbb{E}W) < \mathbb{E}g(W)$.*

Now, observe that,

$$\begin{aligned}
&\mathbb{E}[\mathrm{MMD}^2(\delta_Y, P_{Y|X})] = \mathbb{E}\|k_{\mathcal{Y}}(Y, \cdot) - \mathbb{E}[k_{\mathcal{Y}}(Y, \cdot)|X]\|^2_{\mathcal{H}_{\mathcal{Y}}} \\
&= \mathbb{E}\left[\mathbb{E}[\|k_{\mathcal{Y}}(Y, \cdot) - \mathbb{E}(k_{\mathcal{Y}}(Y, \cdot)|X)\|^2_{\mathcal{H}_{\mathcal{Y}}}|X, Z]\right] \geq \mathbb{E}\left[\|\mathbb{E}[k_{\mathcal{Y}}(Y, \cdot) - \mathbb{E}(k_{\mathcal{Y}}(Y, \cdot)|X)|X, Z]\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&= \mathbb{E}\left[\|\mathbb{E}(k_{\mathcal{Y}}(Y, \cdot)|X, Z) - \mathbb{E}(k_{\mathcal{Y}}(Y, \cdot)|X)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] = \mathbb{E}[\mathrm{MMD}^2(P_{Y|XZ}, P_{Y|X})].
\end{aligned}$$

where we have applied the above Jensen's inequality to the function $g : f \mapsto \|f\|^2_{\mathcal{H}_{\mathcal{Y}}}$ and $W := k_{\mathcal{Y}}(Y, \cdot) - \mathbb{E}[k_{\mathcal{Y}}(Y, \cdot)|X]$. Hence $\rho^2 \leq 1$.

Step 2. Next we show that $\rho^2 = 1$ if and only if Y is a measurable function of Z and X.

If $\rho^2 = 1$, then for a.e. x, z, the above Jensen's inequality attains equality, which means that (Lemma 5): $\mathbb{P}\left(k_{\mathcal{Y}}(Y, \cdot) = \mathbb{E}_{Y|X=x, Z=z} k_{\mathcal{Y}}(Y, \cdot) \middle| X = x, Z = z\right) = 1$. Hence Y is degenerate given $(X, Z) = (x, z)$, and thus a measurable function of X, Z (Remark 16).

23. By strictly convexity we mean: $g(\lambda x + (1 - \lambda)y) < \lambda g(x) + (1 - \lambda)g(y),\ \forall x \neq y, \lambda \in (0, 1)$.

Conversely, if $Y = f(X, Z)$ for some measurable function f, then

$$\mathbb{E}[\mathrm{MMD}^2(P_{Y|XZ}, P_{Y|X})] = \mathbb{E}[\mathrm{MMD}^2(\delta_{f(X,Z)}, P_{Y|X})] = \mathbb{E}[\mathrm{MMD}^2(\delta_Y, P_{Y|X})],$$

and so $\rho^2(Y, Z|X) = 1$.

Step 3. Now we have to show that $\rho^2(Y, Z|X) = 0$ if and only if $P_{Y|XZ} = P_{Y|X}$ a.s.

If $P_{Y|XZ} = P_{Y|X}$ a.s., then $\mathbb{E}[\mathrm{MMD}^2(P_{Y|XZ}, P_{Y|X})] = 0$, and thus $\rho^2(Y, Z|X) = 0$. Conversely, if $\rho^2(Y, Z|X) = 0$, since $k_{\mathcal{Y}}$ is characteristic, $P_{Y|XZ} = P_{Y|X}$ almost surely. ■

B.4 Proof of Proposition 1

For notational clarity, write $r \equiv \rho_{YZ\cdot X}$. By assumption, there exist $\mu_1, \mu_2, \sigma_1, \sigma_2$ such that:

$$(Y, Z)^\top | X \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & r\sigma_1\sigma_2\\ r\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right).$$

Further, $Y|X, Z \sim N\left(\mu_1 + r\frac{\sigma_1}{\sigma_2}(Z - \mu_2), (1 - r^2)\sigma_1^2\right)$. Letting $\xi \sim N(0, 2\sigma_1^2)$, $\rho^2(Y, Z|X)$ can be simplified to:

$$\rho^2(Y, Z|X) = \frac{\mathbb{E}[h_3(\sqrt{1 - r^2}|\xi|)] - \mathbb{E}[h_3(|\xi|)]}{h_3(0) - \mathbb{E}[h_3(|\xi|)]},$$

which is strictly increasing in r^2 (as h_3 is strictly decreasing), if $\sigma_1^2 = \mathrm{Var}(Y|X)$ is fixed.

Note that $\rho^2(Y, Z|X) = 0$ if and only if $r = 0$, and $\rho^2(Y, Z|X) = 1$ if and only if $r^2 = 1$. For the linear kernel, $h_3(u) = -u^2/2$ (here $u \geq 0$), and therefore $\rho^2(Y, Z|X) = r^2$. ■

B.5 Proof of Proposition 2

Lemma 6 *Suppose $Y \overset{d}{=} -Y$ and ξ is an independent noise which is symmetric and unimodal about 0. Fix $c \geq 0$, $\mathbb{P}(|\lambda Y + \xi| \leq c)$ is a nonincreasing function of λ, for $\lambda \in [0, \infty)$.*

Proof It suffices to show that for any $c \geq 0$, (a) $\mathbb{P}(|\xi| \leq c) \geq \mathbb{P}(|Y + \xi| \leq c)$, and (b) for $\lambda > 1$, $\mathbb{P}(|\lambda Y + \xi| \leq c) \leq \mathbb{P}(|Y + \xi| \leq c)$. Note that

$$\mathbb{P}(|Y + \xi| \leq c) = \mathbb{P}(Y = 0)\mathbb{P}(\xi \in [-c, c]) + 2\int_{(0,\infty)} \mathbb{P}(\xi \in [y - c, y + c])\mathrm{d}P_Y(y),$$

where we have used the fact $\mathbb{P}(\xi \in [-y - c, -y + c]) = \mathbb{P}(\xi \in [y - c, y + c])$, and that Y has a symmetric distribution. Note that $h(y) := \mathbb{P}(\xi \in [y - c, y + c])$ is a nonincreasing function on $[0, \infty)$. Hence,

$$\begin{aligned}\mathbb{P}(|Y + \xi| \leq c) &= \mathbb{P}(Y = 0)\mathbb{P}(\xi \in [-c, c]) + 2\mathbb{E}[h(Y)1_{Y>0}]\\ &\geq \mathbb{P}(Y = 0)\mathbb{P}(\xi \in [-c, c]) + 2\mathbb{E}[h(\lambda Y)1_{Y>0}] = \mathbb{P}(|\lambda Y + \xi| \leq c).\end{aligned}$$

Also, $\mathbb{P}(|Y + \xi| \leq c) \leq \mathbb{P}(Y = 0)\mathbb{P}(\xi \in [-c, c]) + 2\mathbb{E}[h(0)1_{Y>0}] = \mathbb{P}(|\xi| \leq c)$. These complete the proof of the lemma. ■

Now let us prove Proposition 2. Recall the notation $\xi \sim \epsilon - \epsilon'$. Under the assumption of Proposition 2,

$$\rho^2(Y, Z|X) = \frac{\mathbb{E}[h_3(|\xi|)] - \mathbb{E}\left[\mathbb{E}[h_3\left(|\lambda[f(X, Z_1) - f(X, Z_2)] + \xi|\right)|X]\right]}{h_3(0) - \mathbb{E}\left[\mathbb{E}[h_3\left(|\lambda[f(X, Z_1) - f(X, Z_2)] + \xi|\right)|X]\right]},$$

By Lemma 6 applied to $Y = f(X, Z_1) - f(X, Z_2)$, we know that conditional on X, $|\lambda_1[f(X, Z_1) - f(X, Z_2)] + \xi|$ is stochastically less than $|\lambda_2[f(X, Z_1) - f(X, Z_2)] + \xi|$ whenever $\lambda_1 \leq \lambda_2$. Since h_3 is a decreasing function, $\mathbb{E}\left[\mathbb{E}[h_3\left(|\lambda[f(X, Z_1) - f(X, Z_2)] + \xi|\right)|X]\right]$ is decreasing in λ, and hence $\rho^2(Y, Z|X)$ is increasing in λ. ■

B.6 Proof of Proposition 8

From the form of $\rho^2(Y, Z|X)$ in Lemma 2 we see that $\rho^2(Y, Z|X)$ does not change by bijectively transforming X and Z.

With the kernel given in (8) and using the notation in Lemma 2,

$$\rho^2(Y, Z|X) = \frac{\mathbb{E}h_3(\|Y_2 - Y_2'\|) - \mathbb{E}h_3(\|Y_1 - Y_1'\|)}{h_3(0) - \mathbb{E}h_3(\|Y_1 - Y_1'\|)}.$$

Hence replacing Y_i, Y_i', for $i = 1, 2$, by $OY_i + b, OY_i' + b$, where O is an orthogonal matrix and b is a vector, does not change $\rho^2(Y, Z|X)$. This shows the desired invariance.

