DISTANCE AND KERNEL-BASED NONPARAMETRIC TESTS FOR INDEPENDENCE AND HOMOGENEITY OF DISTRIBUTIONS, AND THEIR APPLICATIONS

A Dissertation

by

SHUBHADEEP CHAKRABORTY

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Xianyang Zhang
Committee Members,	Jianhua Huang
	Mohsen Pourahmadi
	Ruihong Huang
Head of Department,	Daren B.H. Cline

August 2020

Major Subject: Statistics

ABSTRACT

Measuring and testing for independence and homogeneity of distributions are some fundamental problems in statistics, finding applications in a wide variety of areas like independent component analysis, gene selection, graphical modeling, causal inference, goodness-of-fit testing, change-point detection and so on.

Székely et al. (2007), in their seminal paper, introduced the notion of distance covariance (dCov) as a measure of dependence between two random vectors of arbitrary (but fixed) dimensions. The innovative feature of dCov is the fact that dCov between two random vectors takes the value zero if and only if they are independent, thereby completely characterizing independence between two random vectors.

However, many statistical applications, such as independent component analysis, diagnostic checking for structural equation modeling, etc., require the quantification of joint independence among $d \geq 2$ random vectors, which is a quite different and more ambitious task than testing for pairwise independence of a collection of random vectors. The first work (Chapter 2) proposes a new dependence metric called the Joint Distance Covariance (JdCov) which generalizes or extends the notion of distance covariance to quantify joint dependence among $d \geq 2$ random vectors of arbitrary (but fixed) dimensions. JdCov takes the value zero if and only if the d random vectors are jointly independent, and thereby completely characterizes their joint independence. We propose empirical estimators of JdCov, study their asymptotic behaviors and consequently propose a consistent bootstrap-based nonparametric test for joint independence. The proposed dependence metrics are employed to perform model selection in causal inference, based on the joint independence testing of the residuals from the fitted structural equation models. The effectiveness of the method is illustrated via both simulated and real datasets.

The second work (Chapter 3) proposes nonparametric tests for homogeneity and independence between two high-dimensional random vectors. Energy distance (proposed by Székely and Rizzo (2004)) is a classical measure of equality of two multivariate distributions, taking the value

zero if and only if the two random vectors are identically distributed. Our work shows that energy distance based on the usual Euclidean distance cannot completely characterize the homogeneity of two high-dimensional distributions in the sense that it can only detect the equality of means and the traces of covariance matrices of two high-dimensional random vectors. In other words, the classical energy distance fails to detect inhomogeneity between two high-dimensional distributions beyond the first two moments. Also it has been pointed out very recently by Zhu et al. (2019) that the classical distance covariance can only capture component-wise linear dependence between two high-dimensional random vectors. Such limitations of the classical energy distance and distance covariance arise due to the use of Euclidean distance, and we propose a new class of distance metrics for high-dimensional Euclidean spaces to overcome the drawbacks.

We propose a new class of homogeneity/dependence metrics based on the new distance metrics, which inherit the desirable properties of the classical energy distance/distance covariance in the low-dimensional setting. And more importantly, in the high-dimensional setup the new metrics are capable of completely characterizing the homogeneity/independence between the low-dimensional marginal distributions, going above and beyond the scope of the classical energy distance/distance covariance. Moreover we propose t-tests based on the new metrics to perform high-dimensional two-sample testing/independence testing in a fully nonparametric framework and study their asymptotic properties. We use our methodology to analyze cross-sector independence of (high-dimensional) stock prices data.

Change-point detection has been a classical problem in statistics, finding applications in a wide variety of fields. A nonparametric change-point detection procedure is concerned with detecting abrupt distributional changes in the data generating distribution, rather than only changes in mean. In the third work (Chapter 4), we consider the problem of detecting an unknown number of change-points in an independent sequence of high-dimensional observations and testing for the significance of the estimated change-point locations. Our approach essentially rests upon nonparametric tests for the homogeneity of two high-dimensional distributions. We construct a single change-point location estimator via defining a cumulative sum process in an embedded Hilbert space. As the

key theoretical innovation, we rigorously derive its limiting distribution under the high dimension medium sample size (HDMSS) framework. Subsequently we combine our statistic with the idea of wild binary segmentation to recursively estimate and test for multiple change-point locations. The superior performance of our methodology compared to several other existing procedures is illustrated via both simulated and real datasets.

DEDICATION

To my mother, Mrs. Sudipta Chakraborty; my father, Mr. Girindranath Chakraborty;

&

all the wonderful friends who taught me not all relations are blood related;

some are much more than that.

ACKNOWLEDGMENTS

I believe a person entering the PhD life and the person completing it, are two very different souls. Truly it has been a journey of it's own kind, and I take this as an opportunity to thank everyone who stood beside me all along through thick and thin, giving me a lot of deep lessons of life, changing my perspectives, and instilling in me a lot of hope, encouragement and confidence, when I needed those dearly.

