Kernel Integrated R^2: A Measure of Dependence

Pouya Roudaki[1] **Shakeel Gavioli-Akilagun**[1,2] **Florian Kalinke**[3] **Mona Azadkia**[1] **Zoltán Szabó**[1]

[1]Department of Statistics, London School of Economics, London, UK
[2]Department of Decision Analytics and Operations, City University of Hong Kong, Hong Kong, China
[3]Chair of Information Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany

Abstract

We introduce *kernel integrated* R^2, a new measure of statistical dependence that combines the local normalization principle of the recently introduced *integrated* R^2 with the flexibility of reproducing kernel Hilbert spaces (RKHSs). The proposed measure extends integrated R^2 from scalar responses to responses taking values on general spaces equipped with a characteristic kernel, allowing to measure dependence of multivariate, functional, and structured data, while remaining sensitive to tail behaviour and oscillatory dependence structures. We establish that (i) this new measure takes values in $[0, 1]$, (ii) equals zero if and only if independence holds, and (iii) equals one if and only if the response is almost surely a measurable function of the covariates. Two estimators are proposed: a graph-based method using K-nearest neighbours and an RKHS-based method built on conditional mean embeddings. We prove consistency and derive convergence rates for the graph-based estimator, showing its adaptation to intrinsic dimensionality. Numerical experiments on simulated data and a real data experiment in the context of dependency testing for media annotations demonstrate competitive power against state-of-the-art dependence measures, particularly in settings involving non-linear and structured relationships.

1 INTRODUCTION

Measuring the degree of dependence between two random variables is a long-standing problem in machine learning and statistics, and numerous methods have been proposed over the years; see, for example, the recent surveys by Josse and Holmes [2016], Han [2021], Chatterjee [2024]. Among the most widely-used classical measures of statistical association are Pearson's correlation coefficient, Spearman's ρ, and Kendall's τ. These coefficients are highly effective for detecting monotonic relationships, and their asymptotic behaviour is well understood. However, they perform poorly when the underlying association is non-monotonic.

To overcome this limitation, many alternative measures have been proposed, including the maximal correlation coefficient [Hirschfeld, 1935, Gebelein, 1941, Rényi, 1959, Breiman and Friedman, 1985], methods based on joint cumulative distribution functions and ranks [Hoeffding, 1948, Blum et al., 1961, Yanagimoto, 1970, Puri and Sen, 1971, Rosenblatt, 1975, Csörgő, 1985, Romano, 1988, Bergsma and Dassios, 2014, Nandy et al., 2016, Weihs et al., 2016, Han et al., 2017, Wang et al., 2017, Gamboa et al., 2018, Weihs et al., 2018, Drton et al., 2020, Deb and Sen, 2023, Zhou and Müller, 2025], entropy- and mutual information-based measures [Linfoot, 1957, Kraskov et al., 2004, Pál et al., 2010, Póczos and Schneider, 2011, Reshef et al., 2011, Kandasamy et al., 2015], copula-based coefficients [Sklar, 1959, Schweizer and Wolff, 1981, Kirshner and Póczos, 2008, Póczos et al., 2010, Dette et al., 2013, Lopez-Paz et al., 2013, Kong et al., 2019, Zhang, 2019, Griessenberger et al., 2022], measures based on pairwise distances [Friedman and Rafsky, 1983, Székely et al., 2007, Székely and Rizzo, 2009, Heller et al., 2013, Lyons, 2013, Pan et al., 2020], and kernel-based methods [Gretton et al., 2005, 2008, Póczos et al., 2012, Sen and Sen, 2014, Pfister et al., 2018, Zhang et al., 2018]. Notice that, while developed independently in the machine learning and statistics communities, Hilbert-Schmidt independence criterion (HSIC; Gretton et al. 2005, 2008, based on kernels) and distance covariance (Székely et al. 2007, Székely and Rizzo 2009, Lyons 2013, based on metrics) are now known to be equivalent [Sejdinovic et al., 2013].

More recently, Chatterjee [2021] introduced a new correlation coefficient (i) that is as simple to compute as classical measures, (ii) yet serves as a consistent estimator of a dependence measure $\xi(X, Y)$, (iii) it takes values in $[0, 1]$, (iv) it equals 0 if and only if X and Y are independent, and

(iv) it equals 1 if and only if one variable is a measurable function of the other. While $\xi(X, Y)$ was already known as the limit of a copula-based estimator when both Y and X are continuous random variables [Dette et al., 2013], the simplicity, computational efficiency, and interpretability of Chatterjee's correlation have generated substantial interest, leading to a rapidly growing literature on its theoretical properties and extensions to more complex settings [Cao and Bickel, 2020, Deb et al., 2020, Azadkia and Chatterjee, 2021, Griessenberger et al., 2022, Shi et al., 2022, Bickel, 2022, Gamboa et al., 2022, Huang et al., 2022, Lin and Han, 2023, Zhang, 2023, Auddy et al., 2024, Lin and Han, 2024, Fuchs, 2024, Han and Huang, 2024, Shi et al., 2024, Strothmann et al., 2024, Bücher and Dette, 2024, Tran and Han, 2024, Kroll, 2025, Ansari and Fuchs, 2025, Dette and Kroll, 2025, Zhang, 2025, Yang et al., 2025, Huang et al., 2026].

In particular, Deb et al. [2020] extended Chatterjee's correlation to allow handling random variables X and Y taking values in topological spaces under mild conditions.[1] They employ the kernel mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007] and its conditional variant [Fukumizu et al., 2007, Song et al., 2009, Klebanov et al., 2020, Park and Muandet, 2020], which permit mapping (conditional) probability measures into a reproducing kernel Hilbert space (RKHS; Aronszajn 1950, Steinwart and Christmann 2008, Paulsen and Raghupathi 2016) by using a symmetric positive definite function, the kernel function. If the mapping is injective, the kernel is called characteristic [Fukumizu et al., 2007, Sriperumbudur et al., 2010] and the RKHS distance of mean embeddings induces a metric on the space of probability measures, which underpins the well-known maximum mean discrepancy (MMD; Smola et al. 2007, Gretton et al. 2012); it is also known as Hilbert-Schmidt independence criterion [Gretton et al., 2005, Quadrianto et al., 2009, Pfister et al., 2018] if applied to measuring the distance of a joint distribution to the product of its marginals. In this sense, the measure proposed by Deb et al. [2020] can be interpreted as the (normalized) average (w.r.t. X) MMD distance of the distribution of Y and the distribution of Y given X; the computational tractability of RKHS methods allows estimating this quantity. Besides rigorously analysing their proposed population quantity and different families of estimators, Deb et al. [2020] demonstrated empirically that their extension can exhibit greater power for detecting dependence than the original scalar-based coefficient. Additionally, as, for example, kernels for strings [Watkins, 1999, Lodhi et al., 2002] or more generally for sequences [Király and Oberhauser, 2019], sets [Haussler, 1999, Gärtner et al., 2002], rankings [Jiao and Vert, 2016], fuzzy domains [Guevara et al., 2017] and graphs [Borgwardt et al., 2020] are known, their approach is broadly applicable.

[1] More precisely, Y must take values in a Hausdorff space enriched with a characteristic kernel and the regular conditional distribution of Y given X must exist. The latter can be guaranteed if Y takes values in a Polish space. Additional assumptions depend on the respective estimator; we do not recall these here and refer to their article for more details.

Following the development of Chatterjee's correlation, Azadkia and Roudaki [2025] recently introduced a new dependence measure, denoted by $\nu(Y, X)$. This measure retains all the desirable properties of Chatterjee's correlation while exhibiting enhanced sensitivity to dependence structures that manifest in the tails of the distribution or display oscillatory behaviour. However, the structural form of the measure and its associated estimator restrict its applicability to settings in which $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$.

Motivated by Deb et al. [2020], we leverage the flexibility of RKHSs to introduce a dependence measure that preserves the core structural features of $\nu(Y, X)$, while extending the applicability beyond real-valued responses and (finite-dimensional) Euclidean covariates. This extension presents non-trivial technical challenges. Most notably, unlike Deb et al. [2020], our construction employs a local normalization rather than a global one, which necessitates a more delicate proof strategy. In particular, we make the following **contributions**.

1. We introduce a kernel-based generalization of $\nu(Y, X)$, extending it beyond the setting of $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. In particular, our population quantity is well-defined if X takes values in a topological space and Y takes values in a Polish space equipped with a continuous characteristic kernel.
2. We prove that our proposed generalization has the properties expected of a dependence measure, that is, the quantity takes values between 0 and 1, is 0 if and only if (iff.) X and Y are independent, and is 1 iff. Y is almost surely (a.s.) a measurable function of X.
3. Under mild additional assumptions, we present a graph-based estimator using nearest neighbours, and an RKHS-based estimator of our kernel-based quantity. We provide consistency guarantees and convergence rates for the graph-based estimator.
4. Experiments on synthetic and real-world datasets show that independence tests using our estimators perform competitively w.r.t. the state-of-the-art, especially when considering non-linear and structured associations.

The remainder of the paper has the following **outline**. In Section 2, we introduce the main notations. Section 3 reviews some closely-related measures of dependence. In Section 4, we introduce our general measure of dependence, termed "kernel integrated R^2". Section 5 presents two estimation procedures, and Section 6 establishes theoretical guarantees, including consistency and convergence rates. The empirical performance of the proposed method is demonstrated in Section 7. All proofs are collected in the appendix.

2 NOTATIONS

Next we introduce our notations used throughout the main body of the article: $[n]$, $\mathcal{O}$, $\mathbb{1}_A$, $\mathbf{1}_n$, $\mathbf{I}_n$, $\mathbf{A}^\top$, $\mathrm{Tr}(\mathbf{A})$, $\mathbf{A}^{-1}$, $\circ$, $\mathcal{M}_1^+(\mathcal{Z})$, $\mathrm{supp}(\mathbb{P})$, δ_z, $\mathcal{O}_\mathbb{P}$, $\mathbb{E}_\mathbb{P}[\cdot]$, $\mathbb{V}_\mathbb{P}(\cdot)$, $\mathbb{E}_Z[\cdot]$, $\mathbb{V}_Z(\cdot)$, Cov, $\mathbb{P}_X$, $\mathbb{P}_Y$, $\mathbb{P}_{XY}$, $\mathbb{P}_X \otimes \mathbb{P}_Y$, $\mathbb{P}_{Y|X}$, $\mathcal{H}_k$, $k(\cdot,z)$, μ_k, MMD_k, $\mathcal{H}_\mathcal{Y}$, $\mathcal{H}_\mathcal{X}$.

General conventions. For a positive integer $n \in \mathbb{N} \coloneqq \{1,2,\ldots\}$, $[n] \coloneqq \{1,\ldots,n\}$. For positive sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we write $a_n = \mathcal{O}(b_n)$ if there exist constants $C > 0$ and $N \in \mathbb{N}$ such that $a_n \leq C\, b_n$ for all $n \geq N$. We denote by $\mathbb{1}_A$ the indicator of a set A: $\mathbb{1}_A(x) = 1$ if $x \in A$; $\mathbb{1}_A(x) = 0$ otherwise. The n-dimensional vector of ones is denoted by $\mathbf{1}_n$. The identity matrix is $\mathbf{I}_n \in \mathbb{R}^{n\times n}$. For a matrix $\mathbf{A} \in \mathbb{R}^{n_1\times n_2}$, its transpose is written as $\mathbf{A}^\top \in \mathbb{R}^{n_2\times n_1}$; the trace of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is denoted by $\mathrm{Tr}(\mathbf{A})$; for a non-singular matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, its inverse is denoted by $\mathbf{A}^{-1} \in \mathbb{R}^{n\times n}$. For two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n_1\times n_2}$, we write their Hadamard product as $\mathbf{A} \circ \mathbf{B} = [A_{i,j}B_{i,j}]_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1\times n_2}$.

Probability measures and conditioning. Let $(\mathcal{Z}, \tau_\mathcal{Z})$ be a topological space and $\mathcal{B}(\tau_\mathcal{Z})$ its Borel σ-algebra. We write $\mathcal{M}_1^+(\mathcal{Z})$ for the set of Borel probability measures on the measurable space $(\mathcal{Z}, \mathcal{B}(\tau_\mathcal{Z}))$. The support of $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{Z})$ is denoted by $\mathrm{supp}(\mathbb{P})$; it is the set of points $z \in \mathcal{Z}$ for which every open neighbourhood of z has positive $\mathbb{P}$ measure. We write δ_z for the Dirac delta distribution at $z \in \mathcal{Z}$. A distribution $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{Z})$ is called degenerate iff. $\mathbb{P} = \delta_z$ for some $z \in \mathcal{Z}$. For a sequence of random variables X_n and sequence of positive reals $(a_n)_{n\in\mathbb{N}}$, $X_n = \mathcal{O}_\mathbb{P}(a_n)$ means that X_n/a_n is stochastically bounded, that is for any $\varepsilon > 0$ there exists a finite $M > 0$ and a finite $N > 0$ such that $\mathbb{P}(|X_n/a_n| > M) < \varepsilon$ for all $n > N$. Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_\mathcal{H})$ be a Hilbert space, $f : \mathcal{Z} \to \mathcal{H}$ measurable, and $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{Z})$. If $\int_\mathcal{Z} \|f(z)\|_\mathcal{H}\, \mathrm{d}\mathbb{P}(z) < \infty$, the expectation of f w.r.t. $\mathbb{P}$ is defined as $\mathbb{E}_\mathbb{P}[f] \coloneqq \int_\mathcal{Z} f(z)\, \mathrm{d}\mathbb{P}(z) \in \mathcal{H}$, where the integral is meant in Bochner's sense. If additionally $\int_\mathcal{Z} \|f(z)\|_\mathcal{H}^2\, \mathrm{d}\mathbb{P}(z) < \infty$, then the variance of f w.r.t. $\mathbb{P}$ is $\mathbb{V}_\mathbb{P}(f) \coloneqq \int_\mathcal{Z} \|f(z) - \mathbb{E}_\mathbb{P}[f]\|_\mathcal{H}^2\, \mathrm{d}\mathbb{P}(z)$. When $\mathbb{P}$ is the law of a random variable Z, we write $\mathbb{E}_Z[\cdot]$ and $\mathbb{V}_Z(\cdot)$ for $\mathbb{E}_\mathbb{P}[\cdot]$ and $\mathbb{V}_\mathbb{P}(\cdot)$, respectively. The covariance of two real-valued random variables Z_1 and Z_2 with joint law $\mathbb{P}$ and marginal laws $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively, is $\mathrm{Cov}(Z_1, Z_2) = \mathbb{E}_{Z_1,Z_2}[(Z_1 - \mathbb{E}_{Z_1}[Z_1])(Z_2 - \mathbb{E}_{Z_2}[Z_2])]$. Let $(\mathcal{X}, \mathcal{B}(\tau_\mathcal{X}))$ and $(\mathcal{Y}, \mathcal{B}(\tau_\mathcal{Y}))$ be measurable spaces and (X, Y) a pair of random variables taking values in $\mathcal{X}\times\mathcal{Y}$. We denote their marginal and joint laws by $\mathbb{P}_X \in \mathcal{M}_1^+(\mathcal{X})$, $\mathbb{P}_Y \in \mathcal{M}_1^+(\mathcal{Y})$, and $\mathbb{P}_{XY} \in \mathcal{M}_1^+(\mathcal{X}\times\mathcal{Y})$, respectively. Their product distribution is denoted by $\mathbb{P}_X \otimes \mathbb{P}_Y$. If $\mathcal{Y}$ is a Polish space, that is, a complete separable metrizable topological space, a (regular) conditional distribution of Y given X exists, which we write as $\mathbb{P}_{Y|X}$.

Kernels and RKHS. Let $\mathcal{H}_k$ be the RKHS on $\mathcal{Z}$ with (reproducing) kernel $k : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}$; it is the Hilbert space of functions $f : \mathcal{Z} \to \mathbb{R}$ such that $k(\cdot,z) \in \mathcal{H}_k$ and $\langle f, k(\cdot,z)\rangle_{\mathcal{H}_k} = f(z)$ for all $z \in \mathcal{Z}$ and $f \in \mathcal{H}_k$, where the canonical feature map $k(\cdot,z)$ stands for $z' \mapsto k(z',z) \in \mathbb{R}$ for fixed z and any $z' \in \mathcal{Z}$. Throughout this manuscript, we assume all kernels to be Borel measurable and bounded (for a kernel $k : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}$, the latter property is meant as $\sup_{z,z'\in\mathcal{Z}} k(z,z') < \infty$). For $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{Z})$, the kernel mean embedding of $\mathbb{P}$ w.r.t. k is $\mu_k(\mathbb{P}) \coloneqq \int_\mathcal{Z} k(\cdot,z)\, \mathrm{d}\mathbb{P}(z) \in \mathcal{H}_k$, where the integral is meant in Bochner's sense; the assumed boundedness of k ensures its existence. The maximum mean discrepancy of $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{Z})$ w.r.t. k is $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}$. A kernel k is called characteristic if the map $\mathbb{P} \mapsto \mu_k(\mathbb{P})$ is injective on $\mathcal{M}_1^+(\mathcal{Z})$. In this case, MMD_k induces a metric on $\mathcal{M}_1^+(\mathcal{Z})$. In the main text, we work with the RKHS $\mathcal{H}_\mathcal{Y} \coloneqq \mathcal{H}_{k_\mathcal{Y}}$ induced by the kernel $k_\mathcal{Y} : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$; for the RKHS-based estimator (Section 5.2), we additionally use the RKHS $\mathcal{H}_\mathcal{X} \coloneqq \mathcal{H}_{k_\mathcal{X}}$ induced by the kernel $k_\mathcal{X} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$.