Next we show the continuity result. By Skorokhod's representation theorem, we can assume the convergence in distribution is actually a.s. convergence. The boundedness of $1 + \varepsilon$ moments in our assumption implies uniform integrability for large n. Together with the a.s. convergence, we have L_1 convergence, so $\rho^2(Y_n, Z_n|X_n) \to \rho^2(Y, Z|X)$ follows. ■

B.7 Proof of Lemma 3

We first show the general case, where $k_{\mathcal{Y}}(y_1, y_2) := \int 1_{y_1 \geq t} 1_{y_2 \geq t} \mathrm{d}P_Y(t)$, $y_1, y_2 \in \mathbb{R}$. We use the equivalent formulation (5) to calculate ρ^2. For Y_2, Y_2' as in (7),

$$\mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_2, Y_2')|X, Z]\right] = \mathbb{E}\left[\mathbb{E}\left[\int 1_{Y_2 \geq t} 1_{Y_2' \geq t} \mathrm{d}P_Y(t) \Big| X, Z\right]\right] = \mathbb{E}\left[\int \mathbb{P}\left(Y \geq t|X, Z\right)^2 \mathrm{d}P_Y(t)\right]. \tag{27}$$

Likewise, $\mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]\right] = \mathbb{E}\left[\int \mathbb{P}\left(Y \geq t|X\right)^2 dP_Y(t)\right]$. For $t \in \mathbb{R}$, let $g_t(X, Z) := \mathbb{P}\left(Y \geq t|X, Z\right)$. Then, we have,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_2, Y_2')|X, Z]\right] - \mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]\right] = \int \mathbb{E}\left[g_t(X, Z)^2 - (\mathbb{E}[g_t(X, Z)|X])^2\right] \mathrm{d}P_Y(t) \\
&= \int \mathbb{E}\left[\mathrm{Var}(g_t(X, Z)|X)\right] \mathrm{d}P_Y(t) = \int \mathbb{E}\left[\mathrm{Var}(\mathbb{P}(Y \geq t|X, Z)|X)\right] \mathrm{d}P_Y(t),
\end{aligned}$$

which is exactly the numerator of $T(Y, Z|X)$. The denominator of ρ^2 in (5) is

$$\begin{aligned}
&\mathbb{E}[k_{\mathcal{Y}}(Y, Y)] - \mathbb{E}[\mathbb{E}[k_{\mathcal{Y}}(Y_1, Y_1')|X]] = \mathbb{E}\left[\int 1_{Y \geq t}^2 \mathrm{d}P_Y(t)\right] - \mathbb{E}\left[\int \mathbb{P}\left(Y \geq t|X\right)^2 \mathrm{d}P_Y(t)\right] \\
&= \int \mathbb{E}\left[1_{Y \geq t}^2 - \mathbb{P}\left(Y \geq t|X\right)^2\right] \mathrm{d}P_Y(t) = \int \mathbb{E}\left[\mathrm{Var}(1_{Y \geq t}|X)\right] \mathrm{d}P_Y(t),
\end{aligned}$$

which coincides with the denominator of $T(Y,Z|X)$.

Now, suppose Y is continuous, and $k_{\mathcal{Y}}(y,y') := \frac{1}{2}\left(|y|+|y'|-|y-y'|\right) = \min\{y,y'\} = \int_0^1 1_{y\geq t}1_{y'\geq t}\mathrm{d}t$ for all $y,y'\in[0,1]$. The right-inverse $F_Y^{-1}(t) := \inf\{y : F_Y(y)\geq t\}$ is non-decreasing and satisfies $F_Y^{-1}(F_Y(Y)) \overset{a.s.}{=} Y$. Let Y' be an independent copy of Y. Then,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(F_Y(Y_2),F_Y(Y_2'))|X,Z]\right] &= \mathbb{E}\left[\mathbb{E}\left[\int_0^1 1_{F_Y(Y_2)\geq t}1_{F_Y(Y_2')\geq t}\mathrm{d}t\middle| X,Z\right]\right] \\
&= \mathbb{E}\left[\int_0^1 \mathbb{P}\left(F_Y(Y)\geq t|X,Z\right)^2\mathrm{d}t\right] = \mathbb{E}\left[\mathbb{P}\left(F_Y(Y)\geq F_Y(Y')|X,Z\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(Y\geq Y'|X,Z\right)^2\right] = \mathbb{E}\left[\int \mathbb{P}\left(Y\geq t|X,Z\right)^2\mathrm{d}P_Y(t)\right].
\end{aligned}$$

The rest of the proof is the same as that after (27). ■

B.8 Proof of Proposition 3

By Deb et al. (2020, Proposition 3.1), both $\sqrt{n}\left(\frac{1}{n}\sum_i\sum_{(i,j)\in\mathcal{E}(\mathcal{G}_n^{\ddot{X}})}\frac{k_{\mathcal{Y}}(Y_i,Y_j)}{d_i^{\ddot{X}}} - \mathbb{E}k_{\mathcal{Y}}(Y_1,Y_{\ddot{N}(1)})\right)$ and $\sqrt{n}\left(\frac{1}{n}\sum_i\sum_{(i,j)\in\mathcal{E}(\mathcal{G}_n^{X})}\frac{k_{\mathcal{Y}}(Y_i,Y_j)}{d_i^{X}} - \mathbb{E}k_{\mathcal{Y}}(Y_1,Y_{N(1)})\right)$ are $O_p(1)$.

Since we also have $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n k_{\mathcal{Y}}(Y_i,Y_i) - \mathbb{E}k_{\mathcal{Y}}(Y_1,Y_1)\right) = O_p(1)$, the following holds:

$$\sqrt{n}\left(\hat{\rho}^2(Y,Z|X) - \frac{\mathbb{E}k_{\mathcal{Y}}(Y_1,Y_{\ddot{N}(1)}) - \mathbb{E}k_{\mathcal{Y}}(Y_1,Y_{N(1)})}{\mathbb{E}k_{\mathcal{Y}}(Y_1,Y_1) - \mathbb{E}k_{\mathcal{Y}}(Y_1,Y_{N(1)})}\right) = O_p(1).$$

■

B.9 Proof of Theorem 4

Let

$$\begin{aligned}
Q_n &:= \frac{1}{n}\sum_{i=1}^n \sum_{(i,j)\in\mathcal{E}(\mathcal{G}_n^{\ddot{X}})}\frac{k_{\mathcal{Y}}(Y_i,Y_j)}{d_i^{\ddot{X}}} - \frac{1}{n}\sum_{i=1}^n \sum_{(i,j)\in\mathcal{E}(\mathcal{G}_n^{X})}\frac{k_{\mathcal{Y}}(Y_i,Y_j)}{d_i^{X}}, \\
S_n &:= \frac{1}{n}\sum_{i=1}^n k_{\mathcal{Y}}(Y_i,Y_i) - \frac{1}{n}\sum_{i=1}^n \sum_{(i,j)\in\mathcal{E}(\mathcal{G}_n^{X})}\frac{k_{\mathcal{Y}}(Y_i,Y_j)}{d_i^{X}},
\end{aligned}$$

and their population limits be

$$Q := \mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_2,Y_2')|X,Z]\right] - \mathbb{E}\left[\mathbb{E}[k_{\mathcal{Y}}(Y_1,Y_1')|X]\right], \quad S := \mathbb{E}[k_{\mathcal{Y}}(Y,Y)] - \mathbb{E}[\mathbb{E}[k_{\mathcal{Y}}(Y_1,Y_1')|X]].$$

Then $\hat{\rho^2}(Y,Z|X) = \frac{Q_n}{S_n}$, and $\rho^2(Y,Z|X) = \frac{Q}{S}$. From the proof of Deb et al. (2020, Theorem 5.1) and Deb et al. (2020, Corollary 5.1)[24],

$$\frac{1}{n}\sum_i \sum_{(i,j)\in\mathcal{E}(\mathcal{G}_n^{\ddot{X}})}\frac{k_{\mathcal{Y}}(Y_i,Y_j)}{d_i^{\ddot{X}}} \quad \text{and} \quad \frac{1}{n}\sum_i \sum_{(i,j)\in\mathcal{E}(\mathcal{G}_n^{X})}\frac{k_{\mathcal{Y}}(Y_i,Y_j)}{d_i^{X}}$$

24. Note that with the notation in Deb et al. (2020), by Assumption 5 on the intrinsic dimensionality of X, we have $N(\mu_X, B(x^*,t),\varepsilon,0) \leq C_1(t/\varepsilon)^d$, and t_n/K_n being bounded is implied by Assumption 6. Hence the same argument in Deb et al. (2020, Section 5.1) works through.