Before anything, I want to express my heartfelt gratitude to Dr. Xianyang Zhang for being an outstanding advisor; a brilliant mind who taught me Mathematics; someone who taught me what makes it to a top journal, and to aim at high-impact research and publications, always. Someone who has always been there whenever I needed him. After I submitted my second (joint) work, I still vividly remember, he came all the way to my office and told me with a smiling face that "you have become an expert in this field ! Start thinking about problems that you find exciting, think of ways to address them with your acquired expertise !" That moment will stand out in my memory, ever. Deep inside, I knew very well how far far apart I am from being an expert, but it indeed came as a huge moral boost; it slowly made me start believing in myself. Afterwards I slowly started training myself to work independently with as minimal guidance as possible, which I believe he always appreciated, and gave me all the space I needed.

I am extremely thankful to Dr. Jianhua Huang for the innumerable things I learnt from him and the wonderful encouragement and support I have received from him all along. He gave me the opportunity to teach linear algebra, metric spaces and regression analysis in bootcamps arranged for the incoming PhD students in two successive summers, which turned out to be an incredible experience for me. The experience was very different from and way more enjoyable than teaching undergraduate courses, which made me realize my zeal for teaching, adding a lot of motivations to pursue a career in the academia.

I want to thank Dr. Mohsen Pourahmadi and Dr. Ruihong Huang for their continued support. I have been fortunate to have attended courses taught by Dr. Anirban Bhattacharya, Dr. Irina Gaynanova and Dr. Natarajan Sivakumar, who are some of the best teachers and presenters I have ever seen. Attending their classes and talks gave me fantastic lessons on how to give an organized, lively and captivating talk. I am indebted to Dr. Natarajan Sivakumar for teaching me metric topology in the most beautiful way I could have ever imagined, which helped me all along in my research. I am also sincerely thankful to Dr. Bani Mallick, Dr. Samiran Sinha, Dr. Debdeep Pati and Dr. Suhasini Subbarao for the wonderful pieces of advices I got from them at different points of time, which went a long way. Last but not the least, I can never forget what an amazingly nice, approachable and supportive person and Associate Head Dr. Michael Longnecker has been, all along. It is indeed a pleasure to thank them all for making this journey so memorable and sweet.

As I look down the memory lane, I feel immensely grateful to Dr. Kiranmoy Das at the Indian Statistical Institute, without whose kind guidance, support and encouragement, I wouldn't have probably made it here. I also feel indebted to Mr. Dipankar Sarkar, Mr. Biman Bhadra and Mr. Sankar Prasad Ghosh whose magical teaching abilities didn't leave me with any other choice than to fall in love with mathematics and statistics during my school days. Without their caring encouragement, probably I wouldn't have dared to think of a career in Statistics.

Beyond academics, I feel hugely privileged to have got some wonderful friends who never made me feel dull or lonely, while staying thousands of miles away from my family. I truly regret for not being able to mention all their names, just fearing the list would become a little too long. The friday late night parties, the weekend road trips, cooking experiments and the frequent group dinners, group-watching of cricket matches : all gave me the much needed oxygen whenever I felt stressed out. And a special thanks to Trinetri, who has always been a marvellous friend through thick and thin, ever since the undergraduate days.

Finally, lots of love for my sweet mom and dad, who sacrificed a lot of pleasures of their lives just to see their son scaling newer and newer heights in life; and my family and relatives, who take immense pride in whatever little I succeed to achieve.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by Dr. Xianyang Zhang, the thesis advisor. All work for the thesis was completed independently by the student.

Funding Sources

Graduate study was supported by the Department of Statistics at Texas A&M University.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Literature Review

1.1.1 Nonparametric tests for independence

Measuring and testing dependence is of central importance in statistics, which has found applications in a wide variety of areas including independent component analysis, gene selection, graphical modeling and causal inference. Statistical tests of independence can be associated with widely many dependence measures. Two of the most classical measures of association between two ordinal random variables are Spearman's rho and Kendall's tau. However, tests for (pairwise) independence using these two classical measures of association are not consistent, and only have power for alternatives with monotonic association. Contingency table-based methods, and in particular the power-divergence family of test statistics (Read and Cressie, 1988), are the best known general purpose tests of independence, but are limited to relatively low dimensions, since they require a partitioning of the space in which each random variable resides. Another classical measure of dependence between two random vectors is the mutual information (Cover and Thomas, 1991), which can be interpreted as the Kullback-Leibler divergence between the joint density and the product of the marginal densities. The idea originally dates back to the 1950's, in groundbreaking works by Shannon and Weaver (1949), Mcgill (1954) and Fano (1961). Mutual information completely characterizes independence and generalizes to more than two random vectors. However, test based on mutual information involves distributional assumptions for the random vectors and hence is not robust to model misspecification.