3 RELATED DEPENDENCE MEASURES

For a real-valued random variable $Y \in \mathbb{R}$ and a random vector $X \in \mathbb{R}^d$ for $d \geq 1$, Azadkia and Roudaki [2025] introduced $\nu(Y,X)$ to quantify the extent of dependence of Y on X. If Y admits a continuous density[2], $\nu(Y,X)$ can be written as

$$\nu(Y,X) = \int_\mathbb{R} \frac{\mathbb{V}_X\big(\mathbb{E}_{Y|X}[\mathbb{1}_{\{Y>t\}}]\big)}{\mathbb{V}_Y\big(\mathbb{1}_{\{Y>t\}}\big)}\, \mathrm{d}\mathbb{P}_Y(t), \qquad (1)$$

which is closely related to Chatterjees correlation coefficient [Chatterjee, 2021],

$$\xi(X,Y) = \int_\mathbb{R} \frac{\mathbb{V}_X\big(\mathbb{E}_{Y|X}[\mathbb{1}_{\{Y>t\}}]\big)}{\int_\mathbb{R} \mathbb{V}_Y\big(\mathbb{1}_{\{Y>u\}}\big)\, \mathrm{d}\mathbb{P}_Y(u)}\, \mathrm{d}\mathbb{P}_Y(t).$$

While introduced for $X, Y \in \mathbb{R}$, ξ can be extended to allow handling $X \in \mathbb{R}^d$ [Azadkia and Chatterjee, 2021]. [3]

Although the difference between ν and ξ may appear marginal, in practice ν is often more powerful for detecting dependence. The distinction lies in the normalization of $\mathbb{V}_X(\mathbb{E}_{Y|X}[\mathbb{1}_{\{Y>t\}}])$. In $\nu(Y,X)$, this quantity is normalized pointwise by $\mathbb{V}_Y(\mathbb{1}_{\{Y>t\}})$, so the conditional variation of $\mathbb{1}_{\{Y>t\}}$ given X is compared directly to its marginal variation, rather than to an average over all thresholds $t \in \mathbb{R}$.

[2] The continuous density ensures that (1) is well-defined. For the exact definition of ν in the general case, see Azadkia and Roudaki [2025].

[3] Note that $\xi(X,Y)$ measures the extent of dependence of Y on X. Although similar measures are typically written with arguments ordered as (Y,X), we retain the ordering (X,Y) for ξ to remain consistent with Chatterjee [2021], where ξ was originally introduced.

As a consequence, even when $\mathbb{V}_Y(\mathbb{1}_{\{Y>t\}})$ is small, the dependence of $\mathbb{1}_{\{Y>t\}}$ on X is not masked by averaging over values of t with larger marginal variability.

In a different line of work, Deb et al. [2020] proposed $\eta_{k_\mathcal{Y}}$, a general measure of association inspired by $\xi(X,Y)$, by leveraging RKHS methods. Let Y and Y' be conditionally i.i.d. given X, and let Y, Y_1, and Y_2 be marginally i.i.d. Then the measure $\eta_{k_\mathcal{Y}}$ is defined as

$$\eta_{k_\mathcal{Y}}(Y,X) := 1 - \frac{\mathbb{E}_X\Big[\mathbb{E}_{Y,Y'|X}\big[\big\|k_\mathcal{Y}(\cdot,Y)-k_\mathcal{Y}(\cdot,Y')\big\|^2_{\mathcal{H}_\mathcal{Y}}\big]\Big]}{\mathbb{E}_{Y_1,Y_2}\big[\big\|k_\mathcal{Y}(\cdot,Y_1)-k_\mathcal{Y}(\cdot,Y_2)\big\|^2_{\mathcal{H}_\mathcal{Y}}\big]} \tag{2}$$

$$= \int_\mathcal{X} \frac{\mathrm{MMD}^2_{k_\mathcal{Y}}\left(\mathbb{P}_{Y|X=x}, \mathbb{P}_Y\right)}{\int_\mathcal{Y} \|k_\mathcal{Y}(\cdot,y)-\mathbb{E}_Y\left[k_\mathcal{Y}(\cdot,Y)\right]\|^2_{\mathcal{H}_\mathcal{Y}} \mathrm{d}\mathbb{P}_Y(y)} \mathrm{d}\mathbb{P}_X(x).$$

An advantage of $\eta_{k_\mathcal{Y}}$ over ξ is that, by leveraging the RKHS framework, it naturally extends to multivariate responses Y and general data types such as graphs, manifolds, and functional data. Moreover, an appropriate choice of kernel allows domain knowledge about similarity to be incorporated into the measure. Note that for $Y, X \in \mathbb{R}$, and the Brownian kernel $k_\mathcal{Y}(y_1,y_2) = |y_1| + |y_2| - |y_1 - y_2|$, we get $\eta_{k_\mathcal{Y}}(Y,X) = \xi(X,Y)$, hence $\eta_{k_\mathcal{Y}}$ can be viewed as a generalization of ξ.

Motivated by the complementary strengths of ν and $\eta_{k_\mathcal{Y}}$, we combine the power of both approaches by introducing a kernelized version of ν, with the goal of further enhancing its ability to detect dependence and to broaden its applicability.

In doing so, we tackle two key challenges. First, the denominator of the integrand in $\nu(Y,X)$ is $\mathbb{V}_Y(\mathbb{1}_{\{Y>t\}})$, which can be small in the tails and therefore requires careful control. In contrast, $\eta_{k_\mathcal{Y}}(Y,X)$ involves a single global normalization term. When extending $\nu(Y,X)$ to an RKHS-based framework, one must therefore ensure uniform control of the corresponding denominator.

Second, note that for i.i.d. random variables Z and Z', we have

$$\mathbb{V}_Z(Z) = \frac{1}{2}\mathbb{E}_{Z,Z'}\left[(Z-Z')^2\right].$$

Therefore the numerator and denominator in (2) are related to the conditional variance of Y given X and the variance of Y, respectively. Hence, (2) shows that $\eta_{k_\mathcal{Y}}(Y,X)$ can be interpreted as the ratio of two variances in a Hilbert space. By contrast, constructing a kernel analogue of ν requires considering the variance of conditional objects across the range of values of the response variable, which does not admit an immediate representation in an RKHS.

Finally, observe that the integrand in (1) can be interpreted as the R^2 (coefficient of determination) from the linear regression of $\mathbb{1}_{\{Y>t\}}$ on X,

$$R^2_t = \frac{\mathbb{V}_X(\mathbb{E}_{Y|X}[\mathbb{1}_{\{Y>t\}}])}{\mathbb{V}_Y(\mathbb{1}_{\{Y>t\}})} = 1 - \frac{\mathbb{E}_X[\mathbb{V}_{Y|X}(\mathbb{1}_{\{Y>t\}})]}{\mathbb{V}_Y(\mathbb{1}_{\{Y>t\}})}.$$

Consequently, $\nu(Y,X) = \int_\mathbb{R} R^2_t \mathrm{d}\mathbb{P}_Y(t)$ can be viewed as an integrated R^2 over the thresholded responses $\mathbb{1}_{\{Y>t\}}$. Motivated by this perspective, we introduce in the next section a kernel-based analogue, which we refer to as the kernel integrated R^2.

4 KERNEL INTEGRATED R^2

In this section, we introduce our measure of dependence, the kernel integrated R^2. We first state the set of assumptions that we require for our measure to be well-defined.

Assumption 1. *(i) $\mathcal{X}$ is a topological space and $\mathcal{Y}$ is a Polish space. (ii)* $\mathrm{supp}(\mathbb{P}_Y) = \mathcal{Y}$ *and $\mathbb{P}_Y$ is non-degenerate. (iii) $k_\mathcal{Y}$ is continuous, characteristic, and there exists no $y \in \mathcal{Y}$ such that $k_\mathcal{Y}(\cdot,y)$ is a constant function.*[4]

The following remark elaborates our assumptions.

Remark 1.

1. *The assumption that $\mathcal{Y}$ is Polish ensures the existence of regular conditional probabilities. [Dudley, 2004, Theorem 10.2.2].*
2. *As $k_\mathcal{Y}$ is continuous and* $\mathrm{supp}(\mathbb{P}_Y) = \mathcal{Y}$*, for any $f \in \mathcal{H}_\mathcal{Y}$, we have $f = 0$ $\mathbb{P}_Y$-a.e. iff. $f \equiv 0$ on $\mathcal{Y}$ [Klebanov et al., 2020, Assumption 2.1(f) and footnote 7].*
3. *The characteristic property implies that $k_\mathcal{Y}$ is point-separating [Bonnier et al., 2023, p. 5], that is, $y \mapsto k_\mathcal{Y}(\cdot,y) \in \mathcal{H}_\mathcal{Y}$ is injective for $y \in \mathcal{Y}$.*[5]

Together, these properties pave the way to ensuring that the following definition of our quantity of main interest satisfies various natural requirements of a dependence measure, as we elaborate in Theorem 1.

Definition 1 (Kernel integrated R^2)**.** *Under Assumption 1, let*

$$D(Y,X) := D(Y,X;k_\mathcal{Y}) := 1 - \int_\mathcal{Y} \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}[k_\mathcal{Y}(Y,y)]\right]}{\mathbb{V}_Y[k_\mathcal{Y}(Y,y)]} \mathrm{d}\mathbb{P}_Y(y). \tag{3}$$

[4] This condition holds, for instance, if $\mathcal{Y} = \mathbb{R}^d$ and $k_\mathcal{Y}$ is the Gaussian kernel.

[5] The point-separating property follows from the characteristicness as $k_\mathcal{Y}(\cdot,y) = \mu_{k_\mathcal{Y}}(\delta_y) \overset{\text{(char.)}}{\neq} \mu_{k_\mathcal{Y}}(\delta_{y'}) = k_\mathcal{Y}(\cdot,y')$ for any distinct $y, y' \in \mathcal{Y}$.

Remark 2.

1. *When ignoring that $k_{\mathcal{Y}}$ must be a kernel, and when $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, then choosing $k_{\mathcal{Y}}(u, y) = \mathbb{1}_{\{u>y\}}$, we have*[6]

$$D(Y, X) = 1 - \int_{\mathbb{R}} \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}[\mathbb{1}_{\{Y>y\}}]\right]}{\mathbb{V}_Y[\mathbb{1}_{\{Y>y\}}]} \mathrm{d}\mathbb{P}_Y(y),$$

   which is formally equivalent to $\nu(Y, X)$. In this sense, D can be thought of as a "kernel" version of ν.

2. *$k_{\mathcal{Y}}(\cdot, y)$ being non-constant for any $y \in \mathcal{Y}$ and $\mathrm{supp}(\mathbb{P}_Y) = \mathcal{Y}$ ensure that the denominator $\mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)]$ is non-zero for all $y \in \mathcal{Y}$; we prove this claim as part of the following Theorem 1.*

Our first theorem shows that $D(Y, X)$ is well-defined and satisfies the standard properties expected of a dependence measure.

Theorem 1. *Under Assumption 1, $D(Y, X)$ is well-defined and*

(i) $D(Y, X) \in [0, 1]$,

(ii) $D(Y, X) = 0$ *iff. Y and X are independent, and*

(iii) $D(Y, X) = 1$ *iff. there exists a Borel measurable function $f : \mathcal{X} \to \mathcal{Y}$ such that $Y = f(X)$ holds $\mathbb{P}_{XY}$-a.s.*

Note that D, like ξ, $\eta_{k_{\mathcal{Y}}}$, and ν, is not symmetric in its arguments; it can be advantageous to work with a directional measure of dependence, for example, if $X \sim N(0, 1)$ and $Y = X^2$, then Y is a measurable function of X, but X is not a measurable function of Y. This directionality sets these measures apart from many classical or widely-used symmetric measures of dependence, such as Spearman's ρ, mutual information, or HSIC. While kernel integrated R^2 can be symmetrized,[7] many other measures, e.g., HSIC, are intrinsically symmetric and do not encode a direction of dependence. Table 1 summarizes key properties of D in comparison with related measures.

Property	ξ	ν	$\eta_{k_{\mathcal{Y}}}$	D
Value in $[0, 1]$	✓	✓	✓	✓
0-independent	✓	✓	✓	✓
1-full dependence	✓	✓	✓	✓
Bijective invariance	✓	✓	×	×
Multivariate Y	×	×	✓	✓
Kernel-endowed domain X, Y	×	×	✓	✓

Table 1: Comparison of dependence measures and their properties.

In the next section, we develop estimators for the population quantity $D(Y, X)$.

[6]The function $k_{\mathcal{Y}}$ is not symmetric, hence it is not a kernel.

[7]For example, when both directions are well-defined, one may set $D^{\mathrm{sym}}(Y, X) := \max\{D(Y, X), D(X, Y)\}$.

5 ESTIMATION

In this section, we introduce two approaches for estimating $D(Y, X)$ from an i.i.d. sample $(X_i, Y_i)_{i=1}^n$. Both estimators follow the same general strategy. In estimating $D(Y, X)$, we can approximate the integral with respect to $\mathbb{P}_Y$ by an average over the observed sample $\{Y_j\}_{j=1}^n$,

$$\int_{\mathcal{Y}} \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}\left[k_{\mathcal{Y}}(Y, y)\right]\right]}{\mathbb{V}_Y\left[k_{\mathcal{Y}}(Y, y)\right]} \mathrm{d}\mathbb{P}_Y(y) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}\left[k_{\mathcal{Y}}(Y, Y_i)\right]\right]}{\mathbb{V}_Y\left[k_{\mathcal{Y}}(Y, Y_i)\right]}, \quad (4)$$

where the random part in each summand is Y and Y_i is the observed sample. Hence one needs to provide an estimate for each summand.

While both estimators rely on (4), we emphasize that the two estimators are fundamentally different in nature. The first estimator relies on the nearest-neighbour graph structure induced by the X_i's and therefore requires $\mathcal{X}$ to be a metric space. The second estimator does not impose such a requirement on $\mathcal{X}$ but rather assumes the existence of a kernel $k_{\mathcal{X}}$ on $\mathcal{X}$, and makes use of the conditional mean embedding.

5.1 NEAREST-NEIGHBOUR ESTIMATOR

We begin by introducing an estimator based on approximating the conditional distribution $\mathbb{P}_{Y|X}$ using the K-nearest neighbours method. To formalize this construction, we impose the following assumption.

Assumption 2. *The space $(\mathcal{X}, d_{\mathcal{X}})$ is a metric space.*

Note that the numerator of (4) contains a conditional variance, which for any $y \in \mathcal{Y}$ can be expressed as

$$\mathbb{V}_{Y_j|X_j}[k_{\mathcal{Y}}(Y_j, y)] = \frac{1}{2}\mathbb{E}_{Y_j, Y_j'|X_j}\left[\left(k_{\mathcal{Y}}(Y_j, y) - k_{\mathcal{Y}}(Y_j', y)\right)^2\right],$$

where Y_j' is an i.i.d. copy of Y_j conditional on X_j. To estimate this quantity, we approximate Y_j' using a surrogate. Given a sample $\{(X_i, Y_i)\}_{i=1}^n$, for each Y_j, we select Y_k as a surrogate for Y_j' whenever $d_{\mathcal{X}}(X_k, X_j)$ is sufficiently small. To make the notion of "small" precise, we restrict attention to those X_k that belong to the set of K-nearest neighbours of X_j. More precisely, for distinct indices i and j, let $\mathcal{N}_j^{\setminus i}$ denote the set of the K nearest neighbours of X_j among $\{X_k\}_{k \neq i,j}$, with ties broken uniformly at random.

We let for each $i \in [n]$,

$$E_{n,i}^{\text{K-NN}} := \frac{1}{2K(n-1)} \sum_{j \neq i} \sum_{l \in \mathcal{N}_j^{\setminus i}} \big(k_{\mathcal{Y}}(Y_j, Y_i) - k_{\mathcal{Y}}(Y_l, Y_i)\big)^2, \tag{5}$$

$$V_{n,i}^{\text{K-NN}} := \frac{1}{n-1} \sum_{j \neq i} k_{\mathcal{Y}}^2(Y_j, Y_i) - \left[\frac{1}{n-1} \sum_{j \neq i} k_{\mathcal{Y}}(Y_j, Y_i)\right]^2, \tag{6}$$

be the estimators for $\mathbb{E}_X[\mathbb{V}_{Y|X}[k_{\mathcal{Y}}(Y, Y_i)]]$ and $\mathbb{V}_Y[k_{\mathcal{Y}}(Y, Y_i)]$, respectively. Consequently, we construct an estimator of $D(Y, X)$ as follows.

Definition 2 (Nearest-neighbour estimator)**.** *Suppose that Assumption 2 holds. Given an i.i.d. sample $(X_i, Y_i)_{i=1}^n$ from $\mathbb{P}_{XY}$, let*

$$\hat{D}^{\text{K-NN}}(Y, X) := 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{E_{n,i}^{\text{K-NN}}}{V_{n,i}^{\text{K-NN}}},$$

with $E_{n,i}^{\text{K-NN}}$ and $V_{n,i}^{\text{K-NN}}$ as in (5) *and* (6)*, respectively.*

We emphasize that the use of the K-nearest neighbours is not essential to the construction. In principle, it may be replaced by any geometric graph built on $(X_i)_{i=1}^n$; see Bhattacharya [2019] for a general framework.

5.2 RKHS ESTIMATOR

Next, by interpreting the conditional expectations in (3) as conditional mean embeddings, we present an alternative estimator of $D(Y, X)$. We begin by introducing the key additional assumptions and necessary notation.