are within $O_p\left(\sqrt{\nu_n}\right)$ distance of their theoretical limits. Since $\frac{1}{n}\sum_{i=1}^n k_{\mathcal{Y}}(Y_i, Y_i)$ is also within $O_p(n^{-1/2}) \lesssim O_p\left(\sqrt{\nu_n}\right)$ distance from its theoretical limit, Q_n (resp. S_n) is within $O_p\left(\sqrt{\nu_n}\right)$ distance to Q (resp. S). Hence:

$$\left|\hat{\rho}^2(Y, Z|X) - \rho^2(Y, Z|X)\right| = \left|\frac{S \cdot Q_n - Q \cdot S_n}{S \cdot S_n}\right| = \left|\frac{S \cdot O_p\left(\sqrt{\nu_n}\right) - Q \cdot O_p\left(\sqrt{\nu_n}\right)}{S \cdot S_n}\right| = O_p\left(\sqrt{\nu_n}\right).$$

The last equality follows as $S > 0$. ■

B.10 Proof of Proposition 5

We use the same notation as in Sheng and Sriperumbudur (2019). Define $S_X : \mathcal{H}_{\mathcal{X}} \to \mathbb{R}^n$ such that $S_X : f \mapsto (f(X_1), \ldots, f(X_n))^\top$. Then, for $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$, $\langle S_X f, \alpha\rangle_{\mathbb{R}^n} = \sum_i \alpha_i f(X_i) = \langle f, S_X^*\alpha\rangle_{\mathcal{H}_{\mathcal{X}}}$ where we have $S_X^* : \mathbb{R}^n \to \mathcal{H}_{\mathcal{X}}$ given by $S_X^* : \alpha \mapsto \sum_i \alpha_i k_{\mathcal{X}}(X_i, \cdot)$. Similarly define S_Y, S_Y^*, S_Z and S_Z^*. For $f \in \mathcal{H}_{\mathcal{X}}$,

$$\begin{aligned}\hat{C}_{YX} f &= \frac{1}{n}\sum_{i=1}^n k_{\mathcal{Y}}(Y_i, \cdot) f(X_i) - \left(\frac{1}{n}\sum_{i=1}^n k_{\mathcal{Y}}(Y_i, \cdot)\right)\left(\frac{1}{n}\sum_{i=1}^n f(X_i)\right) \\ &= \frac{1}{n} S_Y^* S_X f - \frac{1}{n^2} S_Y^* \mathbf{1}\mathbf{1}^\top S_X f = \frac{1}{n} S_Y^* (I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top) S_X f \;=\; \frac{1}{n} S_Y^* H S_X f.\end{aligned}$$

Hence $\hat{C}_{YX} = \frac{1}{n} S_Y^* H S_X$. Similarly, we have $\hat{C}_{Y\ddot{X}} = \frac{1}{n} S_Y^* H S_{\ddot{X}}$, $\hat{C}_X = \frac{1}{n} S_X^* H S_X$, $\hat{C}_{\ddot{X}} = \frac{1}{n} S_{\ddot{X}}^* H S_{\ddot{X}}$. Note that for all $\alpha \in \mathbb{R}^n$, $S_X S_X^* \alpha = \sum_{i=1}^n \alpha_i S_X k_{\mathcal{X}}(X_i, \cdot) = K_X \alpha$, where K_X is the kernel matrix with $(K_X)_{ij} = k_{\mathcal{X}}(X_i, X_j)$. Hence $S_X S_X^* = K_X$. Letting e_i denote the i-th unit vector in $\mathbb{R}^n$, for $i = 1, \ldots, n$, we have

$$\begin{aligned}&\hat{C}_{YX}(\hat{C}_X + \varepsilon I)^{-1}(k_{\mathcal{X}}(X_i, \cdot) - \hat{\mu}_X) = \hat{C}_{YX}(\hat{C}_X + \varepsilon I)^{-1} S_X^*(e_i - \frac{1}{n}\mathbf{1}) \\ &= \frac{1}{n} S_Y^* H S_X \left(\frac{1}{n} S_X^* H S_X + \varepsilon I\right)^{-1} S_X^* H e_i = \frac{1}{n} S_Y^* H S_X S_X^* \left(\frac{1}{n} H S_X S_X^* + \varepsilon I\right)^{-1} H e_i\end{aligned}$$

where we have used the fact that, for operators A and B, $(BA + \varepsilon I)^{-1} B = B(AB + \varepsilon I)^{-1}$, which holds by direct verification. Now, using $K_X = S_X S_X^*$, we can show that the right side of the above display equals

$$\begin{aligned}&S_Y^* H K_X \left(H K_X + n\varepsilon I\right)^{-1} H e_i = S_Y^* H K_X \left(H^2 K_X + n\varepsilon I\right)^{-1} H e_i \\ &= S_Y^* H K_X H \left(H K_X H + n\varepsilon I\right)^{-1} e_i = S_Y^* \tilde{K}_X \left(\tilde{K}_X + n\varepsilon I\right)^{-1} e_i,\end{aligned}$$

where $\tilde{K}_X := H K_X H$ is the centered kernel matrix. Similarly, we have

$$\hat{C}_{Y\ddot{X}}(\hat{C}_{\ddot{X}} + \varepsilon I)^{-1}\left(k_{\ddot{\mathcal{X}}}(\ddot{X}_i, \cdot) - \hat{\mu}_{\ddot{X}}\right) = S_Y^* \tilde{K}_{\ddot{X}} \left(\tilde{K}_{\ddot{X}} + n\varepsilon I\right)^{-1} e_i.$$

Recalling that $M = \tilde{K}_X\left(\tilde{K}_X + n\varepsilon I\right)^{-1} - \tilde{K}_{\ddot{X}}\left(\tilde{K}_{\ddot{X}} + n\varepsilon I\right)^{-1}$, the numerator of $\tilde{\rho^2}$ reduces to $\sum_{i=1}^n \|S_Y^* M e_i\|_{\mathcal{H}_{\mathcal{Y}}}^2$. Note that, for $\alpha \in \mathbb{R}^n$, $\|S_Y^*\alpha\|_{\mathcal{H}_{\mathcal{Y}}}^2 = \langle \sum_i \alpha_i k_{\mathcal{Y}}(Y_i, \cdot), \sum_j \alpha_j k_{\mathcal{Y}}(Y_j, \cdot)\rangle_{\mathcal{H}_{\mathcal{Y}}} = \sum_{i,j=1}^n \alpha_i \alpha_j k_{\mathcal{Y}}(Y_i, Y_j) = \alpha^\top K_Y \alpha$. Thus,

$$\sum_{i=1}^n \|S_Y^* M e_i\|_{\mathcal{H}_{\mathcal{Y}}}^2 = \sum_{i=1}^n e_i^\top M^\top K_Y M e_i = \mathrm{Tr}(M^\top K_Y M) = \mathrm{Tr}(M^\top \tilde{K}_Y M),$$

where we have used the fact that $H^2 = H$. Now the denominator of $\tilde{\rho^2}$ can be simplified as $\sum_{i=1}^n \|S_Y^* H e_i - S_Y^* \tilde{K}_X \left(\tilde{K}_X + n\varepsilon I\right)^{-1} e_i\|_{\mathcal{H}_\mathcal{Y}}^2$, where we have used the fact that $k_\mathcal{Y}(Y_i, \cdot) - \hat{\mu}_Y = S_Y^* H e_i$. Letting $N_0 := H - \tilde{K}_X \left(\tilde{K}_X + n\varepsilon I\right)^{-1}$, and recalling that $N = I - \tilde{K}_X \left(\tilde{K}_X + n\varepsilon I\right)^{-1}$, the above display can be expressed as

$$\sum_{i=1}^n \|S_Y^* N_0 e_i\|_{\mathcal{H}_\mathcal{Y}}^2 = \sum_{i=1}^n e_i^\top N_0^\top K_Y N_0 e_i = \mathrm{Tr}(N_0^\top K_Y N_0) = \mathrm{Tr}(N^\top \tilde{K}_Y N).$$

This proves the desired result. ■

B.11 Proof of Proposition 7

By assumption we have $K_Y = \mathbf{Y}\mathbf{Y}^\top$. As M, N are symmetric matrices,

$$\tilde{\rho_u^2} = \frac{\mathrm{Tr}(M^\top \mathbf{Y}\mathbf{Y}^\top M)}{\mathrm{Tr}(N^\top \mathbf{Y}\mathbf{Y}^\top N)} = \frac{\|M\mathbf{Y}\|^2}{\|N\mathbf{Y}\|^2} = \frac{\left\|K_X(K_X + n\varepsilon I)^{-1}\mathbf{Y} - K_{\ddot{X}}(K_{\ddot{X}} + n\varepsilon I)^{-1}\mathbf{Y}\right\|^2}{\|\mathbf{Y} - K_X(K_X + n\varepsilon I)^{-1}\mathbf{Y}\|^2}. \quad (28)$$

Kernel ridge regression yields $\hat{\mathbf{Y}} = K(K + \lambda I)^{-1}\mathbf{Y}$ for a generic kernel matrix and regularization parameter λ (see e.g., Kung, 2014, Section 7.3.4). Hence (28) reduces to $\tilde{\rho_u^2} = \frac{\|\hat{\mathbf{Y}}_x - \hat{\mathbf{Y}}_{xz}\|^2}{\|\hat{\mathbf{Y}}_x - \mathbf{Y}\|^2}$, with the regularization parameter $\rho = n\varepsilon$. ■