Assumption 3. *(i) $\mathcal{X}$ is separable, (ii) $\operatorname{supp}(\mathbb{P}_X) = \mathcal{X}$, (iii) $k_{\mathcal{X}}$ is characteristic and continuous, (iv) for all $g \in \mathcal{H}_{\mathcal{Y}}$ there exists $h_g \in \mathcal{H}_{\mathcal{X}}$ such that $\operatorname{Cov}(h_g(X) - f_g(X), h(X)) = 0$ for all $h \in \mathcal{H}_{\mathcal{X}}$, where $f_g(x) = \mathbb{E}_{Y|X=x}[g(Y)]$ for $x \in \mathcal{X}$, and (v) $k_{\mathcal{Y}}^2$ is characteristic.*

Notice that Assumption 3 features requirements for rigorously handling the conditional mean embedding, as detailed in Klebanov et al. [2020]. The characteristic property of a kernel is a relatively mild assumption with various sufficient conditions [Sriperumbudur et al., 2011]. We further elaborate the assumptions in the following.

Remark 3. *Suppose that Assumption 1 and Assumption 3 both hold.*

1. *The separability of $\mathcal{X}$ and $\mathcal{Y}$, together with the continuity of $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$, imply the separability of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively [Steinwart and Christmann, 2008, Lemma 4.33].*
2. *As in Remark 1(2), it holds for any $f \in \mathcal{H}_{\mathcal{X}}$ that $f = 0$ $\mathbb{P}_X$-a.e. iff. $f \equiv 0$ on $\mathcal{X}$.*
3. *Condition (iv) concerns the richness of $\mathcal{H}_{\mathcal{X}}$ relative to $\mathcal{H}_{\mathcal{Y}}$. It corresponds to Assumption C of Klebanov et al. [2020] and is needed to ensure that for any $x \in \mathcal{X}$, the conditional mean embedding*

$$\mu_{k_{\mathcal{Y}}}(\mathbb{P}_{Y|X=x}) = \mathbb{E}_{Y|X=x}[k_{\mathcal{Y}}(\cdot, Y)] \in \mathcal{H}_{\mathcal{Y}}, \tag{7}$$

   can be realized via linear algebra involving covariance operators [Klebanov et al., 2020, Theorem 4.3].
4. *To define* (7)*, Klebanov et al. [2020, (2.4)] assign $\mu_{k_{\mathcal{Y}}}(\mathbb{P}_{Y|X=x}) := 0$ for $x \in \mathcal{X} \setminus \mathcal{X}_{\mathcal{Y}}$, with $\mathcal{X}_{\mathcal{Y}} := \{x \in \mathcal{X} \mid \mathbb{E}_{Y|X=x}\|k_{\mathcal{Y}}(\cdot, Y)\|_{\mathcal{H}_{\mathcal{Y}}}^2 < \infty\}$. In our case $\mathcal{X}_{\mathcal{Y}} = \mathcal{X}$ as the boundedness of $k_{\mathcal{Y}}$ implies that of $\|k_{\mathcal{Y}}(\cdot, y)\|_{\mathcal{H}_{\mathcal{Y}}}$ for $y \in \mathcal{Y}$ [Steinwart and Christmann, 2008, p. 124]; hence, this distinction is not needed.*

Given an i.i.d. sample $(X_i, Y_i)_{i=1}^n$ from $\mathbb{P}_{XY}$ and $\epsilon_n > 0$, define the $n \times n$ matrices

$$\begin{aligned} &\mathbf{K}_X = [k_{\mathcal{X}}(X_i, X_j)]_{i,j=1}^n, \qquad &&\tilde{\mathbf{K}}_X = \mathbf{H}\mathbf{K}_X\mathbf{H}, \\ &\mathbf{K}_Y = [k_{\mathcal{Y}}(Y_i, Y_j)]_{i,j=1}^n, \\ &\mathbf{M} = \mathbf{K}_Y\tilde{\mathbf{K}}_X(\tilde{\mathbf{K}}_X + n\epsilon_n\mathbf{I}_n)^{-1}. \end{aligned} \tag{8}$$

where $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \in \mathbb{R}^{n \times n}$ is the centering matrix. Denote the j-th canonical basis vector by $\mathbf{e}_j \in \mathbb{R}^n$. With these notations in place, $\mathbb{E}_X[\mathbb{V}_{Y|X}[k_{\mathcal{Y}}(Y, Y_i)]]$ and $\mathbb{V}_Y[k_{\mathcal{Y}}(Y, Y_i)]$ ($i \in [n]$) can be estimated, respectively, via

$$\begin{aligned} E_{n,i}^{\text{RKHS}} := \frac{1}{n}\big\langle \mathbf{e}_i, &(\mathbf{K}_Y \circ \mathbf{K}_Y)\mathbf{1}_n \\ &+ (\mathbf{K}_Y \circ \mathbf{K}_Y)\tilde{\mathbf{K}}_X(\tilde{\mathbf{K}}_X + n\epsilon_n\mathbf{I}_n)^{-1}\mathbf{1}_n \\ &- \frac{1}{n}\mathbf{K}_Y\mathbf{1}_n\mathbf{1}_n^\top\mathbf{K}_Y\mathbf{e}_i - \frac{2}{n}\mathbf{K}_Y\mathbf{1}_n\mathbf{1}_n^\top\mathbf{M}^\top\mathbf{e}_i \\ &- \mathbf{M}\mathbf{M}^\top\mathbf{e}_i\big\rangle_{\mathbb{R}^n}, \end{aligned} \tag{9}$$

$$V_{n,i}^{\text{RKHS}} := \frac{1}{n}\mathbf{1}_n^\top(\mathbf{K}_Y \circ \mathbf{K}_Y)\mathbf{e}_i - \left(\frac{1}{n}\mathbf{1}_n^\top\mathbf{K}_Y\mathbf{e}_i\right)^2. \tag{10}$$

The underlying idea is to use the linear algebraic formulation of (7) to obtain a quantity which then permits plug-in estimation. The first step relies on Assumption 1 and Assumption 3; see the formal justification in Appendix A.2.

Using the above notations, we are ready to introduce the following RKHS-based estimator of $D(Y, X)$.

Definition 3 (RKHS-based estimator)**.** *Given a sample $(X_i, Y_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{XY}$, $\epsilon_n > 0$, and $\mathbf{K}_X$, $\mathbf{K}_Y$, $\tilde{\mathbf{K}}_X$, and $\mathbf{M}$ as in* (8)*, let*

$$\hat{D}^{\text{RKHS}}(Y, X) := 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{E_{n,i}^{\text{RKHS}}}{V_{n,i}^{\text{RKHS}}},$$

where $E_{n,i}^{\text{RKHS}}$ is as in (9) *and $V_{n,i}^{\text{RKHS}}$ as in* (10)*.*

5.3 COMPUTATIONAL COMPLEXITY

In this section, we establish the runtimes of the nearest-neighbour estimator (Definition 2) and the RKHS-based estimator (Definition 3).

Let us start with the nearest-neighbour estimator. Assume that we are given access to a data structure that allows retrieving the K nearest neighbours of any X_j point in $\mathcal{O}(\log n)$, for example, a vantage point tree has been set up; the latter costs $\mathcal{O}(n \log n)$. Then, in the numerator (5), the computational cost consists of, for each $j \neq i$, retrieving the K nearest neighbours and performing $\mathcal{O}(K)$ elementary operations in the inner sum, which adds up to $\mathcal{O}(n(K+\log n))$. The computation of the denominator (6) has a cost of $\mathcal{O}(n)$. The dominant cost of the nearest neighbour estimator is thus repeating the computations in the numerator n times, which yields a total runtime complexity of $\mathcal{O}(n^2(K + \log n))$.

For the RKHS-based estimator, one must first compute the matrices in (8), costing $\mathcal{O}(n^3)$ by the matrix multiplications in $\mathbf{M}$ and by the cost of matrix inversion encountered in practice. Next, to compute the numerator (9) for $i \in [n]$, notice that the second line in (9) is independent of i; hence, while its computation costs $\mathcal{O}(n^3)$, it must only be computed once. The remaining operations in (9) are matrix-vector and vector-vector products, which have a cost of at most $\mathcal{O}(n^2)$. Similarly, the denominator (10) has a computational cost of $\mathcal{O}(n^2)$. Repeating these n times adds up to a total complexity of $\mathcal{O}(n^3)$ for the RKHS-based estimator.

While we expect that common approaches for accelerating kernel machines (incomplete Cholesky decomposition [Fine and Scheinberg, 2001, Bach and Jordan, 2002], random Fourier features [Rahimi and Recht, 2007, Sriperumbudur and Szabó, 2015], Nyström sampling [Williams and Seeger, 2001, Rudi et al., 2015, Chatalic et al., 2022, Kalinke and Szabó, 2023]) can readily be employed for accelerating the RKHS-based estimator (see also Remark 4.5 and Appendix A.4 of Huang et al. [2022]), their use in this setting does not come with theoretical guarantees.

6 CONSISTENCY AND RATE OF CONVERGENCE

In this section, we establish the consistency and convergence rate of the nearest-neighbour estimator in Definition 2, under the following assumptions.

Assumption 4. *Given a sample $\{X_i\}_{i=1}^n$, let $\delta_\ell := |\{j : \ell \in \mathcal{N}_j\}|$ for $\ell \in [n]$ where $\mathcal{N}_j$ is the set of the K-nearest neighbours of X_j in $\{X_i\}_{i\neq j}$ with ties broken at random. There exists constants $C_\mathcal{X} > 0$ such that $\delta_\ell \leq C_\mathcal{X} K$ for all $\ell \in [n]$.*

Assumption 5.

(i) There exists a finite constant $C_3 > 0$ such that $\int_\mathcal{Y} (\mathbb{V}_Y(k_\mathcal{Y}(Y,y))^{-3} \mathrm{d}\mathbb{P}_Y(y) < C_3$.

(ii) There exists an $x^* \in \mathcal{X}$, and constants $\alpha, C_1, C_2 > 0$ such that for any $t > 0$ $\mathbb{P}_X\left(d_\mathcal{X}(X, x^*) \geq t\right) \leq C_1 \exp(-C_2 t^\alpha)$.

(iii) There exist constants $d > 0$ and $c_0 > 0$ such that for every radius $T > 0$, for all $x \in \mathcal{X}$ with $d_\mathcal{X}(x^*, x) \leq T$ and all $r > 0$, $\mathbb{P}_X\left(d_\mathcal{X}(X, x) \leq r\right) \geq c_0 r^d$.

(iv) For $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, let $m_{1,y}(x) := \mathbb{E}_{Y|X=x}[k_\mathcal{Y}(Y, y)]$ and $m_{2,y}(x) := \mathbb{E}_{Y|X=x}[k^2_\mathcal{Y}(Y, y)]$. Then there exist constants $L, \beta > 0$ such that for any $y \in \mathcal{Y}$ and for any $x, x' \in \mathcal{X}$ $|m_{1,y}(x) - m_{1,y}(x')| \leq L d^\beta_\mathcal{X}(x, x')$, and $|m_{2,y}(x) - m_{2,y}(x')| \leq L d^\beta_\mathcal{X}(x, x')$.

We elaborate the assumptions in the following remarks.

Remark 4. *Assumption 4 ensures that replacing (X_ℓ, Y_ℓ) for any $\ell \in [n]$ by an i.i.d. copy (X'_ℓ, Y'_ℓ) affects only finitely many terms in $E^{K\text{-NN}}_{n,i}$. In Euclidean-like spaces, such a condition follows from Stone [1977]; in more general metric-space settings, analogous assumptions are commonly used in graph-based asymptotic analyses, e.g., Bhattacharya [2019], Huang et al. [2022].*

Remark 5. *Assumption 5 (i) can be verified analytically in simple settings. For example, suppose that $\mathcal{Y} = \mathbb{R}$, $\mathbb{P}_Y = N(0, \sigma^2)$, and $k_\mathcal{Y}(y, y') = \exp(-\gamma(y - y')^2)$ with $\gamma > 0$. If $\gamma < 1/(8\sigma^2)$, then $\int_\mathbb{R} \mathbb{V}_Y[k_\mathcal{Y}(Y, y)]^{-3} \mathrm{d}\mathbb{P}_Y(y) < \infty$ (proved in Lemma B.3). Part (ii) can be interpreted as a tail bound on X, while part (iii) asserts that the distribution of X is nowhere too thin, meaning that it is locally d-dimensional and hence d can be viewed as the intrinsic dimension. Equivalently, parts (ii)–(iii) require well-behaved tails and sufficient local mass of $\mathbb{P}_X$, so that nearest neighbours are informative. We illustrate the dependence on the intrinsic dimension in Appendix D. Part (iv) assumes that the first and second conditional moments of the canonical feature maps depend smoothly on x. This is exactly what ensures that replacing X_j by a nearby X_k introduces only an $\mathcal{O}(d^\beta_\mathcal{X}(X_j, X_k))$ bias for the K-NN estimator. As noted by Azadkia and Roudaki [2025, Lemma 4] and Deb et al. [2020, Section 5], without regularity conditions of this type, the convergence rate may be arbitrarily slow. Comparable smoothness assumptions in related settings are imposed by Dasgupta and Kpotufe [2014], Bhattacharya [2019], Deb et al. [2020].*

Under these assumptions, we obtain the following result on the rate of convergence of the nearest neighbour estimator.

Theorem 2. *Under Assumptions 1, 2, 4, and 5,*

$$|\hat{D}^{K\text{-NN}}(Y, X) - D(Y, X)| = \mathcal{O}_\mathbb{P}\left(\frac{1}{\sqrt{n}} + \left(\frac{K}{n}\right)^{\beta/d} + \frac{(\log n)^{\beta/\alpha}}{n^2}\right).$$

Theorem 2 shows that the rate of convergence of $\hat{D}^{\text{K-NN}}(Y, X)$ adapts to the intrinsic dimension d of X. For instance, the result guarantees, for suitable β, faster convergence if $\mathcal{X} = \mathbb{R}^{d_0}$ but X is only supported on a $d < d_0$ dimensional hyperplane.

7 NUMERICAL ILLUSTRATIONS

In this section, we apply our proposed estimators to simulated and real-world datasets. In real-world data analysis, we also normalize each real-valued variable to have mean 0 and variance 1.

7.1 SIMULATION STUDIES

We assess the performance of independence tests based on $\hat{D}^{\text{RKHS}}$ and $\hat{D}^{\text{K-NN}}$, benchmarking them against several competitive tests based on the related dependence measures discussed in Section 3. The competing statistics are: Chatterjee's ξ_n (estimator of ξ) correlation coefficient [Chatterjee, 2021], the T_n (estimator of ξ for the case where X can be multidimensional) statistic [Azadkia and Chatterjee, 2021], the integrated R^2 dependence measure ν_n (estimator for ν) [Azadkia and Roudaki, 2025], and the kernel measure of association in both its K-nearest-neighbour form $\hat{\eta}^{\text{K-NN}}$ and its RKHS form $\hat{\eta}^{\text{RKHS}}$ (estimators for $\eta_{k_\mathcal{Y}}$) [Huang et al., 2022]. These are computed using the `R` packages `XICOR`, `FOCI`, `FORD`, and `KPC`, respectively. Our proposed estimators are computed using the implementation provided in `https://github.com/PouyaRoudaki/KernelIR`, which also contains the code for reproducing our experiments. The sample size is $n = 100$; all p-values are obtained via 1000 independent permutations, and power is estimated from 500 simulation replications at the 5% significance level.

Euclidean data. We draw n i.i.d. samples $(X_i, Y_i)_{i=1}^n$ from a distribution on $\mathbb{R}^2$, following the experimental setup of Chatterjee [2021, Example 6.4(6)]. Indeed, $X \sim \text{Uniform}[-1, 1]$, the noise $\varepsilon \sim N(0, 1)$ is independent of X, and the alternative hypothesis is $Y = 3\big(\sigma(X)(1-\lambda)+\lambda\big)\varepsilon$, where $\sigma(X) = \mathbb{1}_{\{|X|\leq 0.5\}}$ and the noise level $\lambda \in [0, 1]$, that is, we consider a heteroscedastic setting. The statistics $\hat{\eta}^{\text{K-NN}}$, $\hat{\eta}^{\text{RKHS}}$, $\hat{D}^{\text{K-NN}}$, and $\hat{D}^{\text{RKHS}}$ all use the Gaussian kernel with the median of pairwise distances as bandwidth, and their regularization is $\epsilon_n = 10^{-4}$. The graph-based methods use $K = 5$ nearest neighbours.

Non-Euclidean data. Following Huang et al. [2022, Section 6], we additionally consider the setting where Y takes values in the special orthogonal group $\text{SO}(3)$, the manifold of 3×3 orthogonal matrices with determinant 1. We equip $\text{SO}(3)$ with the characteristic kernel $k_\mathcal{Y}(\mathbf{A}, \mathbf{B}) = \frac{\pi\theta(\pi-\theta)}{8\sin\theta}$, where $\theta \in [0, \pi]$ is defined by $\cos\theta = (\text{Tr}(\mathbf{B}^{-1}\mathbf{A}) - 1)/2$, so that $e^{\pm\sqrt{-1}\theta}$ are the eigenvalues of $\mathbf{B}^{-1}\mathbf{A}$. Let $R_1(x)$ and $R_2(z)$ denote rotations in the $y - z$ plane and $x - y$ plane with angle x and z, respectively. The predictor is $X \sim N(\mathbf{0}, \mathbf{I}_3)$. Let the independent noise variables be $\varepsilon_1, \varepsilon_2 \overset{\text{i.i.d.}}{\sim} N(0, 1)$, the noise scale $\lambda \in [0, 1]$, and the alternative hypothesis is $Y = R_1(X_1 + \lambda\varepsilon_1)R_2(X_2X_3 + \lambda\varepsilon_2) \in \text{SO}(3)$. For $\hat{D}^{\text{RKHS}}$ the kernel $k_\mathcal{X}$ is the Gaussian kernel with the median of pairwise distances as bandwidth, and their regularization parameter is set to $\epsilon_n = 10^{-4}$. Moreover, $\hat{D}^{\text{K-NN}}$ uses $K = 5$ nearest neighbours.

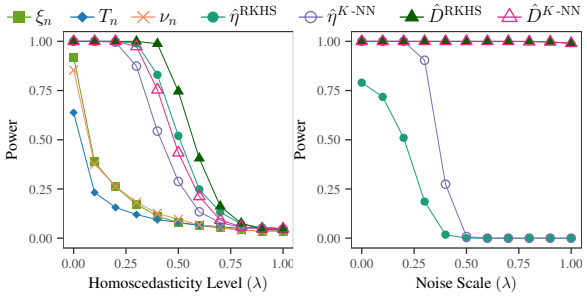

Figure 1: Comparison of power of independence tests for the heteroscedastic (left) and SO(3) (right) alternatives as a function of homoscedasticity (left) and the level of noise (right).

The results in Figure 1 show that $\hat{D}^{\text{K-NN}}$ and $\hat{D}^{\text{RKHS}}$ consistently outperform all competing tests in both the Euclidean and non-Euclidean examples. The kernel-based measures $\hat{\eta}^{\text{K-NN}}$ and $\hat{\eta}^{\text{RKHS}}$ already improve upon non-kernel methods in the heteroscedastic case—owing to their greater flexibility, as discussed in Section 3—and naturally extend to non-Euclidean data. Nevertheless, $\hat{D}^{\text{K-NN}}$ and $\hat{D}^{\text{RKHS}}$ achieve uniformly higher power while retaining the same generality.

Runtime comparison. In this experiment, we compare the runtimes of $\hat{D}^{\text{K-NN}}$ and $\hat{D}^{\text{RKHS}}$ with that of closely-related measures (elaborated in Section 3). The right plot in Figure 2 shows the average runtime of each estimator over 100 repetitions for sample size n ranging from 10 to 5000. We observe the higher runtime of RKHS-based methods compared to graph-based methods, in line with Section 5.3.

7.2 REAL DATA EXAMPLE

Million Song Dataset The Million Song Dataset [Bertin-Mahieux et al., 2011] contains 515,345 songs. Each song is described by 90 features X and its year of release Y; the latter ranges from 1992 to 2011. The objective is to detect statistical dependence between the features and the release year. To assess empirical power as a function of the sample size (ranging from 50 to 1500), we proceed as follows: for each fixed sample size n, we draw 200 independent subsamples from the full dataset and estimate the power at

significance level 0.01. The p-values are computed using 200 permutations. For the RKHS-based estimators, we use the Gaussian kernel with the median of pairwise distances as the bandwidth, and the regularization parameter $\epsilon_n = 10^{-4}$ (kept fixed for simplicity). For the graph-based estimators, we consider $K = 5$ nearest neighbours. The kernel-based methods are generally expected to provide greater flexibility in capturing complex dependence structures and to achieve higher power as it is showed in left plot of Figure 2; we also observe in this example that the RKHS-based estimator outperforms the graph-based approach.

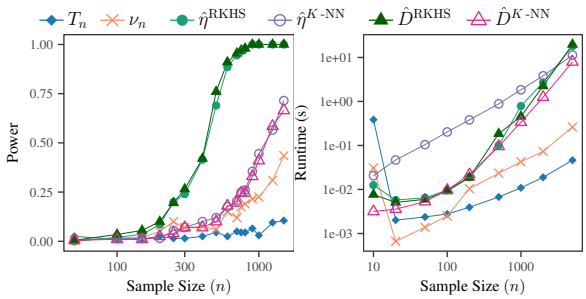

Figure 2: (Left) Comparison of power of independence tests for Million Song Data as a function of the sample size (n). (Right) Comparison of time complexity of the compared dependence measures as a function of the sample size (n).

Acknowledgements

FK is supported by the pilot program Core-Informatics of the Helmholtz Association (HGF). This work used the Cirrus UK National Tier-2 HPC Service at EPCC (http://www.cirrus.ac.uk) funded by The University of Edinburgh, the Edinburgh and South East Scotland City Region Deal, and UKRI via EPSRC.

References

Jonathan Ansari and Sebastian Fuchs. A direct extension of Azadkia & Chatterjee's rank correlation to multi-response vectors. Technical report, 2025. https://arxiv.org/abs/2212.01621.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68: 337–404, 1950.

Arnab Auddy, Nabarun Deb, and Sagnik Nandy. Exact detection thresholds and minimax optimality of Chatterjee's correlation coefficient. *Bernoulli*, 30(2):1640–1668, 2024.

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

Mona Azadkia and Pouya Roudaki. A new measure of dependence: Integrated R^2. Technical report, 2025. https://arxiv.org/abs/2505.18146.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

Wicher Bergsma and Angelos Dassios. A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 20(2):1006–1028, 2014.

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

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 591–596, 2011.

Bhaswar B. Bhattacharya. A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3):575–602, 2019.

Peter J. Bickel. Measures of independence and functional dependence. Technical report, 2022. https://arxiv.org/abs/2206.13663.

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

Patric Bonnier, Harald Oberhauser, and Zoltán Szabó. Kernelized cumulants: Beyond kernel mean embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11049–11074, 2023.

Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-López, Leslie O'Bray, and Bastian Riec. Graph kernels: State-of-the-art and future challenges. *Foundations and Trends in Machine Learning*, 13(5-6):531–712, 2020.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities*. Oxford University Press, 2013.

Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391): 580–619, 1985.

Axel Bücher and Holger Dette. On the lack of weak continuity of Chatterjee's correlation coefficient. Technical report, 2024. https://arxiv.org/abs/2410.11418.

Sky Cao and Peter J. Bickel. Correlations with tailored extremal properties. Technical report, 2020. https://arxiv.org/abs/2008.10177.

Antoine Chatalic, Nicolas Schreuder, Alessandro Rudi, and Lorenzo Rosasco. Nyström kernel mean embeddings. In *International Conference on Machine Learning (ICML)*, pages 3006–3024, 2022.

Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536): 2009–2022, 2021.

Sourav Chatterjee. A survey of some recent developments in measures of association. In *Probability and stochastic processes—a volume in honour of Rajeeva L. Karandikar*, pages 109–128. Springer, 2024.

Sándor Csörgő. Testing for independence by the empirical characteristic function. *Journal of Multivariate Analysis*, 16(3):290–299, 1985.

Sanjoy Dasgupta and Samory Kpotufe. Optimal rates for k-NN density and mode estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2555–2563, 2014.

Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 118(541):192–207, 2023.

Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Measuring association on topological spaces using kernels and geometric graphs. Technical report, 2020. https://arxiv.org/abs/2010.01768.

H. Dette and M. Kroll. A simple bootstrap for Chatterjee's rank correlation. *Biometrika*, 112(1):asae045, 2025.

Holger Dette, Karl F. Siburg, and Pavel A. Stoimenov. A copula-based non-parametric measure of regression dependence. *Scandinavian Journal of Statistics. Theory and Applications*, 40(1):21–41, 2013.

Mathias Drton, Fang Han, and Hongjian Shi. High-dimensional consistent independence testing with maxima of rank correlations. *The Annals of Statistics*, 48(6): 3206–3227, 2020.

Richard M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2004.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.

Shai Fine and Katya Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

Jerome H. Friedman and Lawrence C. Rafsky. Graph-theoretic measures of multivariate association and prediction. *The Annals of Statistics*, 11(2):377–391, 1983.

Sebastian Fuchs. Quantifying directed dependence via dimension reduction. *Journal of Multivariate Analysis*, 201: Paper No. 105266, 21, 2024.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 489–496, 2007.

Fabrice Gamboa, Thierry Klein, and Agnès Lagnoux. Sensitivity analysis based on Cramér–von Mises distance. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2): 522–548, 2018.

Fabrice Gamboa, Pierre Gremaud, Thierry Klein, and Agnès Lagnoux. Global sensitivity analysis: a novel generation of mighty estimators based on rank statistics. *Bernoulli*, 28(4):2345–2374, 2022.

Thomas Gärtner, Peter Flach, Adam Kowalczyk, and Alexander Smola. Multi-instance kernels. In *International Conference on Machine Learning (ICML)*, pages 179–186, 2002.

Hans Gebelein. Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *Zeitschrift für Angewandte Mathematik und Mechanik. Ingenieurwissenschaftliche Forschungsarbeiten*, 21:364–379, 1941.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 63–78, 2005.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 585–592, 2008.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25): 723–773, 2012.

Florian Griessenberger, Robert R. Junker, and Wolfgang Trutschnig. On a multivariate copula-based dependence measure and its estimation. *Electronic Journal of Statistics*, 16(1):2206–2251, 2022.

Jorge Guevara, Roberto Hirata, and Stéphane Canu. Cross product kernels for fuzzy set similarity. In *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2017.

Fang Han. On extensions of rank correlation coefficients to multivariate spaces. *Bernoulli News*, 28(2):7–11, 2021.

Fang Han and Zhihan Huang. Azadkia-Chatterjee's correlation coefficient adapts to manifold data. *The Annals of Applied Probability*, 34(6):5172–5210, 2024.

Fang Han, Shizhe Chen, and Han Liu. Distribution-free tests of independence in high dimensions. *Biometrika*, 104(4):813–828, 2017.

David Haussler. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz, 1999. `https://tr.soe.ucsc.edu/sites/default/files/technical-reports/UCSC-CRL-99-10.pdf`.

Ruth Heller, Yair Heller, and Malka Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2013.

H. O. Hirschfeld. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520524, 1935.

Wassily Hoeffding. A non-parametric test of independence. *Annals of Mathematical Statistics*, 19:546–557, 1948.

Wenjie Huang, Zonghan Li, and Yuhao Wang. A multivariate extension of Azadkia-Chatterjee's rank coefficient. Technical report, 2026. `https://arxiv.org/abs/2512.07443`.

Zhen Huang, Nabarun Deb, and Bodhisattva Sen. Kernel partial correlation coefficient — a measure of conditional dependence. *Journal of Machine Learning Research*, 23 (216):1–58, 2022.

Yunlong Jiao and Jean-Philippe Vert. The Kendall and Mallows kernels for permutations. In *International Conference on Machine Learning (ICML)*, pages 2982–2990, 2016.

Julie Josse and Susan Holmes. Measuring multivariate association and beyond. *Statistics Surveys*, 10:132–167, 2016.

Florian Kalinke and Zoltán Szabó. Nyström M-Hilbert-Schmidt independence criterion. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1005–1015, 2023.

Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabás Póczos, Larry Wasserman, and James M. Robins. Nonparametric von Mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 397–405, 2015.

Franz J. Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31):1–45, 2019.

Sergey Kirshner and Barnabás Póczos. ICA and ISA using Schweizer-Wolff measure of dependence. In *International Conference on Machine Learning (ICML)*, pages 464–471, 2008.

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

Efang Kong, Yingcun Xia, and Wei Zhong. Composite coefficient of determination and its application in ultrahigh dimensional variable screening. *Journal of the American Statistical Association*, 114(528):1740–1751, 2019.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E. Statistical, Nonlinear, and Soft Matter Physics*, 69(6): 066138, 16, 2004.

Marius Kroll. Asymptotic normality of Chatterjee's rank correlation. Technical report, 2025. `https://arxiv.org/abs/2408.11547`.

Z. Lin and F. Han. On boosting the power of Chatterjee's rank correlation. *Biometrika*, 110(2):283–299, 2023.

Zhexiao Lin and Fang Han. On the failure of the bootstrap for Chatterjee's rank correlation. *Biometrika*, 111(3): 1063–1070, 2024.

E. H. Linfoot. An informational measure of correlation. *Information and Control*, 1:85–89, 1957.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2: 419–444, 2002.

David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–9, 2013.

Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.

Preetam Nandy, Luca Weihs, and Mathias Drton. Large-sample theory for the Bergsma-Dassios sign covariance. *Electronic Journal of Statistics*, 10(2):2287–2311, 2016.

Wenliang Pan, Xueqin Wang, Heping Zhang, Hongtu Zhu, and Jin Zhu. Ball covariance: a generic measure of dependence in Banach space. *Journal of the American Statistical Association*, 115(529):307–317, 2020.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21247–21259, 2020.

Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):5–31, 2018.

Barnabás Póczos, Zoubin Ghahramani, and Jeff Schneider. Copula-based kernel dependency measures. In *International Conference on Machine Learning (ICML)*, pages 775–782, 2012.

Madan Lal Puri and Pranab Kumar Sen. *Nonparametric Methods in Multivariate Analysis*. John Wiley & Sons, Inc., 1971.

Dávid Pál, Barnabás Póczos, and Csaba Szepesvári. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1849–1857, 2010.

Barnabás Póczos and Jeff Schneider. On the estimation of α-divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 609–617, 2011.

Barnabás Póczos, Sergey Kirshner, and Csaba Szepesvári. REGO: Rank-based estimation of Rényi information using Euclidean graph optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 605–612, 2010.

Novi Quadrianto, Le Song, and Alex Smola. Kernelized sorting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1289–1296, 2009.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1177–1184, 2007.

A. Rényi. On measures of dependence. *Acta Mathematica. Academiae Scientiarum Hungaricae*, 10:441–451, 1959.

David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

Joseph P. Romano. A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708, 1988.

M. Rosenblatt. A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *The Annals of Statistics*, 3:1–14, 1975.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1657–1665, 2015.

B. Schweizer and E. F. Wolff. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4):879–885, 1981.

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

A. Sen and B. Sen. Testing independence and goodness-of-fit in linear models. *Biometrika*, 101(4):927–942, 2014.

H. Shi, M. Drton, and F. Han. On the power of Chatterjee's rank correlation. *Biometrika*, 109(2):317–333, 2022.

Hongjian Shi, Mathias Drton, and Fang Han. On Azadkia-Chatterjee's conditional dependence coefficient. *Bernoulli*, 30(2):851–877, 2024.

M. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

Alexander Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

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

Bharath K. Sriperumbudur and Zoltán Szabó. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1144–1152, 2015.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50): 1517–1561, 2010.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–645, 1977.

Christopher Strothmann, Holger Dette, and Karl Friedrich Siburg. Rearranged dependence measures. *Bernoulli*, 30 (2):1055–1078, 2024.

Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3 (4):1236–1265, 2009.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Leon Tran and Fang Han. On a rank-based Azadkia-Chatterjee correlation coefficient. Technical report, 2024. https://arxiv.org/abs/2412.02668.

X. Wang, B. Jiang, and J. S. Liu. Generalized R-squared for detecting dependence. *Biometrika*, 104(1):129–139, 2017.

Chris Watkins. Dynamic alignment kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 39–50, 1999.

L. Weihs, M. Drton, and N. Meinshausen. Symmetric rank covariances: a generalized framework for nonparametric measures of dependence. *Biometrika*, 105(3):547–562, 2018.

Luca Weihs, Mathias Drton, and Dennis Leung. Efficient computation of the Bergsma-Dassios sign covariance. *Computational Statistics*, 31(1):315–328, 2016.

Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 682–688, 2001.

Takemi Yanagimoto. On measures of association and a related problem. *Annals of the Institute of Statistical Mathematics*, 22(1):57–63, 1970.

Xuzhi Yang, Mona Azadkia, and Tengyao Wang. Coverage correlation: detecting singular dependencies between random variables. Technical report, 2025. https://arxiv.org/abs/2508.06402.

Kai Zhang. BET on independence. *Journal of the American Statistical Association*, 114(528):1620–1637, 2019.

Qingyang Zhang. On the asymptotic null distribution of the symmetrized Chatterjee's correlation coefficient. *Statistics & Probability Letters*, 194:Paper No. 109759, 7, 2023.

Qingyang Zhang. On relationships between Chatterjee's and Spearman's correlation coefficients. *Communications in Statistics. Theory and Methods*, 54(1):259–279, 2025.

Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

Hang Zhou and Hans-Georg Müller. Association and independence test for random objects. Technical report, 2025. https://arxiv.org/abs/2505.01983.

Kernel Integrated R^2: A Measure of Dependence (Supplementary Material)

Pouya Roudaki[1] **Shakeel Gavioli-Akilagun**[1,2] **Florian Kalinke**[3] **Mona Azadkia**[1] **Zoltán Szabó**[1]

[1]Department of Statistics, London School of Economics, London, UK
[2]Department of Decision Analytics and Operations, City University of Hong Kong, Hong Kong, China
[3]Chair of Information Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany

A PROOFS

This appendix collects our proofs. We prove Theorem 1 in Appendix A.1, derive the RKHS-based estimator (Definition 3) in Appendix A.2, and prove Theorem 2 in Appendix A.3.

A.1 PROOF OF THEOREM 1

We prove the well-definedness and parts (i)–(iii) one by one.

Proof of well-definedness. By Assumption 1, the kernel $k_{\mathcal{Y}}$ is continuous and by the blanket assumption in Section 2, it is also bounded. Hence, $k_{\mathcal{Y}}(\cdot, y)$ is Lebesgue-integrable for every $y \in \mathcal{Y}$ and so is its square; therefore the numerator is finite. It remains to show that the denominator $\mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)]$ is non-zero for all $y \in \mathcal{Y}$, which holds if $k_{\mathcal{Y}}(Y, y)$ is not a.s. constant. We prove this in the following.