B.12 Proof of Proposition 6

The same argument as (28) shows that the centered estimator $\tilde{\rho^2}$ can be expressed as

$$\frac{\left\|\tilde{K}_X(\tilde{K}_X + n\varepsilon I)^{-1} H\mathbf{Y} - \tilde{K}_{\ddot{X}}(\tilde{K}_{\ddot{X}} + n\varepsilon I)^{-1} H\mathbf{Y}\right\|^2}{\|H\mathbf{Y} - \tilde{K}_X(\tilde{K}_X + n\varepsilon I)^{-1} H\mathbf{Y}\|^2} = \frac{\left\|\left[(\tilde{K}_X + n\varepsilon)^{-1} - (\tilde{K}_{\ddot{X}} + n\varepsilon I)^{-1}\right] H\mathbf{Y}\right\|^2}{\|(\tilde{K}_X + n\varepsilon I)^{-1} H\mathbf{Y}\|^2}.$$

With linear kernels, $K_X = \mathbf{X}\mathbf{X}^\top$ and $K_{\ddot{X}} = \ddot{\mathbf{X}}\ddot{\mathbf{X}}^\top$. Let $\mathbf{Y}_c = H\mathbf{Y}$, $\mathbf{X}_c = H\mathbf{X}$, $\ddot{\mathbf{X}}_c = H\ddot{\mathbf{X}}$ be the centered versions of $\mathbf{Y}$, $\mathbf{X}$, $\ddot{\mathbf{X}}$ by subtracting the mean from all columns. The empirical classical partial correlation first fits two linear regressions of $\mathbf{Y}$ on $\mathbf{X}$ and $\mathbf{Z}$ on $\mathbf{X}$ (an intercept term is added so the design matrix is $[\mathbf{1}_n\ \mathbf{X}]$), and then outputs the correlation of the two resulting residuals r_Y, r_Z. Therefore, the partial correlation remains unchanged when we replace $\mathbf{Y}, \mathbf{Z}$ by $\mathbf{Y}_c, \mathbf{Z}_c$. Without loss of generality, suppose $\mathbf{Y} = \mathbf{Y}_c$, $\mathbf{Z} = \mathbf{Z}_c$ have been centered. By the matrix identity $(\mathbf{X}_c\mathbf{X}_c^\top + n\varepsilon I)^{-1} = \frac{1}{n\varepsilon} I - \frac{1}{n\varepsilon}\mathbf{X}_c(n\varepsilon I + \mathbf{X}_c^\top\mathbf{X}_c)^{-1}\mathbf{X}_c^\top$ we have

$$\hat{\rho^2}(Y, Z|X) = \frac{\left\|\left[(\mathbf{X}_c\mathbf{X}_c^\top + n\varepsilon)^{-1} - (\ddot{\mathbf{X}}_c\ddot{\mathbf{X}}_c^\top + n\varepsilon I)^{-1}\right]\mathbf{Y}_c\right\|^2}{\|(\mathbf{X}_c\mathbf{X}_c^\top + n\varepsilon I)^{-1}\mathbf{Y}_c\|^2}$$

$$= \frac{\left\|\left[\mathbf{X}_c(n\varepsilon I + \mathbf{X}_c^\top\mathbf{X}_c)^{-1}\mathbf{X}_c^\top - \ddot{\mathbf{X}}_c(n\varepsilon I + \ddot{\mathbf{X}}_c^\top\ddot{\mathbf{X}}_c)^{-1}\ddot{\mathbf{X}}_c^\top\right]\mathbf{Y}_c\right\|^2}{\|(I - \mathbf{X}_c(n\varepsilon I + \mathbf{X}_c^\top\mathbf{X}_c)^{-1}\mathbf{X}_c^\top)\mathbf{Y}_c\|^2} \xrightarrow{\varepsilon \to 0} \frac{\|\mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c - \mathrm{Proj}_{\ddot{\mathbf{X}}_c}\mathbf{Y}_c\|^2}{\|\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c\|^2},$$

where $\mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c$ is the projection of $\mathbf{Y}_c$ onto the column space of $\mathbf{X}_c$. Note that r_Z is in the column space of $\ddot{\mathbf{X}}_c$ and is orthogonal to the column space of $\mathbf{X}_c$, and $\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c = r_Y$.

Hence $\mathrm{Cor}(r_Y, r_Z) = \mathrm{Cor}(\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c, r_Z) = \mathrm{Cor}(\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c, \mathrm{Proj}_{\ddot{\mathbf{X}}_c}\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c)$.
Note that $\big(\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c\big) - \Big(\mathrm{Proj}_{\ddot{\mathbf{X}}_c}\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c\Big)$ is orthogonal to $\mathrm{Proj}_{\ddot{\mathbf{X}}_c}\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c$.
Hence $\mathrm{Cor}(r_Y, r_Z)^2 = \mathrm{Cor}(\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c, \mathrm{Proj}_{\ddot{\mathbf{X}}_c}\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c)^2 = \frac{\|\mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c - \mathrm{Proj}_{\ddot{\mathbf{X}}_c}\mathbf{Y}_c\|^2}{\|\mathbf{Y}_c - \mathrm{Proj}_{\mathbf{X}_c}\mathbf{Y}_c\|^2}$. ■

B.13 Proofs of Theorems 5 and 6

We start with some preliminaries. The following result (Fukumizu et al., 2007, Lemma 5) are known that the cross-covariance operator C_{YX} and its sample version are close in the Hilbert-Schmidt norm and are consequently close in the operator norm, since $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_{\mathrm{HS}}$.

Lemma 7 *Suppose that $\mathbb{E}[k_{\mathcal{X}}(X,X)] < \infty$ and $\mathbb{E}[k_{\mathcal{Y}}(Y,Y)] < \infty$, and $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}$ are both separable. Then $\|\hat{C}_{YX} - C_{YX}\|_{\mathrm{HS}} = O_p(n^{-1/2})$.*

We divide the proof into several steps. Recall that the operator norm of a linear map $A : \mathcal{V} \to \mathcal{W}$ (for two given normed vector spaces $\mathcal{V}$ and $\mathcal{W}$) is defined as $\|A\|_{\mathrm{op}} := \inf\{c \ge 0 : \|Av\| \le c\|v\| \text{ for all } v \in \mathcal{V}\}$. For notational simplicity, $\|\cdot\|$ will denote either the norm in an RKHS or the operator norm.

Step 1. $\mathbb{E}\left[\|[(C_X^\dagger C_{XY})^* - C_{YX}(C_X + \varepsilon I)^{-1}](k_{\mathcal{X}}(X,\cdot) - \mu_X)\|_{\mathcal{H}_{\mathcal{Y}}}^2\right] \to 0$ as $\varepsilon \to 0^+$.

Denote by $A := C_X^\dagger C_{XY}$, which is a bounded linear operator by Lemma 4. Further, $\operatorname{ran} C_{XY} \subset \operatorname{ran} C_X$ and $C_X C_X^\dagger C_X = C_X$ imply $C_X A = C_{XY}$. Hence

$$\begin{aligned} & \mathbb{E}\left[\big\|[(C_X^\dagger C_{XY})^* - C_{YX}(C_X + \varepsilon I)^{-1}](k_{\mathcal{X}}(X,\cdot) - \mu_X)\big\|^2\right] \\ = \;& \mathbb{E}\left[\|[A^* - A^* C_X (C_X + \varepsilon I)^{-1}](k_{\mathcal{X}}(X,\cdot) - \mu_X)\|^2\right] \\ \le \;& \|A^*\| \cdot \mathbb{E}\left[\|[I - C_X(C_X + \varepsilon I)^{-1}](k_{\mathcal{X}}(X,\cdot) - \mu_X)\|^2\right] \end{aligned}$$

where in the last display we have used the definition of the operator norm. Let $\{e_j\}_{j\ge1}$ be an eigenbasis of C_X with corresponding eigenvalues $\{\lambda_j\}_{j\ge1}$. Since C_X is trace-class[25], $\sum_j \lambda_j < \infty$. Then, $I - C_X(C_X + \varepsilon I)^{-1}$ has eigenbasis $\{e_j\}_{j\ge1}$ with corresponding eigenvalues $\{\varepsilon(\varepsilon + \lambda_j)^{-1}\}_{j\ge1}$. Thus,

$$\mathbb{E}\left[\big\|[I - C_X(C_X + \varepsilon I)^{-1}](k_{\mathcal{X}}(X,\cdot) - \mu_X)\big\|^2\right] = \mathbb{E}\left[\Big\|\sum_j \frac{\varepsilon}{\varepsilon + \lambda_j}\langle k_{\mathcal{X}}(X,\cdot) - \mu_X, e_j\rangle e_j\Big\|^2\right]$$

$$= \mathbb{E}\left[\sum_j \frac{\varepsilon^2 \langle k_{\mathcal{X}}(X,\cdot) - \mu_X, e_j\rangle^2}{(\varepsilon + \lambda_j)^2}\right] = \sum_j \frac{\varepsilon^2 \mathrm{Var}(e_j(X))}{(\varepsilon + \lambda_j)^2} = \sum_j \frac{\varepsilon^2 \lambda_j}{(\varepsilon + \lambda_j)^2}.$$

It is easily seen that the above quantity converges to 0 as $\varepsilon \to 0^+$.