Indeed, it is known that a random variable is a.s. constant iff. its distribution is degenerate. Hence, it suffices to show that the distribution of $k_{\mathcal{Y}}(Y, y)$ is non-degenerate for all $y \in \mathcal{Y}$. Let us argue by contradiction, in other words, there exists a $y \in \mathcal{Y}$ such that the distribution of $k_{\mathcal{Y}}(Y, y)$ is degenerate, that is, for some $r \in \mathbb{R}$ it holds that $\mathbb{P}_Y \circ k_{\mathcal{Y}}^{-1}(\cdot, y) = \delta_r$. Then there exists $u \in \mathcal{Y}$ such that $k_{\mathcal{Y}}(u, y) = r$; indeed, assuming that $k_{\mathcal{Y}}^{-1}(\{r\}, y) = \emptyset$ would mean that $\mathbb{P}_{\mathcal{Y}} \circ k_{\mathcal{Y}}^{-1}(\{r\}, y) = 0 \neq \delta_r(\{r\}) = 1$, which is a contradiction. Moreover, as by Assumption 1, $k_{\mathcal{Y}}(\cdot, y)$ is non-constant, there exist $v \in \mathcal{Y}$ and $s \in \mathbb{R}$ such that $k_{\mathcal{Y}}(v, y) =: s \neq r$. As $\mathbb{R}$ is Hausdorff and $r \neq s$, we can find disjoint open sets $R, S \subset \mathbb{R}$ such that $r \in R$ and $s \in S$. As $k_{\mathcal{Y}}$ is continuous, so is $k_{\mathcal{Y}}(\cdot, y)$ [Steinwart and Christmann, 2008, Lemma 4.29], which implies that also $U \coloneqq k_{\mathcal{Y}}^{-1}(S, y) \subset \mathcal{Y}$ is open. Hence, by the assumed full support of $\mathbb{P}_{\mathcal{Y}}$, $\mathbb{P}_{\mathcal{Y}}(U) > 0$. But then $\mathbb{P}_{\mathcal{Y}} \circ k_{\mathcal{Y}}^{-1}(S, y) > 0$ while $\delta_r(S) = 0$, contradicting the assumption that $\mathbb{P}_{\mathcal{Y}} \circ k_{\mathcal{Y}}^{-1}(\cdot, y) = \delta_r$. □

Proof of (i). By the law of total variance and the non-negativity of variances, we have

$$\mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)] = \mathbb{V}_X\big[\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y, y)]\big] + \underbrace{\mathbb{E}_X\big[\mathbb{V}_{Y|X}[k_{\mathcal{Y}}(Y, y)]\big]}_{\geq 0} \geq \mathbb{V}_X\big[\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y, y)]\big] \geq 0,$$

implying that

$$\mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)] \geq \mathbb{V}_X\big[\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y, y) \mid X]\big] \geq 0. \tag{A.1}$$

The combination of (A.1) and the alternative expression (B.26) in Lemma B.1 shows that $D(Y, X) \in [0, 1]$. □

Proof of (ii). ($\Longleftarrow$) We will show that in this case the numerator in (B.26) is zero, which implies that (B.26) itself is zero. The claim then follows by the equivalence of (3) and (B.26), established in Lemma B.1.

Indeed, by the assumed independence of X and Y, for any $y \in \mathcal{Y}$,

$$\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y, y)] = \mathbb{E}_Y[k_{\mathcal{Y}}(Y, y)] =: g(y),$$

and the numerator of (B.26) becomes

$$\mathbb{V}_X\left[\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y,y)]\right] = \mathbb{V}_X[g(y)] = 0,$$

proving the first direction.

($\Longrightarrow$) Assume that $D(Y,X) = 0$. Then, using the equivalence of (3) and (B.26), we have that

$$\mathbb{V}_X\left[\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y,y)]\right] = 0 \text{ for } \mathbb{P}_Y\text{-a.e. } y,$$

as the integrand of (B.26) is non-negative. Hence, for $\mathbb{P}_Y$-almost every y, $\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y,y)]$ is $\mathbb{P}_X$-almost surely constant, that is, there exists $c(y) \in \mathbb{R}$ such that

$$c(y) = \mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y,y)] \quad \mathbb{P}_X\text{-a.s. for } \mathbb{P}_Y\text{-a.e. } y. \tag{A.2}$$

Integrating the last expression w.r.t. $\mathbb{P}_X$ and using the tower property of expectations, we get

$$c(y) = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y,y)]\right] = \mathbb{E}_Y[k_{\mathcal{Y}}(Y,y)] = \mu_{k_{\mathcal{Y}}}(\mathbb{P}_Y)(y), \tag{A.3}$$

for $\mathbb{P}_Y$-almost every y. Notice that (A.2) can be written as

$$\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y,y)] = \mathbb{E}_{Y|X}[\langle k_{\mathcal{Y}}(\cdot,Y), k_{\mathcal{Y}}(\cdot,y)\rangle_{\mathcal{H}_{\mathcal{Y}}}] = \left\langle \mathbb{E}_{Y|X}[k_{\mathcal{Y}}(\cdot,Y)], k_{\mathcal{Y}}(\cdot,y)\right\rangle_{\mathcal{H}_{\mathcal{Y}}} = \mu_{k_{\mathcal{Y}}}(\mathbb{P}_{Y|X})(y), \tag{A.4}$$

by the reproducing property, Steinwart and Christmann [2008, (A.32)], and the definition of conditional mean embeddings with the reproducing property. As the l.h.s. of (A.2) and (A.3) coincide, so do the r.h.s. of (A.2) and (A.4), which shows that

$$\mu_{k_{\mathcal{Y}}}(\mathbb{P}_{Y|X})(y) = \mu_{k_{\mathcal{Y}}}(\mathbb{P}_Y)(y) \quad \mathbb{P}_X\text{-a.s. for } \mathbb{P}_Y\text{-a.e. } y,$$

which, by Remark 1(2), implies that $\mu_{k_{\mathcal{Y}}}(\mathbb{P}_{Y|X}) = \mu_{k_{\mathcal{Y}}}(\mathbb{P}_Y)$ holds $\mathbb{P}_X$-a.s. (as $\mu_{k_{\mathcal{Y}}}(\mathbb{P}_{Y|X}) - \mu_{k_{\mathcal{Y}}}(\mathbb{P}_Y) \in \mathcal{H}_{k_{\mathcal{Y}}}$ and $\mu_{k_{\mathcal{Y}}}(\mathbb{P}_{Y|X})(y) - \mu_{k_{\mathcal{Y}}}(\mathbb{P}_Y)(y) = 0$ holds $\mathbb{P}_X$-a.s. for $\mathbb{P}_Y$-a.e. y). Thus, using that $k_{\mathcal{Y}}$ is characteristic by Assumption 1,

$$\mathbb{P}_{Y|X} = \mathbb{P}_Y \quad \mathbb{P}_X\text{-a.s.} \tag{A.5}$$

To conclude the proof of (ii), we now show that (A.5) implies the independence of X and Y. Indeed, let $A \coloneqq B \times C \in \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y})$ be arbitrary, where $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y})$ denotes the product sigma-algebra of $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{Y})$. Then A, B, and C are measurable w.r.t. $\mathbb{P}_{XY}$, $\mathbb{P}_X$, and $\mathbb{P}_Y$, respectively. Using that $\mathbb{1}_A = \mathbb{1}_B \mathbb{1}_C$, by the decomposition in Dudley [2004, Theorem 10.2.1], we have

$$\begin{aligned}
\mathbb{P}_{XY}(A) &= \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}_A(x,y)\mathrm{d}\mathbb{P}_{XY}(x,y) = \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathbb{1}_A(x,y)\mathrm{d}\mathbb{P}_{Y|X=x}(y)\mathrm{d}\mathbb{P}_X(x) \\
&\overset{\text{(A.5)}}{=} \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathbb{1}_A(x,y)\mathrm{d}\mathbb{P}_Y(y)\mathrm{d}\mathbb{P}_X(x) = \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathbb{1}_B(x)\mathbb{1}_C(y)\mathrm{d}\mathbb{P}_Y(y)\mathrm{d}\mathbb{P}_X(x) \\
&= \int_{\mathcal{X}} \mathbb{1}_B(x)\mathrm{d}\mathbb{P}_X(x) \int_{\mathcal{Y}} \mathbb{1}_C(y)\mathrm{d}\mathbb{P}_Y(y) = \mathbb{P}_X(B)\mathbb{P}_Y(C),
\end{aligned}$$

that is, the joint distribution of (X,Y) factorizes to the product of the marginals, showing the independence of X and Y. □

Proof of (iii). ($\Longleftarrow$) Suppose that $Y = f(X)$ for some Borel measurable function $f : \mathcal{X} \to \mathcal{Y}$. Then, for any fixed $y \in \mathcal{Y}$, $k_{\mathcal{Y}}(Y,y) = k_{\mathcal{Y}}(f(X),y)$ is a Borel measurable function of X as the composition of measurable functions is a measurable function. Then for any $y \in \mathcal{Y}$, by the definition of the conditional variance in (a) and the properties of conditional expectations in (b)

$$\mathbb{V}_{Y|X}(k_{\mathcal{Y}}(Y,y)) \overset{(a)}{=} \mathbb{E}_{Y|X}[k^2_{\mathcal{Y}}(Y,y)] - (\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y,y)])^2 \overset{(b)}{=} k^2_{\mathcal{Y}}(Y,y) - k^2_{\mathcal{Y}}(Y,y) = 0 \text{ holds } \mathbb{P}_X\text{-a.s.}$$

hence,

$$\mathbb{E}_X[\mathbb{V}_{Y|X}(k_{\mathcal{Y}}(Y,y))] = \mathbb{E}_X[0] = 0.$$

As $y \in \mathcal{Y}$ was arbitrary, the numerator is zero everywhere, which implies that

$$D(Y,X) = 1 - \int_{\mathcal{Y}} \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}(k_{\mathcal{Y}}(Y,y))\right]}{\mathbb{V}_Y(k_{\mathcal{Y}}(Y,y))}\mathrm{d}\mathbb{P}_Y(y) = 1.$$

($\implies$) Our goal is to show that if $D(Y,X)=1$, then $Y \mid X=x$ is a.s. constant for a.e. $x \in \mathcal{X}$. This implies that $\mathbb{P}_{Y|X=x}$ is degenerate for a.e. $x \in \mathcal{X}$ and an application of Lemma C.2 then yields the claim.

Indeed, assume that $D(Y,X)=1$ and let us show that then $Y \mid X=x$ is a.s. constant for a.e. $x \in \mathcal{X}$. Rearranging D by using linearity of the integral, and flipping the integrals (permitted by Tonelli's theorem as all terms are non-negative), we obtain

$$D(Y,X) = 1 - \int_{\mathcal{Y}} \frac{\int_{\mathcal{X}} \mathbb{V}_{Y|X=x}[k_{\mathcal{Y}}(Y,y)]\,\mathrm{d}\mathbb{P}(x)}{\mathbb{V}_Y[k_{\mathcal{Y}}(Y,y)]} \mathrm{d}\mathbb{P}_Y(y) = 1 - \int_{\mathcal{X}}\int_{\mathcal{Y}} \frac{\mathbb{V}_{Y|X=x}[k_{\mathcal{Y}}(Y,y)]}{\mathbb{V}_Y[k_{\mathcal{Y}}(Y,y)]} \mathrm{d}\mathbb{P}_Y(y)\mathrm{d}\mathbb{P}(x) \overset{\text{(Ass.)}}{=} 1,$$

which shows that, given $\mathbb{P}_X$ almost any $x \in \mathcal{X}$, $\mathbb{V}_{Y|X=x}[k_{\mathcal{Y}}(Y,y)]=0$ for $\mathbb{P}_Y$-a.e. y. As the variance is zero, it must hold that

$$\mathbb{P}_{Y|X=x}(k_{\mathcal{Y}}(Y,y)=c_x(y))=1 \text{ for } \mathbb{P}_Y\text{-a.e. } y \text{ and for } \mathbb{P}_X\text{-a.e. } x,$$

where $c_x(y)$ denotes a constant depending on $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. Notice that for $\mathbb{P}_X$ almost any $x \in \mathcal{X}$ this defines a function $c_x : \mathcal{Y} \to \mathbb{R}$ for $\mathbb{P}_Y$-a.e. y.

With this setup in place, we now show that $Y \mid X=x$ is a.s. constant for a.e. $x \in \mathcal{X}$. Let us argue by contradiction, that is, suppose that with positive $\mathbb{P}_X$ probability there exists $x \in \mathcal{X}$ such that $Y \mid X=x$ is not a.s. constant. Then there exist distinct ω_1, ω_2 in the preimage of $Y \mid X=x$ satisfying $y_1 \coloneqq (Y \mid X=x)(\omega_1) \neq (Y \mid X=x)(\omega_2) =: y_2$ where y_1 and y_2 are elements in sets with positive $\mathbb{P}_{Y|X=x}$-probability. But then,

$$k_{\mathcal{Y}}(y_1,y)=c_x(y) \text{ for } \mathbb{P}_Y\text{-a.e. } y, \qquad \text{while also} \qquad k_{\mathcal{Y}}(y_2,y)=c_x(y) \text{ for } \mathbb{P}_Y\text{-a.e. } y,$$

that is $k_{\mathcal{Y}}(y_1,y)=k_{\mathcal{Y}}(y_2,y)$ for $\mathbb{P}_Y$-a.e. y. Using the assumed continuity of $k_{\mathcal{Y}}$ and the full support of $\mathbb{P}_Y$, by Remark 1(2), we obtain $k_{\mathcal{Y}}(y_1,\cdot)=k_{\mathcal{Y}}(y_2,\cdot)$, contradicting the point-separating property of the characteristic $k_{\mathcal{Y}}$. Hence, $Y \mid X=x$ is a.s. constant for a.e. $x \in \mathcal{X}$.

As indicated in the beginning of the proof of this direction, $Y \mid X=x$ being a.s. constant for a.e. $x \in \mathcal{X}$ implies that $\mathbb{P}_{Y|X=x}$ is degenerate for a.e. $x \in \mathcal{X}$. An application of Lemma C.2 concludes the proof. □

A.2 DERIVATION OF THE RKHS ESTIMATOR IN DEFINITION 3

We start by introducing the additional notations and background used in this section only.

For a kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ and a sample $Z=(Z_1,\ldots,Z_n) \in \mathcal{Z}^n$ with $Z_i \sim \mathbb{P}_Z \in \mathcal{M}_1^+(\mathcal{Z})$ ($i \in [n]$), denote by $S_{k,Z} : \mathcal{H}_k \to \mathbb{R}^n$ the sampling operator, which is defined by $h \mapsto (h(Z_i))_{i=1}^n$. It has adjoint $S_{k,Z}^* : \mathbb{R}^n \to \mathcal{H}_k$, $\boldsymbol{\alpha}=(\alpha_i)_{i=1}^n \mapsto \sum_{i=1}^n \alpha_i k(\cdot,Z_i)$ [Smale and Zhou, 2007]. Furthermore, it is known that $S_{k,Z}S_{k,Z}^* = [k(Z_i,Z_j)]_{i,j=1}^n = \mathbf{K}_Z \in \mathbb{R}^{n\times n}$. The centered (cross-)covariance operator associated to Y and X (as defined in the main part) is given by

$$C_{YX} \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} [k_{\mathcal{Y}}(\cdot,y)-\mu_{k_{\mathcal{Y}}}(\mathbb{P}_Y)] \otimes [k_{\mathcal{X}}(\cdot,x)-\mu_{k_{\mathcal{X}}}(\mathbb{P}_X)] \mathrm{d}\mathbb{P}_{XY}(x,y) \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}},$$

where, for $f \in \mathcal{H}_{\mathcal{Y}}$ and $g \in \mathcal{H}_{\mathcal{X}}$, $f \otimes g : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ denotes the rank-one operator defined by $h \mapsto f\langle g,h\rangle_{\mathcal{H}_{\mathcal{X}}}$; the operator is an element of the tensor product RKHS $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$. The centered covariance of X is its centered cross-covariance with itself, that is,

$$C_X \coloneqq C_{XX} = \int_{\mathcal{X}} [k_{\mathcal{X}}(\cdot,x)-\mu_{k_{\mathcal{X}}}(\mathbb{P}_X)] \otimes [k_{\mathcal{X}}(\cdot,x)-\mu_{k_{\mathcal{X}}}(\mathbb{P}_X)] \mathrm{d}\mathbb{P}_X(x) \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}.$$

For a bounded linear operator A, we write A^- for its (Moore-Penrose) pseudo-inverse; see, for example, Engl et al. [1996, Definition 2.2].

With the notation established, let us present the derivation. We tackle the denominator and the numerator separately.

- For the **denominator**, using that $\mathbb{V}_Y[k_{\mathcal{Y}}(Y,Y_i)] = \mathbb{E}_Y\left[k_{\mathcal{Y}}^2(Y,Y_i)\right] - \left(\mathbb{E}_Y[k_{\mathcal{Y}}(Y,Y_i)]\right)^2$ and replacing all expectations with their empirical counterparts, we obtain that

$$\mathbb{V}_Y[k_{\mathcal{Y}}(Y,Y_i)] \approx \frac{1}{n}\sum_{j=1}^n k_{\mathcal{Y}}^2(Y_j,Y_i) - \left(\frac{1}{n}\sum_{j=1}^n k_{\mathcal{Y}}(Y_j,Y_i)\right)^2 = \frac{1}{n}\mathbf{1}_n^\top(\mathbf{K}_Y \circ \mathbf{K}_Y)\mathbf{e}_i - \left(\frac{1}{n}\mathbf{1}_n^\top \mathbf{K}_Y \mathbf{e}_i\right)^2 = V_{n,i}^{\mathrm{RKHS}}. \tag{A.6}$$

- To derive the expression for the **numerator**, we first observe that given Assumption 1 and Assumption 3, by Klebanov et al. [2020, Theorem 4.3], for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$

$$\mu_{k_\mathcal{Y}}\left(\mathbb{P}_{Y|X=x}\right) = \mu_{k_\mathcal{Y}}(\mathbb{P}_Y) + C_{YX}C_X^-(k_\mathcal{X}(\cdot,x) - \mu_{k_\mathcal{X}}(\mathbb{P}_X))\,.$$

Using the plug-in estimator and the definition of $S^*_{k_\mathcal{Y},Y}$ for the first term and Huang et al. [2022, p. 48(bottom)] for the second term, we obtain, for $X = X_j$ ($j \in [n]$), the estimator

$$\hat{\mu}_{k_\mathcal{Y}}\left(\mathbb{P}_{Y|X=X_j}\right) \coloneqq \frac{1}{n}S^*_{k_\mathcal{Y},Y}\mathbf{1}_n + S^*_{k_\mathcal{Y},Y}\tilde{\mathbf{K}}_X\left(\tilde{\mathbf{K}}_X + n\epsilon_n\mathbf{I}_n\right)^{-1}\mathbf{e}_j. \tag{A.7}$$

Let $\hat{\mathbb{P}}_{X,n} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical measure associated to the observed X_i-s. Coming back to the quantity which we want to estimate, notice that, for any $i, j \in [n]$, we have

$$\mathbb{V}_{Y|X=X_j}[k_\mathcal{Y}(Y,Y_i)] = \underbrace{\mathbb{E}_{Y|X=X_j}\left[k^2_\mathcal{Y}(Y,Y_i)\right]}_{=:t_1} - \big(\underbrace{\mathbb{E}_{Y|X=X_j}[k_\mathcal{Y}(Y,Y_i)]}_{=:t_2}\big)^2,$$

and will therefore estimate

$$\mathbb{E}_{\hat{\mathbb{P}}_{X,n}}\left[\mathbb{V}_{Y|X=X_j}[k_\mathcal{Y}(Y,Y_i)]\right] = \mathbb{E}_{\hat{\mathbb{P}}_{X,n}}[t_1] - \mathbb{E}_{\hat{\mathbb{P}}_{X,n}}\left[t_2^2\right];$$

we will use (A.7) to approximate t_1 and t_2, respectively. Having obtained these approximations, we approximate the expectation of t_1 (resp. the expectation of t_2^2) by plug-in estimation.

  - **Term t_1.** One has for $i, j \in [n]$ that

$$\begin{aligned}
\mathbb{E}_{Y|X=X_j}\left[k^2_\mathcal{Y}(Y,Y_i)\right] &\overset{(a)}{=} \mathbb{E}_{Y|X=X_j}\left[\langle k_\mathcal{Y}(\cdot,Y), k_\mathcal{Y}(\cdot,Y_i)\rangle^2_{\mathcal{H}_\mathcal{Y}}\right] \\
&\overset{(b)}{=} \mathbb{E}_{Y|X=X_j}\left[\langle k_\mathcal{Y}(\cdot,Y)\otimes k_\mathcal{Y}(\cdot,Y), k_\mathcal{Y}(\cdot,Y_i)\otimes k_\mathcal{Y}(\cdot,Y_i)\rangle_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Y}}\right] \\
&\overset{(c)}{=} \left\langle \mathbb{E}_{Y|X=X_j}[k_\mathcal{Y}(\cdot,Y)\otimes k_\mathcal{Y}(\cdot,Y)], k_\mathcal{Y}(\cdot,Y_i)\otimes k_\mathcal{Y}(\cdot,Y_i)\right\rangle_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Y}} \\
&\overset{(d)}{=} \left\langle \mu_{k_\mathcal{Y}\otimes k_\mathcal{Y}}\left(\mathbb{P}_{Y|X=X_j}\right), k_\mathcal{Y}(\cdot,Y_i)\otimes k_\mathcal{Y}(\cdot,Y_i)\right\rangle_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Y}},
\end{aligned}$$

    where the reproducing property implies (a), the properties of tensor products yield (b), and Steinwart and Christmann [2008, (A.32)] allows flipping the expectation and inner product in (c). We apply the definition of the conditional mean embedding with the product kernel in (d); it is characteristic by Assumption 3 and continuous, which allows the estimation by adapting (A.7).
    Indeed, replacing $\mu_{k_\mathcal{Y}\otimes k_\mathcal{Y}}\left(\mathbb{P}_{Y|X=X_j}\right)$ by $\hat{\mu}_{k_\mathcal{Y}\otimes k_\mathcal{Y}}\left(\mathbb{P}_{Y|X=X_j}\right)$, we get

$$\begin{aligned}
t_1 &\approx \left\langle \hat{\mu}_{k_\mathcal{Y}\otimes k_\mathcal{Y}}\left(\mathbb{P}_{Y|X=X_j}\right), k_\mathcal{Y}(\cdot,Y_i)\otimes k_\mathcal{Y}(\cdot,Y_i)\right\rangle_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Y}} \overset{(a)}{=} \left\langle \hat{\mu}_{k_\mathcal{Y}\otimes k_\mathcal{Y}}\left(\mathbb{P}_{Y|X=X_j}\right), S^*_{k_\mathcal{Y}\otimes k_\mathcal{Y},Y}\mathbf{e}_i\right\rangle_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Y}} \\
&\overset{(b)}{=} \left\langle S_{k_\mathcal{Y}\otimes k_\mathcal{Y},Y}\hat{\mu}_{k_\mathcal{Y}\otimes k_\mathcal{Y}}\left(\mathbb{P}_{Y|X=X_j}\right), \mathbf{e}_i\right\rangle_{\mathbb{R}^n} \\
&\overset{(A.7)}{=} \left\langle \frac{1}{n}S_{k_\mathcal{Y}\otimes k_\mathcal{Y},Y}S^*_{k_\mathcal{Y}\otimes k_\mathcal{Y},Y}\mathbf{1}_n + S_{k_\mathcal{Y}\otimes k_\mathcal{Y},Y}S^*_{k_\mathcal{Y}\otimes k_\mathcal{Y},Y}\tilde{\mathbf{K}}_X\left(\tilde{\mathbf{K}}_X + n\epsilon_n\mathbf{I}_n\right)^{-1}\mathbf{e}_j, \mathbf{e}_i\right\rangle_{\mathbb{R}^n} \\
&\overset{(c)}{=} \left\langle \frac{1}{n}(\mathbf{K}_Y\circ\mathbf{K}_Y)\,\mathbf{1}_n + (\mathbf{K}_Y\circ\mathbf{K}_Y)\,\tilde{\mathbf{K}}_X\left(\tilde{\mathbf{K}}_X + n\epsilon_n\mathbf{I}_n\right)^{-1}\mathbf{e}_j, \mathbf{e}_i\right\rangle_{\mathbb{R}^n},
\end{aligned} \tag{A.8}$$

    by using the definition of the sampling operator in (a), the defining property of adjoint operators in (b), and $S_{k_\mathcal{Y}\otimes k_\mathcal{Y},Y}S^*_{k_\mathcal{Y}\otimes k_\mathcal{Y},Y} = \mathbf{K}_Y\circ\mathbf{K}_Y$ in (c).
    Let us now consider the outer expectation $\mathbb{E}_{\hat{\mathbb{P}}_{X,n}}$ of t_1's approximation. We sum (A.8) over $j \in [n]$, divide by n, and, by the linearity of the inner product and as $\frac{1}{n}\sum_{j=1}^n \mathbf{e}_j = \frac{1}{n}\mathbf{1}_n$, obtain

$$\mathbb{E}_{\hat{\mathbb{P}}_{X,n}}[t_1] \approx \left\langle \frac{1}{n}(\mathbf{K}_Y\circ\mathbf{K}_Y)\,\mathbf{1}_n + \frac{1}{n}(\mathbf{K}_Y\circ\mathbf{K}_Y)\,\tilde{\mathbf{K}}_X\left(\tilde{\mathbf{K}}_X + n\epsilon_n\mathbf{I}_n\right)^{-1}\mathbf{1}_n, \mathbf{e}_i\right\rangle_{\mathbb{R}^n} \tag{A.9}$$

– **Term t_2.** We have for $i, j \in [n]$ that

$$\mathbb{E}_{Y|X=X_j}[k_{\mathcal{Y}}(Y, Y_i)] \overset{(a)}{=} \mathbb{E}_{Y|X=X_j}\left[\langle k_{\mathcal{Y}}(\cdot, Y), k_{\mathcal{Y}}(\cdot, Y_i)\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] \overset{(b)}{=} \left\langle \mathbb{E}_{Y|X=X_j}[k_{\mathcal{Y}}(\cdot, Y)], k_{\mathcal{Y}}(\cdot, Y_i)\right\rangle_{\mathcal{H}_{\mathcal{Y}}}$$
$$\overset{(c)}{=} \left\langle \mu_{k_{\mathcal{Y}}}\left(\mathbb{P}_{Y|X=X_j}\right), k_{\mathcal{Y}}(\cdot, Y_i)\right\rangle_{\mathcal{H}_{\mathcal{Y}}},$$

by using the reproducing property in (a), flipping the integral and the inner product using Steinwart and Christmann [2008, (A.32)] in (b), and by the definition of the conditional mean embedding in (c).
Again replacing $\mu_{k_{\mathcal{Y}}}\left(\mathbb{P}_{Y|X=X_j}\right)$ by its empirical counterpart $\hat{\mu}_{k_{\mathcal{Y}}}\left(\mathbb{P}_{Y|X=X_j}\right)$, we obtain the approximation

$$\begin{aligned}
t_2 &\approx \left\langle \hat{\mu}_{k_{\mathcal{Y}}}\left(\mathbb{P}_{Y|X=X_j}\right), k_{\mathcal{Y}}(\cdot, Y_i)\right\rangle_{\mathcal{H}_{\mathcal{Y}}} \overset{(a)}{=} \left\langle \hat{\mu}_{k_{\mathcal{Y}}}\left(\mathbb{P}_{Y|X=X_j}\right), S^*_{k_{\mathcal{Y}},Y}\mathbf{e}_i\right\rangle_{\mathcal{H}_{\mathcal{Y}}} \\
&\overset{(b)}{=} \left\langle S_{k_{\mathcal{Y}},Y}\hat{\mu}_{k_{\mathcal{Y}}}\left(\mathbb{P}_{Y|X=X_j}\right), \mathbf{e}_i\right\rangle_{\mathbb{R}^n} \\
&\overset{(A.7)}{=} \left\langle \frac{1}{n} S_{k_{\mathcal{Y}},Y} S^*_{k_{\mathcal{Y}},Y}\mathbf{1}_n + S_{k_{\mathcal{Y}},Y} S^*_{k_{\mathcal{Y}},Y}\tilde{\mathbf{K}}_X\left(\tilde{\mathbf{K}}_X + n\epsilon_n \mathbf{I}_n\right)^{-1}\mathbf{e}_j, \mathbf{e}_i\right\rangle_{\mathbb{R}^n} \\
&\overset{(c)}{=} \left\langle \frac{1}{n}\mathbf{K}_Y\mathbf{1}_n + \mathbf{K}_Y\tilde{\mathbf{K}}_X\left(\tilde{\mathbf{K}}_X + n\epsilon_n \mathbf{I}_n\right)^{-1}\mathbf{e}_j, \mathbf{e}_i\right\rangle_{\mathbb{R}^n} \overset{(d)}{=} \left\langle \frac{1}{n}\mathbf{K}_Y\mathbf{1}_n + \mathbf{M}\mathbf{e}_j, \mathbf{e}_i\right\rangle_{\mathbb{R}^n} \\
&\overset{(e)}{=} \frac{1}{n}\mathbf{e}_i^\top \mathbf{K}_Y\mathbf{1}_n + \mathbf{e}_i^\top \mathbf{M}\mathbf{e}_j,
\end{aligned} \tag{A.10}$$

where (a) is by the definition of the sampling operator, (b) comes from the definition of the adjoint operator, and in (c), we use that $S_{k_{\mathcal{Y}},Y} S^*_{k_{\mathcal{Y}},Y} = \mathbf{K}_Y$, (d) comes from the definition of $\mathbf{M}$ in (8), (e) follows from the definition and the linearity of the inner product in $\mathbb{R}^n$. Squaring (A.10) gives

$$\begin{aligned}
t_2^2 &\approx \frac{1}{n^2}\mathbf{e}_i^\top \mathbf{K}_Y\mathbf{1}_n\mathbf{e}_i^\top \mathbf{K}_Y\mathbf{1}_n + \frac{2}{n}\mathbf{e}_i^\top \mathbf{K}_Y\mathbf{1}_n\mathbf{e}_i^\top \mathbf{M}\mathbf{e}_j + \mathbf{e}_i^\top \mathbf{M}\mathbf{e}_j\mathbf{e}_i^\top \mathbf{M}\mathbf{e}_j \\
&= \frac{1}{n^2}\mathbf{e}_i^\top \mathbf{K}_Y\mathbf{1}_n\mathbf{1}_n^\top \mathbf{K}_Y\mathbf{e}_i + \frac{2}{n}\mathbf{e}_i^\top \mathbf{K}_Y\mathbf{1}_n\mathbf{e}_j^\top \mathbf{M}^\top\mathbf{e}_i + \mathbf{e}_i^\top \mathbf{M}\mathbf{e}_j\mathbf{e}_j^\top \mathbf{M}^\top\mathbf{e}_i,
\end{aligned} \tag{A.11}$$

where we use that a real number equals its transpose and the symmetry of Gram matrices.
To get the outer expectation, we sum (A.11) over $j \in [n]$, divide by n, and obtain

$$\mathbb{E}_{\hat{\mathbb{P}}_{X,n}}\left[t_2^2\right] \approx \frac{1}{n^2}\mathbf{e}_i^\top \mathbf{K}_Y\mathbf{1}_n\mathbf{1}_n^\top \mathbf{K}_Y\mathbf{e}_i + \frac{2}{n^2}\mathbf{e}_i^\top \mathbf{K}_Y\mathbf{1}_n\mathbf{1}_n^\top \mathbf{M}^\top\mathbf{e}_i + \frac{1}{n}\mathbf{e}_i^\top \mathbf{M}\mathbf{I}_n\mathbf{M}^\top\mathbf{e}_i. \tag{A.12}$$

Hence, subtracting (A.12) from (A.9) and rearranging, we have for the numerator

$$\begin{aligned}
E^{\text{RKHS}}_{n,i} = \frac{1}{n}\Big\langle \mathbf{e}_i, &\left(\mathbf{K}_Y \circ \mathbf{K}_Y\right)\mathbf{1}_n + \left(\mathbf{K}_Y \circ \mathbf{K}_Y\right)\tilde{\mathbf{K}}_X\left(\tilde{\mathbf{K}}_X + n\epsilon_n\mathbf{I}_n\right)^{-1}\mathbf{1}_n \\
&- \frac{1}{n}\mathbf{K}_Y\mathbf{1}_n\mathbf{1}_n^\top \mathbf{K}_Y\mathbf{e}_i - \frac{2}{n}\mathbf{K}_Y\mathbf{1}_n\mathbf{1}_n^\top \mathbf{M}^\top\mathbf{e}_i - \mathbf{M}\mathbf{M}^\top\mathbf{e}_i\Big\rangle_{\mathbb{R}^n}.
\end{aligned} \tag{A.13}$$

Combining the numerator (A.13) and denominator (A.6) concludes the derivation.

A.3 PROOF OF THEOREM 2

Proof. In the sequel we let C denote an absolute constant which may change from line to line, and we use the symbol $\lesssim$ to indicate weak inequality up to an absolute constant which again may change from line to line. We will also denote the complement of an event A by A^c. Finally, we recall that in the main text the kernel $k_{\mathcal{Y}}$ is assumed to be bounded and let $\kappa \coloneqq \sup_{y,y' \in \mathcal{Y}} k_{\mathcal{Y}}(y, y')$. For each $i \in [n]$, remember that

$$E^{\text{K-NN}}_{n,i} = \frac{1}{2K(n-1)}\sum_{j \neq i}\sum_{l \in \mathcal{N}_j^{\setminus i}}\left(k_{\mathcal{Y}}(Y_j, Y_i) - k_{\mathcal{Y}}(Y_l, Y_i)\right)^2,$$

$$V^{\text{K-NN}}_{n,i} = \frac{1}{n-1}\sum_{j \neq i} k^2_{\mathcal{Y}}(Y_j, Y_i) - \Big[\frac{1}{n-1}\sum_{j \neq i} k_{\mathcal{Y}}(Y_j, Y_i)\Big]^2.$$

Additionally for $i \in [n]$ let

$$V_i \coloneqq \mathbb{V}_Y^{\setminus i}[k_{\mathcal{Y}}(Y, Y_i)], \qquad E_i \coloneqq \mathbb{E}_X^{\setminus i}\mathbb{V}_{Y|X}^{\setminus i}[k_{\mathcal{Y}}(Y, Y_i)], \tag{A.14}$$

where (X_i, Y_i) is treated as a fixed observed value and the expectation is taken over (X, Y). With this, let

$$Q \coloneqq \int_{\mathcal{Y}} \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}\left[k_{\mathcal{Y}}\left(Y, y\right)\right]\right]}{\mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)]} \mathrm{d}\mathbb{P}_Y(y), \qquad \hat{Q}_n \coloneqq \frac{1}{n}\sum_{i=1}^n \frac{E_{n,i}^{\text{K-NN}}}{V_{n,i}^{\text{K-NN}}}, \qquad \hat{Q}'_n \coloneqq \frac{1}{n}\sum_{i=1}^n \frac{E_i^{\text{K-NN}}}{V_i}, \qquad \hat{Q}''_n \coloneqq \frac{1}{n}\sum_{i=1}^n \frac{E_i}{V_i}. \tag{A.15}$$

Note that $|\hat{D}_n^{\text{K-NN}} - D| = |Q - \hat{Q}_n|$. Introduce the event

$$\Omega_n = \bigcap_{i=1}^n \left\{ |V_{n,i}^{\text{K-NN}} - V_i| \leq \frac{1}{2} V_i \right\}. \tag{A.16}$$

Using triangle inequality, we have

$$|Q - \hat{Q}_n| \leq |Q - \hat{Q}''_n| + |\hat{Q}''_n - \hat{Q}'_n| + |\hat{Q}'_n - \hat{Q}_n|. \tag{A.17}$$