Step 2. $\frac{1}{n}\sum_{i=1}^n \big\|[(C_X^\dagger C_{XY})^* - C_{YX}(C_X + \varepsilon I)^{-1}](k_{\mathcal{X}}(X_i,\cdot) - \mu_X)\big\|^2 \xrightarrow{p} 0$ as $\varepsilon \to 0^+$.

This is a direct consequence of Step 1 and Markov's inequality.

Step 3. $\frac{1}{n}\sum_{i=1}^n \big\|[\hat{C}_{YX}(\hat{C}_X + \varepsilon_n I)^{-1} - C_{YX}(C_X + \varepsilon_n I)^{-1}](k_{\mathcal{X}}(X_i,\cdot) - \hat{\mu}_X)\big\|^2 \xrightarrow{p} 0$.

This is the only step where we use $\varepsilon_n n^{1/2} \to \infty$. We will use the following simple result.

25. $\sum_i \langle C_X e_i, e_i\rangle_{\mathcal{H}_{\mathcal{X}}} = \sum_i \mathrm{Var}(e_i(X)) \le \sum_i \mathbb{E}[e_i(X)^2] = \sum_i \mathbb{E}\left[\langle e_i, k_{\mathcal{X}}(X,\cdot)\rangle_{\mathcal{H}_{\mathcal{X}}}^2\right] = \mathbb{E}[\|k_{\mathcal{X}}(X,\cdot)\|_{\mathcal{H}_{\mathcal{X}}}^2] < \infty$. The last equality follows from Parseval's identity.

Lemma 8 *Suppose that* $\{U_i^{(n)} : 1 \le i \le n\}_{n\ge 1}$ *and* $\{V_i^{(n)} : 1 \le i \le n\}_{n\ge 1}$ *are random elements taking values in a Hilbert space with norm* $\|\cdot\|$. *If* $\frac{1}{n}\sum_{i=1}^n \|U_i^{(n)} - V_i^{(n)}\|^2 \xrightarrow{p} 0$ *and* $\frac{1}{n}\sum_{i=1}^n \|U_i^{(n)}\|^2 \xrightarrow{p} U$, *then* $\frac{1}{n}\sum_{i=1}^n \|V_i^{(n)}\|^2 \xrightarrow{p} U$.

Proof Since

$$\|V_i^{(n)}\|^2 \le \|U_i^{(n)}\|^2 + 2\|U_i^{(n)}\| \cdot \|U_i^{(n)} - V_i^{(n)}\| + \|U_i^{(n)} - V_i^{(n)}\|^2,$$
$$\|V_i^{(n)}\|^2 \ge \|U_i^{(n)}\|^2 - 2\|U_i^{(n)}\| \cdot \|U_i^{(n)} - V_i^{(n)}\| + \|U_i^{(n)} - V_i^{(n)}\|^2,$$

it suffices to show that $\frac{1}{n}\sum_{i=1}^n 2\|U_i^{(n)}\| \cdot \|U_i^{(n)} - V_i^{(n)}\| \xrightarrow{p} 0$. Let $\delta > 0$.

$$\frac{1}{n}\sum_{i=1}^n 2\|U_i^{(n)}\| \cdot \|U_i^{(n)} - V_i^{(n)}\| \le \frac{1}{n}\sum_{i=1}^n \left[\delta\|U_i^{(n)}\|^2 + \frac{1}{\delta}\|U_i^{(n)} - V_i^{(n)}\|^2\right] \xrightarrow{p} \delta U.$$

Since $\delta > 0$ is arbitrary, this concludes the proof. ■

For $i = 1, \ldots, n$, let

$$\begin{aligned} V_i^{(n)} &:= \left[\hat{C}_{YX}(\hat{C}_X + \varepsilon_n I)^{-1} - C_{YX}(C_X + \varepsilon_n I)^{-1}\right](k_{\mathcal{X}}(X_i, \cdot) - \hat{\mu}_X) \\ &= \left[(\hat{C}_{YX} - C_{YX})(\hat{C}_X + \varepsilon_n I)^{-1} + C_{YX}\left((\hat{C}_X + \varepsilon_n I)^{-1} - (C_X + \varepsilon I)^{-1}\right)\right](k_{\mathcal{X}}(X_i, \cdot) - \hat{\mu}_X). \end{aligned}$$

Letting $U_i^{(n)} := (\hat{C}_{YX} - C_{YX})(\hat{C}_X + \varepsilon_n I)^{-1}(k_{\mathcal{X}}(X_i, \cdot) - \hat{\mu}_X)$, for $i = 1, \ldots, n$, note that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \|U_i^{(n)}\|^2 &= \frac{1}{n}\sum_{i=1}^n \left\|(\hat{C}_{YX} - C_{YX})(\hat{C}_X + \varepsilon_n I)^{-1}(k_{\mathcal{X}}(X_i, \cdot) - \hat{\mu}_X)\right\|^2 \\ &\le \frac{1}{n}\sum_{i=1}^n \|\hat{C}_{YX} - C_{YX}\|^2 \cdot \|(\hat{C}_X + \varepsilon_n I)^{-1}\|^2 \cdot \|k_{\mathcal{X}}(X_i, \cdot) - \hat{\mu}_X\|^2 \\ &\le O_p\left(\frac{1}{n}\right) \cdot \frac{1}{\varepsilon_n^2} \cdot \frac{1}{n}\sum_{i=1}^n \|k_{\mathcal{X}}(X_i, \cdot) - \hat{\mu}_X\|^2 = O_p\left(\frac{1}{n\varepsilon_n^2}\right) = o_p(1), \end{aligned}$$

where we have used Lemma 7, $\|(\hat{C}_X + \varepsilon_n)^{-1}\| \le \frac{1}{\varepsilon_n}$ (since if $\hat{C}_X$ has eigenvalues $\hat{\lambda}_i$, then $(\hat{C}_X + \varepsilon_n)^{-1}$ has eigenvalues $\frac{1}{\hat{\lambda}_i + \varepsilon_n} \le \frac{1}{\varepsilon_n}$), and that $n\varepsilon_n^2 \to \infty$.

Thus, in view of Lemma 8, we only need to show that

$$\frac{1}{n}\sum_{i=1}^n \|V_i^{(n)} - U_i^{(n)}\|^2 = \frac{1}{n}\sum_{i=1}^n \left\|C_{YX}\left((\hat{C}_X + \varepsilon_n I)^{-1} - (C_X + \varepsilon_n I)^{-1}\right)(k_{\mathcal{X}}(X_i, \cdot) - \hat{\mu}_X)\right\|^2 \xrightarrow{p} 0.$$

Using $B^{-1} - C^{-1} = C^{-1}(C - B)B^{-1}$, and $A = C_X^{\dagger}C_{XY}$, $C_{YX} = A^*C_X$ as before,

$$\begin{aligned} &\left\|C_{YX}\left((\hat{C}_X + \varepsilon_n I)^{-1} - (C_X + \varepsilon_n I)^{-1}\right)\right\| = \left\|C_{YX}(C_X + \varepsilon_n I)^{-1}(C_X - \hat{C}_X)(\hat{C}_X + \varepsilon_n I)^{-1}\right\| \\ &\le \left\|A^*C_X(C_X + \varepsilon_n I)^{-1}\right\| \cdot \|C_X - \hat{C}_X\| \cdot \left\|(\hat{C}_X + \varepsilon_n I)^{-1}\right\| \\ &\le \|A^*\| \cdot \|C_X(C_X + \varepsilon_n I)^{-1}\| \cdot O_p(n^{-1/2}) \cdot \frac{1}{\varepsilon_n} = O_p\left(\frac{1}{\varepsilon_n\sqrt{n}}\right) = o_p(1). \end{aligned}$$

This implies that

$$\frac{1}{n}\sum_{i=1}^{n}\|V_i^{(n)}-U_i^{(n)}\|^2 \le \left\|C_{YX}\left((\hat{C}_X+\varepsilon_n I)^{-1}-(C_X+\varepsilon_n I)^{-1}\right)\right\|\cdot\frac{1}{n}\sum_{i=1}^{n}\|k_{\mathcal{X}}(X_i,\cdot)-\hat{\mu}_X\|^2 \xrightarrow{p} 0.$$

Now this step follows from Lemma 8.