Since $V_{n,i}^{\text{K-NN}}$ and V_i are non-negative, Ω_n implies $\frac{1}{2}V_i \geq V_{n,i}^{\text{K-NN}}$ for all $i \in [n]$. Therefore, conditional on the event Ω_n it holds that

$$|\hat{Q}_n - \hat{Q}'_n| = \left| \frac{1}{n}\sum_{i=1}^n (V_{n,i}^{\text{K-NN}} - V_i) \frac{E_{n,i}^{\text{K-NN}}}{V_i V_{n,i}^{\text{K-NN}}} \right| \leq \frac{1}{n}\sum_{i=1}^n |V_{n,i}^{\text{K-NN}} - V_i| \frac{E_{n,i}^{\text{K-NN}}}{V_i V_{n,i}^{\text{K-NN}}} \leq \frac{2}{n}\sum_{i=1}^n |V_{n,i}^{\text{K-NN}} - V_i| \frac{E_{n,i}^{\text{K-NN}}}{V_i^2} \coloneqq \tilde{Q}_n.$$

Therefore, for any $\delta > 0$ it holds that

$$\begin{aligned} \mathbb{P}(|\hat{Q}_n - Q| > \delta) &\leq \mathbb{P}(|Q - \hat{Q}''_n| + |\hat{Q}''_n - \hat{Q}'_n| + |\hat{Q}'_n - \hat{Q}_n| > \delta \mid \Omega_n)\mathbb{P}(\Omega_n) + \mathbb{P}(\Omega_n^c) \\ &\leq \mathbb{P}(|Q - \hat{Q}''_n| + |\hat{Q}''_n - \hat{Q}'_n| + \tilde{Q}_n > \delta) + \mathbb{P}(\Omega_n^c). \end{aligned} \tag{A.18}$$

In the following we bound each of the terms appearing in (A.18), where we frequently use the following equality:

$$\mathbb{E}[|Z|] = \int_0^\infty \mathbb{P}(|Z| > t)\mathrm{d}t, \tag{A.19}$$

which holds for any real-valued random variable Z. For the first term in (A.18) observe that by the law of total variance, for $i \in [n]$ we have $E_i/V_i \in [0, 1]$. Additionally E_i/V_i is an i.i.d. sequence. Therefore by Hoeffding's inequality we have

$$\mathbb{P}(|\hat{Q}''_n - \mathbb{E}[\hat{Q}''_n]| > t) \leq 2e^{-2nt^2}.$$

Moreover $\mathbb{E}[\hat{Q}''_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[E_i/V_i] = Q$. Therefore

$$\mathbb{E}|Q - \hat{Q}''_n| = \mathbb{E}|\hat{Q}''_n - \mathbb{E}[\hat{Q}''_n]| = \int_0^\infty \mathbb{P}(|\hat{Q}''_n - \mathbb{E}[\hat{Q}''_n]| > t)\mathrm{d}t \leq 2\int_0^\infty e^{-2nt^2}\mathrm{d}t \lesssim \frac{1}{\sqrt{n}},$$

and by Markov's inequality $|Q - \hat{Q}''_n| = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$. For the second term in (A.18), observe first that for each $i \in [n]$, due to the boundedness of $k_{\mathcal{Y}}$ the statistic $V_{n,i}^{\text{K-NN}}$ enjoys the bounded difference property (see Definition C.29) with absolute finite difference bounded from above by $(n-1)^{-1}$ up to a multiplicative constant depending on κ. Therefore, by Theorem C.1 we have that

$$\mathbb{P}^{\setminus i}(|V_{n,i}^{\text{K-NN}} - \mathbb{E}^{\setminus i}[V_{n,i}^{\text{K-NN}}]| > t) \leq 2\exp(-Cnt^2) \quad \forall t > 0, \tag{A.20}$$

where the term on the right does not depend on Y_i. Additionally for fixed i note that $\{k_{\mathcal{Y}}(Y_j, Y_i)\}_{j \neq i}$ is an i.i.d. sequence of bounded random variables. Therefore, by standard result, the bias of the estimated variance for each $i \in [n]$ is

$$|\mathbb{E}^{\setminus i}[V_{n,i}^{\text{K-NN}}] - V_i| = \frac{1}{n-1} V_i \lesssim \frac{1}{n}, \tag{A.21}$$

where the final inequality follows from the boundedness of $k_{\mathcal{Y}}$. Consequently we have that

$$\mathbb{E}[\tilde{Q}_n] \lesssim \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{|V_{n,i}^{\text{K-NN}} - V_i|}{V_i^2}\right] \tag{A.22a}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{|V_{n,i}^{\text{K-NN}} - \mathbb{E}[V_{n,i}^{\text{K-NN}}]|}{V_i^2} + \frac{|\mathbb{E}[V_{n,i}^{\text{K-NN}}] - V_i|}{V_i^2}\right]$$

$$\lesssim \frac{1}{n}\int_{\mathcal{Y}}\Big(\mathbb{V}_Y\big(k_{\mathcal{Y}}(Y,y)\big)\Big)^{-2}\mathrm{d}\mathbb{P}_Y(y)$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\int_0^{\infty}\mathbb{P}^{\setminus i}\big(|V_{n,i}^{\text{K-NN}} - V_i| > tV_i^2\big)\mathrm{d}t\Big] \tag{A.22b}$$

$$\leq \frac{1}{n}\int_{\mathcal{Y}}\Big(\mathbb{V}_Y\big(k_{\mathcal{Y}}(Y,y)\big)\Big)^{-2}\mathrm{d}\mathbb{P}_Y(y) + \int_{\mathcal{Y}}\int_0^{\infty}\exp\Big(-Cnt^2\big(\mathbb{V}_Y[k_{\mathcal{Y}}(Y,y)]\big)^4\Big)\mathrm{d}t\mathrm{d}\mathbb{P}_Y(y) \tag{A.22c}$$

$$\lesssim \Big\{\frac{1}{n} + \frac{1}{\sqrt{n}}\Big\}\int_{\mathcal{Y}}\big(\mathbb{V}_Y[k_{\mathcal{Y}}(Y,y)]\big)^{-2}\mathrm{d}\mathbb{P}_Y(y) \lesssim \frac{1}{\sqrt{n}}. \tag{A.22d}$$

In particular (A.22a) follows from the boundedness of $E_{n,i}^{\text{K-NN}}$ for all $i \in [n]$, which holds due to the boundedness of $k_{\mathcal{Y}}$, (A.22b) follows (A.19), (A.22c) follows from (A.20), and (A.22d) follows from part (i) of Assumption 5. Consequently it holds that $\tilde{Q}_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$. For term $|\hat{Q}_n' - \hat{Q}_n''|$ using triangle inequality and then Jensen's inequality we have

$$\mathbb{E}[|\hat{Q}_n' - \hat{Q}_n''|] \leq \mathbb{E}\left[\frac{\sqrt{\mathbb{V}^{\setminus i}(E_{n,i}^{\text{K-NN}})}}{V_i}\right] + \mathbb{E}\left[\frac{|\mathbb{E}^{\setminus i}[E_{n,i}^{\text{K-NN}}] - E_i|}{V_i}\right].$$

For the variance term, note that by Efron-Stein inequality we have

$$\mathbb{V}^{\setminus i}\left(E_{n,i}^{\text{K-NN}}\right) \leq \frac{1}{2}\sum_{\ell\neq i}\mathbb{E}^{\setminus i}\left[\left(E_{n,i}^{\text{K-NN}} - E_{n,i,\ell}^{\text{K-NN}}\right)^2\right],$$

where $E_{n,i,\ell}^{\text{K-NN}}$ is calculated using sample $\{(X_i,Y_i)\}_{i\neq\ell}\cup\{(X_\ell',Y_\ell')\}$, such that (X_ℓ',Y_ℓ') is an i.i.d. copy of (X_ℓ,Y_ℓ). Note that in replacing (X_ℓ,Y_ℓ) by the i.i.d. copy (X_ℓ',Y_ℓ'), observation ℓ appears in two roles:

1. As center $j=\ell$: in this case all K terms in the neighbourhood ℓ can change which contributes to at most $8K\kappa^2/2K(n-1) = 4\kappa^2/(n-1)$,
2. As neigbour $k=\ell$: in this case by Assumption 4, X_ℓ is at most in the neighbourhood of CK other X_j's and the therefore the total contributed value of this change is $8CK\kappa^2/2K(n-1) = 4C\kappa^2/(n-1)$.

Therefore

$$|E_{n,i}^{\text{K-NN}} - E_{n,i,\ell}^{\text{K-NN}}| \leq \frac{4\kappa^2}{n-1}(1+C).$$

Hence

$$\mathbb{V}^{\setminus i}\left(E_{n,i}^{\text{K-NN}}\right) \lesssim \frac{1}{n}.$$

This gives us

$$\mathbb{E}\left[\frac{\sqrt{\mathbb{V}^{\setminus i}(E_{n,i}^{\text{K-NN}})}}{V_i}\right] \lesssim \frac{1}{\sqrt{n}}\mathbb{E}\left[\frac{1}{V_i}\right].$$

For the bias term, note that

$$|\mathbb{E}^{\setminus i}[E_{n,i}^{\text{K-NN}}] - E_i| = \left|\mathbb{E}^{\setminus i}\left[\frac{1}{K}\sum_{k\in\mathcal{N}_i^{\setminus j}}(k_{\mathcal{Y}}(Y_j,Y_i) - k_{\mathcal{Y}}(Y_k,Y_i))^2 - \mathbb{E}_{Y|X}^{\setminus i}\left[(k_{\mathcal{Y}}(Y,Y_i) - k_{\mathcal{Y}}(Y',Y_i))^2\right]\right]\right|,$$

where Y and Y' are i.i.d. conditional on X. Let

$$m_1(x) \coloneqq \mathbb{E}^{\setminus i}[k_{\mathcal{Y}}(Y, Y_i) \mid X = x], \quad m_2(x) \coloneqq \mathbb{E}[k^2_{\mathcal{Y}}(Y, Y_i) \mid X = x].$$

Then

$$\mathbb{E}^{\setminus i}\left[\left(k_{\mathcal{Y}}(Y, Y_i) - k_{\mathcal{Y}}(Y', Y_i)\right)^2 \mid X = x\right] = 2\mathbb{V}^{\setminus i}(k_{\mathcal{Y}}(Y, Y_i) \mid X = x) = 2(m_2(x) - m_1(x)^2),$$

and

$$\mathbb{E}^{\setminus i}\left[\left(k_{\mathcal{Y}}(Y_j, Y_i) - k_{\mathcal{Y}}(Y_k, Y_i)\right)^2 \mid X_j = x, X_k = x'\right] = m_2(x) + m_2(x') - 2m_1(x)m_1(x').$$

Let

$$\begin{aligned}\Delta\left(x, x'\right) &\coloneqq \mathbb{E}^{\setminus i}\left[\left(k_{\mathcal{Y}}(Y_j, Y_i) - k_{\mathcal{Y}}(Y_k, Y_i)\right)^2 \mid x, x'\right] - \mathbb{E}^{\setminus i}\left[\left(k_{\mathcal{Y}}(Y, Y_i) - k_{\mathcal{Y}}(Y', Y_i)\right)^2 \mid X = x\right] \\ &= \left(m_2(x') - m_2(x)\right) - 2m_1(x)\left(m_1(x') - m_1(x)\right),\end{aligned}$$

which gives us

$$\left|\Delta\left(x, x'\right)\right| \leq \left|m_2(x') - m_2(x)\right| - 2\kappa\left|m_1(x') - m_1(x)\right| \lesssim d_{\mathcal{X}}(x, x')^{\beta}.$$

where the last inequality is by Assumption 5 . Then by taking average over all the neighbours and using Lemma B.2 together with Assumption 5 we have

$$\left|\mathbb{E}^{\setminus i}[E^{\text{K-NN}}_{n,i}] - E_i\right| \lesssim \mathbb{E}^{\setminus i}\left[\max_{k \in \mathcal{N}_j^{\setminus i}} d^{\beta}_{\mathcal{X}}(X_k, X_i)\right] \lesssim \left[\left(\frac{K}{n}\right)^{\beta/d} + n^{-2}(\log n)^{\beta/\alpha}\right].$$

Consequently,

$$\mathbb{E}\left[\frac{\left|\mathbb{E}^{\setminus i}[E^{\text{K-NN}}_{n,i}] - E_i\right|}{V_i}\right] \lesssim \left[\left(\frac{K}{n}\right)^{\beta/d} + n^{-2}(\log n)^{\beta/\alpha}\right]\mathbb{E}\left[\frac{1}{V_i}\right] \lesssim \left[\left(\frac{K}{n}\right)^{\beta/d} + n^{-2}(\log n)^{\beta/\alpha}\right]. \tag{A.23}$$

Finally for $\mathbb{P}(\Omega_n^c)$, we have that

$$\begin{aligned}\mathbb{P}(\Omega_n^c) &\leq \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{P}^{\setminus i}\left(\left|\hat{V}^{\text{K-NN}}_{n,i} - V_i\right| \geq \frac{1}{2}V_i\right)\right] \\ &\leq \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{P}^{\setminus i}\left(\left|\hat{V}^{\text{K-NN}}_{n,i} - \mathbb{E}^{\setminus i}[\hat{V}^{\text{K-NN}}_{n,i}]\right| + \left|\mathbb{E}^{\setminus i}[\hat{V}^{\text{K-NN}}_{n,i}] - V_i\right| \geq \frac{1}{2}V_i\right)\right]\end{aligned}$$

Note that by (A.21) we have

$$\mathbb{P}^{\setminus i}\left(\left|\hat{V}^{\text{K-NN}}_{n,i} - \mathbb{E}^{\setminus i}[\hat{V}^{\text{K-NN}}_{n,i}]\right| + \left|\mathbb{E}^{\setminus i}[\hat{V}^{\text{K-NN}}_{n,i}] - V_i\right| \geq \frac{1}{2}V_i\right) \leq \mathbb{P}^{\setminus i}\left(\left|\hat{V}^{\text{K-NN}}_{n,i} - \mathbb{E}^{\setminus i}[\hat{V}^{\text{K-NN}}_{n,i}]\right| \geq \left(\frac{1}{2} - \frac{1}{n-1}\right)V_i\right),$$

therefore for $n \geq 4$ we can write

$$\mathbb{P}(\Omega_n^c) \leq \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{P}^{\setminus i}\left(\left|\hat{V}^{\text{K-NN}}_{n,i} - \mathbb{E}[\hat{V}^{\text{K-NN}}_{n,i}]\right| \geq \frac{1}{6}V_i\right)\right] \tag{A.25a}$$

$$\leq n\mathbb{E}\left[\exp\left(-\frac{1}{36}CnV_i^2\right)\right] \tag{A.25b}$$

$$\lesssim \frac{1}{\sqrt{n}}\mathbb{E}\left[\frac{1}{V_i^3}\right] \lesssim \frac{1}{\sqrt{n}}, \tag{A.25c}$$

where (A.25b) follows from (A.20), and finally from $e^{-x} < x^{-3/2}$ for $x > 0$, we have (A.25c).

We already showed $|Q - \hat{Q}_n''| = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$, $\tilde{Q}_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$, and $\mathbb{P}(\Omega_n^c) \lesssim n^{-1/2}$, therefore by (A.23) we have

$$\begin{aligned}\mathbb{P}(|\hat{Q}_n - Q| > \delta) &\leq \mathbb{P}(|Q - \hat{Q}_n''| + |\hat{Q}_n'' - \hat{Q}_n'| + \tilde{Q}_n > \delta) + \mathbb{P}(\Omega_n^c) \\ &\lesssim \mathbb{P}(|Q - \hat{Q}_n''| > \delta/3) + \mathbb{P}(|\hat{Q}_n'' - \hat{Q}_n'| > \delta/3) + \mathbb{P}(\tilde{Q}_n > \delta/3) + \frac{1}{\sqrt{n}} \\ &\lesssim \frac{1}{\delta}\left(\frac{1}{\sqrt{n}} + \left(\frac{K}{n}\right)^{\beta/d} + (\log n)^{\beta/\alpha} n^{-2}\right),\end{aligned}$$

which gives us

$$|\hat{D}^{\text{K-NN}}(X, Y) - D(X, Y)| = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + \left(\frac{K}{n}\right)^{\beta/d} + (\log n)^{\beta/\alpha} n^{-2}\right),$$

and completes the proof. □

B AUXILIARY RESULTS

This section collects our auxiliary results, used in the proofs of the results stated in the main text.