Step 4. $\frac{1}{n}\sum_{i=1}^{n}\left\|C_{YX}(C_X+\varepsilon_n I)^{-1}(k_{\mathcal{X}}(X_i,\cdot)-\hat{\mu}_X)-C_{YX}(C_X+\varepsilon_n I)^{-1}(k_{\mathcal{X}}(X_i,\cdot)-\mu_X)\right\|^2 \xrightarrow{p} 0.$

This follows as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|C_{YX}(C_X+\varepsilon_n I)^{-1}(k_{\mathcal{X}}(X_i,\cdot)-\hat{\mu}_X)-C_{YX}(C_X+\varepsilon_n I)^{-1}(k_{\mathcal{X}}(X_i,\cdot)-\mu_X)\|^2\\
&=\frac{1}{n}\sum_{i=1}^{n}\|C_{YX}(C_X+\varepsilon_n I)^{-1}(\hat{\mu}_X-\mu_X)\|^2 \;\le\; \|A^*\|\cdot\|C_X(C_X+\varepsilon_n I)^{-1}\|\cdot\frac{1}{n}\sum_{i=1}^{n}\|\hat{\mu}_X-\mu_X\|^2\\
&\le \|A^*\|\cdot 1\cdot\|\hat{\mu}_X-\mu_X\|^2 \xrightarrow{a.s.} 0,
\end{aligned}$$

by the strong law of large numbers (Hoffmann-Jørgensen and Pisier, 1976).

Now let us show Theorem 5. By successive applications of Lemma 8,

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|\mu_{Y|X_i}-\hat{\mu}_{Y|X_i}\|_{\mathcal{H}_\mathcal{Y}}^2\\
&=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|\mu_Y+(C_X^\dagger C_{XY})^*\left(k_\mathcal{Y}(X_i,\cdot)-\mu_X\right)-\hat{\mu}_Y-\hat{C}_{YX}(\hat{C}_X+\varepsilon_n I)^{-1}\left(k_\mathcal{Y}(X_i,\cdot)-\hat{\mu}_X\right)\|_{\mathcal{H}_\mathcal{Y}}^2\\
&=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|(C_X^\dagger C_{XY})^*\left(k_\mathcal{Y}(X_i,\cdot)-\mu_X\right)-\hat{C}_{YX}(\hat{C}_X+\varepsilon_n I)^{-1}\left(k_\mathcal{Y}(X_i,\cdot)-\hat{\mu}_X\right)\|_{\mathcal{H}_\mathcal{Y}}^2\\
&\overset{\text{Step 3}}{=}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|(C_X^\dagger C_{XY})^*\left(k_\mathcal{Y}(X_i,\cdot)-\mu_X\right)-C_{YX}(C_X+\varepsilon_n I)^{-1}\left(k_\mathcal{Y}(X_i,\cdot)-\hat{\mu}_X\right)\|_{\mathcal{H}_\mathcal{Y}}^2\\
&\overset{\text{Step 4}}{=}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|(C_X^\dagger C_{XY})^*\left(k_\mathcal{Y}(X_i,\cdot)-\mu_X\right)-C_{YX}(C_X+\varepsilon_n I)^{-1}\left(k_\mathcal{Y}(X_i,\cdot)-\mu_X\right)\|_{\mathcal{H}_\mathcal{Y}}^2\\
&\overset{\text{Step 2}}{=}0,
\end{aligned}$$

where the limit is in probability, and in the second equality we have used $\frac{1}{n}\sum_{i=1}^{n}\|\mu_Y-\hat{\mu}_Y\|_{\mathcal{H}_\mathcal{Y}}^2=\|\mu_Y-\hat{\mu}_Y\|_{\mathcal{H}_\mathcal{Y}}^2\xrightarrow{p}0$.

Then we can use Theorem 5 and Lemma 8 to show Theorem 6. By the expression of $\tilde{\rho^2}$ in (18), we see that the numerator

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|\hat{\mu}_{Y|\ddot{X}_i}-\hat{\mu}_{Y|X_i}\|_{\mathcal{H}_\mathcal{Y}}^2=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|\mu_{Y|\ddot{X}_i}-\mu_{Y|X_i}\|_{\mathcal{H}_\mathcal{Y}}^2=\mathbb{E}[\|\mu_{Y|\ddot{X}}-\mu_{Y|X}\|_{\mathcal{H}_\mathcal{Y}}^2],$$

and the denominator

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|k_\mathcal{Y}(Y_i,\cdot)-\hat{\mu}_{Y|X_i}\|_{\mathcal{H}_\mathcal{Y}}^2=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|k_\mathcal{Y}(Y_i,\cdot)-\mu_{Y|X_i}\|_{\mathcal{H}_\mathcal{Y}}^2=\mathbb{E}[\|k_\mathcal{Y}(Y,\cdot)-\mu_{Y|X}\|_{\mathcal{H}_\mathcal{Y}}^2].$$

Both limits are in probability, and they are exactly the numerator and denominator of $\rho^2(Y, Z|X)$ respectively. ■

B.14 Proof of Theorem 7

The proof is similar to that of Azadkia and Chatterjee (2021, Theorem 6.1). Let $j_1, \ldots, j_p$ be the complete ordering of all variables produced by the algorithm without imposing the stopping rule. Let $S_k := \{j_1, \ldots, j_k\}$ for $1 \leq k \leq p$, $S_0 := \emptyset$, $S_k := S_p$ for $k > p$, and recall that $\kappa = \lfloor \frac{M}{\delta} + 1 \rfloor$. Let $\varepsilon_1, \varepsilon_2 > 0$ be small such that $((1-\varepsilon_2)\delta - 2\varepsilon_1)\lfloor \frac{M}{\delta} + 1 \rfloor > M$ and $(1-\varepsilon_2)\delta + 2\varepsilon_1 < \delta$. Note that $\varepsilon_1, \varepsilon_2$ only depend on δ and M. Define: Event E_0: S_κ is sufficient. Event E: $|T_n(S_k) - T(S_k)| \leq \varepsilon_1$ for $1 \leq k \leq \kappa$. We will show that E implies the selected subset is sufficient and $\mathbb{P}(E)$ is large.

Lemma 9 *Suppose E happens, and for some $1 \leq k \leq \kappa$*

$$T_n(S_k) - T_n(S_{k-1}) \leq (1-\varepsilon_2)\delta. \tag{29}$$

Then S_{k-1} is sufficient.

Proof If $k > p$, then there is nothing to prove. Suppose $k \leq p$. Since the algorithm chooses $j_k \in \{1, \ldots p\} \backslash S_{k-1}$ to maximize $T_n(S_{k-1} \cup \{j\})$, we have for all $j \in \{1, \ldots p\} \backslash S_{k-1}$: $T(S_{k-1} \cup \{j\}) - T(S_{k-1}) \leq T_n(S_{k-1} \cup \{j\}) - T_n(S_{k-1}) + 2\varepsilon_1 \leq T_n(S_k) - T_n(S_{k-1}) + 2\varepsilon_1 \leq (1-\varepsilon_2)\delta + 2\varepsilon_1 < \delta$. By the definition of δ, S_{k-1} is sufficient. ■

Lemma 10 *Event E implies E_0.*

Proof Suppose E holds. If (29) holds for some $1 \leq k \leq \kappa$, then by Lemma 9, S_{k-1} is sufficient and E_0 holds. Suppose (29) is violated for all $1 \leq k \leq \kappa$. Then $T(S_k) - T(S_{k-1}) \geq T_n(S_k) - T_n(S_{k-1}) - 2\varepsilon_1 > (1-\varepsilon_2)\delta - 2\varepsilon_1$. Hence, $T(S_\kappa) = \sum_{k=1}^{\kappa}(T(S_k) - T(S_{k-1})) + T(S_0) \geq \kappa \cdot ((1-\varepsilon_2)\delta - 2\varepsilon_1) + 0 > M$, by the constuction of $\varepsilon_1, \varepsilon_2$. This yields a contradiction since $T(S_\kappa)$ cannot be greater than M, the bound of the kernel. ■

Lemma 11 *Event E implies that $\hat{S}$ is sufficient.*

Proof If the algorithm stopped after κ, then $S_\kappa \subset \hat{S}$. By Lemma 10, E happens implies E_0 happens, i.e., S_κ is sufficient. Hence $\hat{S}$ is also sufficient. If the algorithm stopped at $k < \kappa$, by the stopping rule $T_n(S_{k+1}) < T_n(S_k)$, so (29) holds, and by Lemma 9, S_k is sufficient. ■

Lemma 12 *There exist L_1, L_2 depending only on $\alpha, \beta_1, \beta_2, \gamma, \{C_i\}_{i=1}^6, d, M, \delta$ such that $\mathbb{P}(E) \geq 1 - L_1 p^\kappa e^{-L_2 n}$.*

Proof By Deb et al. (2020, Equation C.38) with the notation therein and Deb et al. (2020, Section 5.1), there exists $\xi_1, \xi_2, \xi_3 > 0$ depending on $\alpha, \beta_1, \beta_2, \gamma, \{C_i\}_{i=1}^6, d, M$ such that for any S of size $\leq \kappa$

$$|\mathbb{E}T_n(S) - T(S)| \leq \xi_1 \left(\varepsilon_n^{\beta_2} + \sqrt{\nu_{1,n}} \right) \leq \xi_1 \frac{(\log n)^{\xi_2}}{n^{\xi_3}}.$$

By Deb et al. (2020, Proposition 3.1), there exists C^* depending on C_2 and M such that for any S of size $\leq \kappa$, $\mathbb{P}\left(|T_n(S) - \mathbb{E}T(S)| \geq t\right) \leq 2\exp\left(-C^* n t^2\right)$. Hence,

$$\mathbb{P}\left(|T_n(S) - T(S)| \geq \xi_1 \frac{(\log n)^{\xi_2}}{n^{\xi_3}} + t\right) \leq 2e^{-C^* n t^2}.$$

By a union bound:

$$\mathbb{P}\left(\bigcup_{|S|\leq\kappa} \left\{|T_n(S) - T(S)| \geq \xi_1 \frac{(\log n)^{\xi_2}}{n^{\xi_3}} + t\right\}\right) \leq 2p^\kappa e^{-C^* n t^2}.$$

Let $t = \frac{\varepsilon_1}{2}$. For large n, say $n \geq n_0$, $\xi_1 \frac{(\log n)^{\xi_2}}{n^{\xi_3}} < \frac{\varepsilon_1}{2}$ so $\mathbb{P}(E) \geq 1 - 2p^\kappa e^{-L_2 n}$. We can adjust the constant 2 (again depending only on $\alpha, \beta_1, \beta_2, \gamma, \{C_i\}_{i=1}^6, d, M, \delta$) so that the above inequality also holds for small n. ■

Combining the previous two lemmas we obtain the proof of Theorem 7. ■

References

J. Aitchison. *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

Vassilis Athitsos, Jonathan Alon, Stan Sclaroff, and George Kollios. Boostmap: A method for efficient approximate similarity rankings. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 2004.

Jean-Pierre Aubin. *Applied Functional Analysis*, volume 47. John Wiley & Sons, 2011.

Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *Ann. Statist.*, 49(6):3070–3102, 2021.

Mona Azadkia, Sourav Chatterjee, and Norman Matloff. *FOCI: Feature Ordering by Conditional Independence*, 2020. R package version 0.1.2.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3(1):1–48, 2003.

Charles R. Baker. Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186: 273–289, 1973.

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085, 2015.

R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

Wicher Pieter Bergsma. *Testing Conditional Independence for Continuous Random Variables*. Eurandom, 2004.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, MA, 2004.

Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 97–104, 2006.

Bhaswar B. Bhattacharya. A general asymptotic framework for distribution-free graph-based two-sample tests. *J. Roy. Statist. Soc. Ser. B*, 81(3):575–602, 2019.

Howard D. Bondell and Lexin Li. Shrinkage inverse regression estimation for model-free variable selection. *J. Roy. Statist. Soc. Ser. B*, 71(1):287–299, 2009.

Leonid Boytsov and Bilegsaikhan Naidan. Learning to prune in metric and non-metric spaces. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2013.

Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

Kevin Buchin and Wolfgang Mulzer. Delaunay triangulations in O(sort(n)) time and more. *Journal of the ACM*, 58(2):1–27, 2011.

Emmanuel Candès and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. Roy. Statist. Soc. Ser. B*, 80(3): 551–577, 2018.

Sourav Chatterjee. A new coefficient of correlation. *J. Amer. Statist. Assoc.*, 116(536):2009–2022, 2021.

Eric Z Chen and Hongzhe Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617, 2016.

George H. Chen and Devavrat Shah. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends in Machine Learning*, 10(5-6):337–588, 2018.

Jiajia Chen, Xiaoqin Zhang, and Shengjia Li. Multiple linear regression with compositional response and covariates. *J. Appl. Stat.*, 44(12):2270–2285, 2017.

Shaobing Chen and David Donoho. Basis pursuit. In *Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

Ze Hua Chen. Fitting multivariate regression functions by interaction spline models. *J. Roy. Statist. Soc. Ser. B*, 55(2):473–491, 1993.

P.A. Chiappori and B. Salanié. Testing for asymmetric information in insurance markets. *Journal of Political Economy*, 108(1):56–78, 2000.

William G Cochran. Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4): 417–451, 1954.

Donald L. Cohn. *Measure Theory*. Springer, New York, second edition, 2013.

R. Dennis Cook. Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.*, 32 (3):1062–1092, 2004.

R. Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. *Ann. Statist.*, 30(2):455–474, 2002.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49, 2002.

Somayeh Danafar, Arthur Gretton, and Jürgen Schmidhuber. Characteristic kernels on structured domains excel in robotics and human action recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 264–279. Springer, 2010.

A. P. Dawid. Conditional independence in statistical theory. *J. Roy. Statist. Soc. Ser. B*, 41(1): 1–31, 1979.

Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Measuring association on topological spaces using kernels and geometric graphs. *arXiv preprint arXiv:2010.01768*, 2020.

J. Diestel and B. Faires. On vector measures. *Trans. Amer. Math. Soc.*, 198:253–271, 1974.

Nicolae Dinculeanu. Vector integration and stochastic integration in Banach spaces. Oxford University Press, 2011.

Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141, 2014.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375. Springer Science & Business Media, 1996.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Jerome H. Friedman. Multivariate adaptive regression splines. *Ann. Statist.*, 19(1):1–141, 1991.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.*, 5:73–99, 2003/04.

Kenji Fukumizu, Francis R. Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *J. Mach. Learn. Res.*, 8:361–383, 2007.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, pages 489–496, 2008.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Kernel dimension reduction in regression. *Ann. Statist.*, 37(4):1871–1905, 2009a.

Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Bharath K. Sriperumbudur. Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems*, volume 21, pages 473–480, 2009b.

Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *J. Mach. Learn. Res.*, 14:3753–3783, 2013.

Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. *VSURF: Variable Selection Using Random Forests*, 2019. R package version 1.1.0.

Edward I. George. The variable selection problem. *J. Amer. Statist. Assoc.*, 95(452):1304–1308, 2000.

V Glahn and K Hron. Simplicial regression. The normal model. *J. Appl. Probab. Stat.*, 6:87–108, 2012.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Nearly optimal classification for semimetrics. *J. Mach. Learn. Res.*, 18:Paper No. 37, 22, 2017.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, volume 3734 of *Lecture Notes in Comput. Sci.*, pages 63–77. Springer, Berlin, 2005.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

Martin S Hanna and Ted Chang. Fitting smooth histories to rotation data. *J. Multivariate Anal.*, 75(1):47–61, 2000.

Trevor Hastie and Brad Efron. *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2013. R package version 1.2.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, second edition, 2009.

A Hebbali. olsrr: Tools for building ols regression models, 2018. R package version 0.5.1.

J.J. Heckman, H. Ichimura, and P.E. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605, 1997.

M Hein and O Bousquet. Kernels, Associated Structures and Generalizations. Technical Report of the Max Planck Institute for Biological Cybernetics 127, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2004.

Rafiq H. Hijazi and Robert W. Jernigan. Modeling compositional data using Dirichlet regression models. *J. Appl. Probab. Stat.*, 4(1):77–91, 2009.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

J. Hoffmann-Jørgensen and G. Pisier. The law of large numbers and the central limit theorem in Banach spaces. *Ann. Probab.*, 4(4):587–599, 1976.

Torsten Hothorn, Kurt Hornik, Carolin Strobl, and Achim Zeileis. Package 'party': A laboratory for recursive partytioning, 2015.

K. Hron, A. Menafoglio, M. Templ, K. Hrůzová, and P. Filzmoser. Simplicial principal component analysis for density functions in Bayes spaces. *Comput. Statist. Data Anal.*, 94:330–350, 2016.

Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 2010.

Tzee-Ming Huang. Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Statist.*, 38(4):2047–2091, 2010.

Zhen Huang. *KPC: Kernel Partial Correlation Coefficient*, 2021. R package version 0.1.0.

Zhen Huang, Nabarun Deb, and Bodhisattva Sen. Kernel partial correlation coefficient—a measure of conditional dependence. *arXiv preprint arXiv:2012.14804*, 2020.

Malini Iyengar and Dipak K. Dey. A semiparametric model for compositional data analysis in presence of covariates on the simplex. *Test*, 11(2):303–315, 2002.

David W Jacobs, Daphna Weinshall, and Yoram Gdalyahu. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):583–600, 2000.

Jeong Min Jeon and Byeong U. Park. Additive regression with Hilbertian responses. *Ann. Statist.*, 48(5):2671–2697, 2020.

Chenlu Ke and Xiangrong Yin. Expected conditional characteristic function-based measures for testing independence. *J. Amer. Statist. Assoc.*, 115(530):985–996, 2020.

Ilja Klebanov, Ingmar Schuster, and T. J. Sullivan. A rigorous theory of conditional mean embeddings. *SIAM J. Math. Data Sci.*, 2(3):583–606, 2020.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models*. MIT Press, 2009.