Lemma B.1 (Alternative expression)**.** *In the setting of Definition 1, it holds that*

$$D(Y, X) = \int_{\mathcal{Y}} \frac{\mathbb{V}_X\left[\mathbb{E}_{Y|X}[k_{\mathcal{Y}}(Y, y)]\right]}{\mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)]} \mathrm{d}\mathbb{P}_Y(y). \tag{B.26}$$

Proof. We have the chain of equalities

$$\begin{aligned}D(Y, X) &\overset{(3)}{=} 1 - \int_{\mathcal{Y}} \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}\left[k_{\mathcal{Y}}(Y, y)\right]\right]}{\mathbb{V}_Y\left[k_{\mathcal{Y}}(Y, y)\right]} \mathrm{d}\mathbb{P}_Y(y) \overset{(a)}{=} 1 + \int_{\mathcal{Y}} \frac{\pm\mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)] - \mathbb{E}_X\left[\mathbb{V}_{Y|X}\left[k_{\mathcal{Y}}(Y, y)\right]\right]}{\mathbb{V}_Y\left[k_{\mathcal{Y}}(Y, y)\right]} \mathrm{d}\mathbb{P}_Y(y) \\ &\overset{(b)}{=} 1 + \int_{\mathcal{Y}} \frac{\mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)] - \mathbb{E}_X\left[\mathbb{V}_{Y|X}\left[k_{\mathcal{Y}}(Y, y)\right]\right]}{\mathbb{V}_Y\left[k_{\mathcal{Y}}(Y, y)\right]} \mathrm{d}\mathbb{P}_Y(y) - 1 \overset{(c)}{=} \int_{\mathcal{Y}} \frac{\mathbb{V}_X\left[\mathbb{E}_{Y|X}\left[k_{\mathcal{Y}}(Y, y)\right]\right]}{\mathbb{V}_Y\left[k_{\mathcal{Y}}(Y, y)\right]} \mathrm{d}\mathbb{P}_Y(y),\end{aligned}$$

where in (a), we add zero; in (b), we split the fraction and simplify; and in (c), we use that $1 - 1 = 0$ together with

$$\mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)] = \mathbb{E}_X\left[\mathbb{V}_{Y|X}\left[k_{\mathcal{Y}}(Y, y)\right]\right] + \mathbb{V}_X\left[\mathbb{E}_{Y|X}\left[k_{\mathcal{Y}}(Y, y)\right]\right]$$

by the law of total variance ($y \in \mathcal{Y}$). □

Lemma B.2 (Nearest-neighbour distance)**.** *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space and $X_1, \ldots, X_n$ i.i.d. Let d_j be the distance from X_j to its K-th nearest neighbour among $\{X_\ell : \ell \neq j\}$, i.e.*

$$d_j \coloneqq \max_{k \in \mathcal{N}_j} d_{\mathcal{X}}(X_k, X_j)$$

where $\mathcal{N}_j$ is the set of K-nearest neighbours of X_j in $\{X_k\}_{k \neq j}$. Assume:

(A1) There exist $x^* \in \mathcal{X}$ and $\alpha, C_1, C_2 > 0$ such that for all $t \geq 0$,

$$\mathbb{P}\left(d_{\mathcal{X}}(X_1, x^*) \geq t\right) \leq C_1 e^{-C_2 t^\alpha}.$$

(A2) There exist constants $d > 0$ and $c_0 > 0$ such that for every radius $T > 0$, for all $x \in \mathcal{X}$ with $d_{\mathcal{X}}(x^*, x) \leq T$ and all $r > 0$,

$$\mathbb{P}\left(d_{\mathcal{X}}(X_1, x) \leq r\right) \geq c_0 r^d$$

Let $\beta > 0$. Then for all $n \geq 3$ and $1 \leq K \leq n/2$,

$$\mathbb{E}[d_j^\beta] \lesssim \left(\frac{K}{n}\right)^{\beta/d} + (\log n)^{\beta/\alpha} n^{-2},$$

where $\lesssim$ hides constants depending only on $\alpha, \beta, d, c_0, C_1, C_2$.

Proof of Lemma B.2. Fix $\delta \in (0,1)$ and define

$$T_{n,\delta} \coloneqq \left(\frac{1}{C_2}\log\frac{C_1 n}{\delta}\right)^{1/\alpha}.$$

By (A1) and a union bound,

$$\mathbb{P}\left(\max_{1\le\ell\le n} d_{\mathcal{X}}\left(X_\ell, x^*\right) > T_{n,\delta}\right) \le nC_1 e^{-C_2 T_{n,\delta}^{\alpha}} = \delta.$$

Introduce the event

$$\mathcal{G}_\delta \coloneqq \left\{\max_{1\le\ell\le n} d_{\mathcal{X}}(X_\ell, x^*) \le T_{n,\delta}\right\}.$$

Then $\mathbb{P}\left(\mathcal{G}_\delta\right) \ge 1-\delta$. On $\mathcal{G}_\delta$, for all sample points X_i we have $d_{\mathcal{X}}(X_i, x^*) \le T_{n,\delta}$, hence for any j,

$$d_j \le 2T_{n,\delta}.$$

Therefore, we have

$$\mathbb{E}[d_j^\beta] = \mathbb{E}[d_j^\beta \mathbb{1}_{\{\mathcal{G}_\delta\}}] + \mathbb{E}[d_j^\beta \mathbb{1}_{\{\mathcal{G}_\delta^c\}}] \le \mathbb{E}[d_j^\beta \mid \mathcal{G}_\delta] + \mathbb{E}[d_j^\beta \mathbb{1}_{\{\mathcal{G}_\delta^c\}}].$$

Note that

$$d_j \le 2 \max_{1\le\ell\le n} d_{\mathcal{X}}(X_\ell, x^*).$$

Therefore

$$\mathbb{E}\left[d_j^\beta \mathbb{1}_{\{\mathcal{G}_\delta^c\}}\right] \le 2^\beta \mathbb{E}\left[\max_\ell d_{\mathcal{X}}^\beta(X_\ell, x^*)\mathbb{1}_{\{\mathcal{G}_\delta^c\}}\right].$$

Now using the tail bound on the maximum we have

$$\mathbb{P}(\max_\ell d_{\mathcal{X}}^\beta(X_\ell, x^*)) \le nC_1 e^{-C_2 t^\alpha},$$

which with our choice of $T_{n,\delta}$ gives us

$$\mathbb{E}[d_j^\beta \mathbb{1}_{\{\mathcal{G}_\delta^c\}}] \lesssim (\log n)^{\beta/\alpha}\delta.$$

Thus

$$\mathbb{E}[d_j^\beta] \lesssim \mathbb{E}[d_j^\beta \mid \mathcal{G}_\delta] + (\log n)^{\beta/\alpha}\delta. \tag{B.27}$$

Now for fixed j, conditioning on $X_j = x$ and on $\mathcal{G}_\delta$ we have $d_{\mathcal{X}}(x, x^*) \le T_{n,\delta}$. For any $r > 0$, note that by (A2)

$$\mathbb{P}(d_{\mathcal{X}}(X_1, x) \le r) \ge c_0 r^d.$$

Let $N_r \coloneqq \sum_{\ell\neq j} \mathbb{1}_{\{d_{\mathcal{X}}(X_\ell, x)\le r\}}$. Conditional on $X_j = x$, $N_r \sim \text{Bin}(n-1, \mathbb{P}(d_{\mathcal{X}}(X_1, x) \le r))$. The event $\{d_j > r\}$ is exactly $\{N_r < K\}$. Therefore, for any r,

$$\mathbb{P}(d_j > r \mid X_j = x) = \mathbb{P}(N_r < K).$$

Now choose $r_* \coloneqq (2K/c_0(n-1))^{1/d}$, so we have

$$\mu \coloneqq \mathbb{E}[N_{r_*} \mid X_j = x] = (n-1)\mathbb{P}(d_{\mathcal{X}}(X_1, x) \le r_*) \ge 2K.$$

Using a standard multiplicative Chernoff bound for binomials,

$$\mathbb{P}(N_{r_*} < K) \le \exp(-\mu/8) \le \exp(-K/4).$$

More generally, for any $u \geq 1$, we have

$$\mathbb{E}[N_{ur_*} \mid X_j = x] \geq (n-1)c_0 u^d r_*^d = 2Ku^d.$$

Using Chernoff bound again we have

$$\mathbb{P}(d_j > ur_* \mid X_j = x) = \mathbb{P}(N_{ur_*} < K) \leq \exp\left(-Ku^d/4\right).$$

Note that these bounds hold for every x such that $d_{\mathcal{X}}(x, x^*) \leq T_{n,\delta}$, hence they remain valid under $\mathcal{G}_\delta$.

Using this tail bound we have

$$\begin{aligned}
\mathbb{E}[d_j^\beta \mid X_j = x] &= \int_0^{r_*} \beta r^{\beta-1} \mathbb{P}(d_j > r \mid X_j = x) \mathrm{d}r + \int_{r_*}^{\infty} \beta r^{\beta-1} \mathbb{P}(d_j > r \mid X_j = x) \mathrm{d}r \\
&\leq r_*^\beta + \beta r_*^\beta \int_1^\infty u^{\beta-1} e^{-Ku^{d/4}} \mathrm{d}u \\
&\lesssim r_*^\beta = \left(\frac{K}{n}\right)^{\beta/d}
\end{aligned}$$

uniformly for all x such that $d_{\mathcal{X}}(x, x^*) \leq T_{n,\delta}$. Therefore

$$\mathbb{E}[d_j^\beta \mid \mathcal{G}_\delta] \lesssim \left(\frac{K}{n}\right)^{\beta/d}. \tag{B.28}$$

Finally combining (B.27) and (B.28), we have

$$\mathbb{E}[d_j^\beta] \lesssim \left(\frac{K}{n}\right)^{\beta/d} + (2T_{n,\delta})^\beta \delta.$$

Choose $\delta = n^{-2}$ and we get

$$\mathbb{E}[d_j^\beta] \lesssim \left(\frac{K}{n}\right)^{\beta/d} + (\log n)^{\beta/\alpha} n^{-2}.$$

□

Lemma B.3 (Integrability of inverse cube of variance)**.** *Let $\mathcal{Y} = \mathbb{R}$, $Y \sim N(0, \sigma^2)$, and $k_{\mathcal{Y}}(x, y) = \exp\{-\gamma(x-y)^2\}$, where $\gamma > 0$. If $\gamma < \frac{1}{8\sigma^2}$, then $\int_{\mathcal{Y}} \mathbb{V}_Y\left[k_{\mathcal{Y}}(Y, y)\right]^{-3} \mathrm{d}\mathbb{P}_Y(y) < \infty$.*

Proof. For $y \in \mathbb{R}$, we have that

$$\begin{aligned}
\mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)] &= \mathbb{E}_Y\left[\exp\left(-\gamma(Y-y)^2\right)\right]^2 - \left[\mathbb{E}_Y\left[\exp\left(-\gamma(Y-y)^2\right)\right]\right]^2 \\
&\overset{(a)}{=} \frac{e^{-\frac{2y^2\gamma}{1+4\sigma^2\gamma}}}{\sqrt{1+4\sigma^2\gamma}} - \frac{e^{-\frac{2y^2\gamma}{1+2\sigma^2\gamma}}}{1+2\sigma^2\gamma} \overset{(b)}{=:} \frac{e^{-\frac{2y^2\gamma}{c_4}}}{\sqrt{c_4}} - \frac{e^{-\frac{2y^2\gamma}{c_2}}}{c_2} \overset{(c)}{\geq} \frac{e^{-\frac{2y^2\gamma}{c_4}}}{\sqrt{c_4}} - \frac{e^{-\frac{2y^2\gamma}{c_4}}}{c_2} \\
&= e^{-\frac{2y^2\gamma}{c_4}} \left(\frac{1}{\sqrt{c_4}} - \frac{1}{c_2}\right) \overset{(d)}{>} 0,
\end{aligned}$$

where (a) is by the properties of Gaussian integrals, we let $c_4 = 1 + 4\sigma^2\gamma$ and $c_2 = 1 + 2\sigma^2\gamma$ in (b), and use that $c_2 < c_4$ in (c). Inequality (d) follows as

$$\begin{aligned}
c_4^{-1/2} - c_2^{-1} > 0 &\iff c_4^{1/2} < c_2 \iff c_4 < c_2^2 \iff 1 + 4\sigma^2\gamma < (1 + 2\sigma^2\gamma)^2 \\
&\iff 1 + 4\sigma^2\gamma < 1 + 4\sigma^2\gamma + 4\sigma^4\gamma^2 \iff 0 < 4\sigma^4\gamma^2.
\end{aligned}$$

Setting $c \coloneqq c_4^{-1/2} - c_2^{-1}$, and using the obtained lower bound on the variance in the expression of Assumption 5 (i), we have that

$$\begin{aligned}
\int_{\mathcal{Y}} \mathbb{V}_Y[k_{\mathcal{Y}}(Y, y)]^{-3} \mathrm{d}\mathbb{P}_Y(y) &\leq c^{-3} \int_{\mathcal{Y}} e^{\frac{6y^2\gamma}{c_4}} \mathrm{d}\mathbb{P}_Y(y) = c^{-3} \int_{\mathbb{R}} e^{\frac{6y^2\gamma}{c_4}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} \mathrm{d}y \\
&= \frac{1}{c^3\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{y^2\left(\frac{6\gamma}{c_4} - \frac{1}{2\sigma^2}\right)} \mathrm{d}y
\end{aligned}$$

which is finite for $6\gamma/c_4 - 1/(2\sigma^2) < 0$. Hence, solving for γ yields that this is equivalent to

$$\frac{6\gamma}{c_4} < \frac{1}{2\sigma^2} \iff \frac{6\gamma}{1+4\sigma^2\gamma} < \frac{1}{2\sigma^2} \iff 12\gamma\sigma^2 < 1 + 4\sigma^2\gamma \iff 8\sigma^2\gamma < 1 \iff \gamma < \frac{1}{8\sigma^2}. \qquad \square$$

C EXTERNAL RESULTS

This section collects the external results that we use. Theorem C.1 recalls McDiarmid's bounded differences inequality. Lemma C.2 gives a condition for the existence of a Borel measurable function relating two random variables a.s.

Theorem C.1 (Bounded differences inequality; Boucheron et al. 2013)**.** *Let $\mathcal{X}$ be a measurable space. A function $f : \mathcal{X}^n \to \mathbb{R}$ has the bounded difference property for some constants $c_1, \ldots, c_n$ if, for each $i = 1, \ldots, n$,*

$$\sup_{\substack{x_1,\ldots,x_n \\ x_i' \in \mathcal{X}}} |f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots x_n)| \le c_i. \tag{C.29}$$

Then, if $X_1, \ldots, X_n$ is a sequence of independently distributed random variables and (C.29) holds, putting $Z = f(X_1, \ldots, X_n)$ and $\nu = \frac{1}{4}\sum_{i=1}^n c_i^2$, for any $t > 0$, it holds that

$$\mathbb{P}(Z - \mathbb{E}(Z) > t) \le e^{-t^2/(2\nu)}.$$

Lemma C.2 (Remark A.2;[1] Huang et al. 2022)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(\mathcal{X}, \tau_{\mathcal{X}})$ a topological space with Borel σ-algebra $\mathcal{B}(\tau_{\mathcal{X}})$, $(\mathcal{Y}, \tau_{\mathcal{Y}})$ a Polish space with Borel σ-algebra $\mathcal{B}(\tau_{\mathcal{Y}})$, and $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B}(\tau_{\mathcal{X}}))$ and $Y : (\Omega, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B}(\tau_{\mathcal{Y}}))$ random variables. Denote the conditional distribution of Y given X by $\mathbb{P}_{Y|X}$. If $\mathbb{P}_{Y|X=x}$ is degenerate for a.e. $x \in \mathcal{X}$, then there exists a Borel measurable function $f : \mathcal{X} \to \mathcal{Y}$ such that $Y = f(X)$ a.s.*

D ADDITIONAL EXPERIMENT: INTRINSIC-DIMENSION EFFECT IN THEOREM 2

Using the notation below Theorem 2, where d and d_0 denote the intrinsic and ambient dimensions, respectively, we investigate the influence of the intrinsic dimension while keeping the ambient dimension fixed.

To isolate the effect of d, both X and Y are constructed to lie on a d-dimensional subspace of $\mathbb{R}^{d_0}$. Specifically, we generate latent variables

$$Z_X \sim N(\mathbf{0}, \mathbf{I}_d), \qquad Z_Y = Z_X + 0.5\,\varepsilon,$$

where $\varepsilon \sim N(\mathbf{0}, \mathbf{I}_d)$ is independent of Z_X. Let $Q_X, Q_Y \in \mathbb{R}^{d_0 \times d_0}$ be independent Haar-distributed orthogonal matrices, and denote by $Q_X^{(d)}, Q_Y^{(d)} \in \mathbb{R}^{d_0 \times d}$ their first d columns. The observed variables are

$$X = Q_X^{(d)} Z_X, \qquad Y = Q_Y^{(d)} Z_Y.$$

Since $Q_X^{(d)}$ has orthonormal columns, it is an isometric embedding. Hence, for any two observations X_i and X_j,

$$\|X_i - X_j\|_{\mathbb{R}^{d_0}} = \|Q_X^{(d)}(Z_{X_i} - Z_{X_j})\|_{\mathbb{R}^{d_0}} = \|Z_{X_i} - Z_{X_j}\|_{\mathbb{R}^d}.$$

Consequently, the K-NN graph constructed from the observed data in $\mathbb{R}^{d_0}$ is identical to that of the latent variables in $\mathbb{R}^d$. The remaining $d_0 - d$ coordinates therefore carry no geometric information, allowing us to isolate the effect of the intrinsic dimension d on the estimator.

The population quantity $D(Y, X)$ is approximated by averaging estimates obtained from 30 independent simulations, each with a sample size of 10 000, yielding a highly accurate approximation of the population value.

Figure 3 shows that, with the ambient dimension fixed at $d_0 = 20$, the estimator becomes more accurate as the intrinsic dimension decreases, in agreement with the dependence of the convergence rate on the intrinsic dimension established in Theorem 2.

[1] This remark appears in the arXiv version.

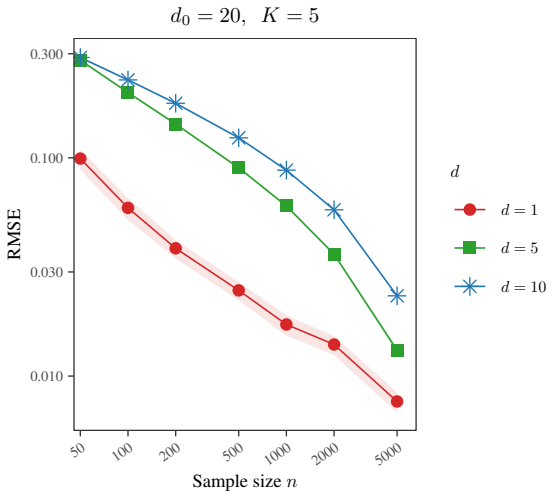

Figure 3: Comparison of the estimation error for fixed ambient dimension $d_0 = 20$ and varying intrinsic dimension $d \in \{1, 5, 10\}$.