Sun Yuan Kung. *Kernel Methods and Machine Learning*. Cambridge University Press, 2014.

Michael H Kutner, Christopher J Nachtsheim, John Neter, and William Wasserman. *Applied Linear Statistical Models*. McGraw-Hill, fourth edition, 2004.

Vincenzo Lagani, Giorgos Athineou, Alessio Farcomeni, Michail Tsagris, and Ioannis Tsamardinos. Feature selection with the R package MXM: Discovering statistically equivalent feature subsets. *Journal of Statistical Software*, 80(7), 2017.

Steffen L. Lauritzen. *Graphical Models*, volume 17. The Clarendon Press, 1996.

Guy Lebanon and John Lafferty. Information diffusion kernels. In *Advances in Neural Information Processing Systems*, volume 15, pages 375–382. MIT Press, 2002.

Bing Li. *Sufficient Dimension Reduction: Methods and Applications with R*. CRC, 2018.

Lexin Li, R. Dennis Cook, and Christopher J. Nachtsheim. Model-free variable selection. *J. Roy. Statist. Soc. Ser. B*, 67(2):285–299, 2005.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002.

Wei Lin, Pixu Shi, Rui Feng, and Hongzhe Li. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014.

Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, 34(5):2272–2297, 2006.

Oliver Linton and Pedro Gozalo. Conditional independence restrictions: testing and estimation. *Cowles Foundation Discussion Paper*, 1140, 1997.

Nikolai Lusin. Sur les propriétés des fonctions mesurables. *CR Acad. Sci. Paris*, 154(25):1688–1690, 1912.

Nathan Mantel and William Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

Jonathan H. Manton and Pierre-Olivier Amblard. A primer on reproducing kernel Hilbert spaces. *Foundations and Trends in Signal Processing*, 8(1-2):1–126, 2014.

Alan Miller. *Subset Selection in Regression*, volume 95 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2002.

Ha Quang Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constr. Approx.*, 32(2):307–338, 2010.

Natalia Miranda, Edgar Chávez, María Fabiana Piccoli, and Nora Reyes. (very) fast (all) k-nearest neighbors in metric and non metric spaces without indexing. In *International Conference on Similarity Search and Applications*, pages 300–311. Springer, 2013.

Joanna Morais, Christine Thomas-Agnan, and Michel Simioni. Using compositional and Dirichlet models for market share regression. *J. Appl. Stat.*, 45(9):1670–1689, 2018.

Jeffrey S Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimal conditional independence testing. *Ann. Statist.*, 49(4):2151–2177, 2021.

Jihwan Oh, Faye Zheng, R. W. Doerge, and Hyonho Chun. Kernel partial correlation: a novel approach to capturing conditional independence in graphical models for noisy data. *J. Appl. Stat.*, 45(14):2677–2696, 2018.

Stephen M Omohundro. *Five Balltree Construction Algorithms*. International Computer Science Institute, Berkeley, 1989.

N Otero, R Tolosana-Delgado, A Soler, Vera Pawlowsky-Glahn, and A Canals. Relative vs. absolute statistical analysis of compositions: a comparative study of surface waters of a mediterranean river. *Water Research*, 39(7):1404–1414, 2005.

Håkon Otneim and Dag Tjøstheim. The locally gaussian partial correlation. *Journal of Business & Economic Statistics,* 1–13, 2021.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *arXiv preprint arXiv:2002.03689*, 2020.

Rohit K. Patra, Bodhisattva Sen, and Gábor J. Székely. On a nonparametric notion of residual and its applications. *Statist. Probab. Lett.*, 109:208–213, 2016.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Michael D. Perlman. Jensen's inequality for a convex vector-valued function on an infinite-dimensional space. *J. Multivariate Anal.*, 4:52–65, 1974.

Alexander Petersen and Hans-Georg Müller. Functional data analysis for density functions by transformation to a Hilbert space. *Ann. Statist.*, 44(1):183–218, 2016.

Alexander Petersen and Hans-Georg Müller. Wasserstein covariance for multiple random densities. *Biometrika*, 106(2):339–351, 2019.

Sumitra Purkayastha. Simple proofs of two results on convolutions of unimodal distributions. *Statist. Probab. Lett.*, 39(2):97–100, 1998.

J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *J. Roy. Statist. Soc. Ser. B*, 71(5):1009–1030, 2009.

James C. Robinson. *Dimensions, Embeddings, and Attractors*, volume 186 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2011.

Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947, 2018.

Bryan P. Rynne and Martin A. Youngson. *Linear Functional Analysis*. Springer-Verlag London, Ltd., London, second edition, 2008.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, 41(5):2263–2291, 2013.

Marta Sestelo, Nora M. Villanueva, and Javier Roca-Pardinas. *FWDselect: Selecting Variables in Regression Models*, 2015. R package version 2.1.0.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.*, 48(3):1514–1538, 2020.

Michael Ian Shamos and Dan Hoey. Closest-point problems. In *16th Annual Symposium on Foundations of Computer Science*, pages 151–162. 1975.

Tianhong Sheng and Bharath K Sriperumbudur. On distance and kernel measures of conditional independence. *arXiv preprint arXiv:1912.01103*, 2019.

Pixu Shi, Anru Zhang, and Hongzhe Li. Regression analysis for microbiome compositional data. *Ann. Appl. Stat.*, 10(2):1019–1040, 2016.

Kyungchul Song. Testing conditional independence via Rosenblatt transforms. *Ann. Statist.*, 37 (6B):4011–4045, 2009.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Annual International Conference on Machine Learning*, pages 961–968, 2009.

Le Song, Byron Boots, Sajid Siddiqi, Geoffrey J. Gordon, and Alex Smola. Hilbert space embeddings of hidden Markov models. In *International Conference on Machine Learning*, 2010a.

Le Song, Arthur Gretton, and Carlos Guestrin. Nonparametric tree graphical models. In *International Conference on Artificial Intelligence and Statistics*, pages 765–772, 2010b.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Oystein Sorensen. hdme: High-dimensional regression with measurement error. *Journal of Open Source Software*, 4(37):1404, 2019.

Jaime Lynn Speiser, Michael E Miller, Janet Tooze, and Edward Ip. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, 134:93–101, 2019.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning. MIT Press, second edition, 2000.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Injective Hilbert space embeddings of probability measures. In *21st Annual Conference on Learning Theory*, pages 111–122, 2008.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, 12:2389–2410, 2011.

Oyvind Stavdahl, Anne Karin Bondhus, Kristin Y. Pettersen, and Kjell E. Malvig. Optimal statistical operators for 3-dimensional rotational data: geometric interpretations and application to prosthesis kinematics. *Model. Identif. Control*, 26(4):185–200, 2005.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008.

Liangjun Su and Halbert White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834, 2007.

Liangjun Su and Halbert White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008.

Liangjun Su and Halbert White. Testing conditional independence via empirical likelihood. *Journal of Econometrics*, 182(1):27–44, 2014.

Antoni Susin, Yiwen Wang, Kim-Anh Lê Cao, and M Luz Calle. Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, 2(2), 2020.

Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *J. Mach. Learn. Res.*, 18(233):1–29, 2017.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 35(6):2769–2794, 2007.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58 (1):267–288, 1996.

Michail Tsagris. Regression analysis with compositional data containing zero values. *Chil. J. Stat.*, 6(2):47–57, 2015.

Michail Tsagris, Abdulaziz Alenazi, and Connie Stewart. The α-k-NN regression for compositional data. *arXiv preprint arXiv:2002.05137*, 2020.

Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186, 2014.

Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *J. Amer. Statist. Assoc.*, 110(512):1726–1734, 2015.

George Wynne and Andrew B Duncan. A kernel two-sample test for functional data. *arXiv preprint arXiv:2008.11095*, 2020.

Fan Xia, Jun Chen, Wing Kam Fung, and Hongzhe Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063, 2013.

Andrew Chi-Chih Yao. On constructing minimum spanning trees in k-dimensional spaces and related problems. *SIAM Journal on Computing*, 11(4):721–736, 1982.

Paul Yau, Robert Kohn, and Sally Wood. Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *J. Comput. Graph. Statist.*, 12(1):23–54, 2003.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B*, 68(1):49–67, 2006.

Joseph E. Yukich. *Probability Theory of Classical Euclidean Optimization Problems*, volume 1675 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1998.

Adriano Zanin Zambom and Michael G. Akritas. NonpModelCheck: An R package for nonparametric lack-of-fit testing and variable selection. *Journal of Statistical Software*, 77(10):1–28, 2017.

Hao Helen Zhang, Grace Wahba, Yi Lin, Meta Voelker, Michael Ferris, Ronald Klein, and Barbara Klein. Variable selection and model building via likelihood basis pursuit. *J. Amer. Statist. Assoc.*, 99(467):659–672, 2004.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Lu Zhang and Lucas Janson. Floodgate: inference for model-free variable importance. *arXiv preprint arXiv:2007.01283*, 2020.

Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, 67(2):301–320, 2005.