In the past fifteen years, kernel-based methods have received considerable attention in both the statistics and machine learning literature. For instance, Bach and Jordan (2002) derived a regularized correlation operator from the covariance and cross-covariance operators and used its largest singular value to conduct independence test. Gretton et al. (2005; 2007) introduced a kernel-based independence measure, namely the Hilbert-Schmidt Independence Criterion (HSIC), to test for independence of two random vectors. This idea was recently extended by Sejdinovic et al. (2013) and Pfister et al. (2018) to quantify the joint independence among more than two random vectors.

Along with a different direction, Székely et al. (2007), in their seminal paper, introduced the notion of distance covariance (dCov) and distance correlation as a measure of dependence between two random vectors of arbitrary dimensions. Given the theoretical appeal of the population quantity and the striking simplicity of the sample version, the idea has been widely extended and analyzed in various ways in Székely and Rizzo (2012; 2014), Lyons (2013), Sejdinovic et al. (2013), Dueck et al. (2014), Bergsma et al. (2014), Wang et al. (2015), and Huo and Székely (2016), to mention only a few.

1.1.2 Nonparametric tests for homogeneity of distributions

Nonparametric two-sample testing of homogeneity of distributions has been a classical problem in statistics, finding a plethora of applications in goodness-of-fit testing, clustering, change-point detection and so on. Some of the most traditional tools in this domain are Kolmogorov-Smirnov test, and Wald-Wolfowitz runs test, whose multivariate and multidimensional extensions have been studied by Darling (1957), David (1958) and Bickel (1969) among others. Friedman and Rafsky (1979) proposed a distribution-free multivariate generalization of the Wald-Wolfowitz runs test applicable for arbitrary but fixed dimensions. Schilling (1986) proposed another distribution-free test for multivariate two-sample problem based on k-nearest neighbor (k-NN) graphs. Maa et al. (1996) suggested a technique for reducing the dimensionality by examining the distribution of interpoint distances. In a recent novel work, Chen and Friedman (2017) proposed graph-based tests for moderate to high dimensional data and non-Euclidean data. The last two decades have seen an abundance of literature on distance and kernel-based tests for equality of distributions. Energy distance (first introduced by Székely (2002)) and maximum mean discrepancy or MMD (see Gretton et al. (2012)) have been widely studied in both the statistics and machine learning communities. Sejdinovic et al. (2013) provided a unifying framework establishing the equivalence between the (generalized) energy distance and MMD.

1.1.3 Nonparametric change-point detection

Change-point detection has been a classical and well-established problem in statistics, aiming to detect lack of homogeneity in a sequence of time-ordered observations. This finds abundance of applications in a wide variety of fields, for example, bioinformatics (see Picard et al. (2005), Curtis et al. (2012)), neuroscience (see Park et al. (2015)), digital speech processing (see Rabiner and Schäfer (2007)), social network analysis (see McCulloh (2009)), and so on. A nonparametric change-point detection procedure is concerned with detecting and localizing quite general types of changes in the data generating distribution, rather than only changes in mean. This challenging problem of detecting abrupt distributional changes in the nonparametric setting has been addressed in the literature over the last couple of decades. But many of the methodologies developed suffer from several limitations, for example, applicability only for real-valued data or in the low-dimensional setting, assumption that the number of true change-points is known, etc. Harchaoui and Cappé (2007) proposed a kernel-based procedure assuming a known number of change-points, which reduces its practical interest. Zou et al. (2014) proposed a nonparametric maximum likelihood approach for detecting multiple (unknown number of) change-points using BIC, but is only applicable for real-valued data. Lung-Yut-Fong et al. (2012) developed a nonparametric approach based on marginal rank statistics, which requires the number of observations to be larger than the dimension of the data. Arlot et al. (2012) proposed a kernel-based multiple change-point detection algorithm for multivariate (but fixed dimensional) or complex (non-Euclidean) data. Some graph-based tests have been proposed recently by Chen and Zhang (2015) and Chu and Chen (2019) for high-dimensional data, which allow us to detect only one or two change-points. Matteson and James (2014) proposed a procedure for estimating multiple change-point locations, namely E-Divisive, built upon an energy distance based test that applies to multivariate observations of arbitrary (but fixed) dimensions. Biau et al. (2016) rigorously derived the asymptotic distribution of the statistic proposed by Matteson and James (2014), thereby adding theoretical justifications to their methodology.

1.2 An overview : distance and kernel-based metrics

In this section, we provide a vivid overview of some classical distance and kernel-based dependence and homogeneity metrics, which serves as the background of Chapters 2, 3 and 4. Let us clarify some notations first.

Notation. Let $X = (X_1, \ldots X_p) \in \mathbb{R}^p$ and $Y = (Y_1, \ldots, Y_q) \in \mathbb{R}^q$ be two random vectors of dimensions p and q respectively. Denote by $\|\cdot\|_p$ the Euclidean norm of $\mathbb{R}^p$ (we shall use it interchangeably with $\|\cdot\|$ when there is no confusion). Let 0_p be the origin of $\mathbb{R}^p$. We use $X \perp\!\!\!\perp Y$ to denote that X is independent of Y, and use "$X \overset{d}{=} Y$" to indicate that X and Y are identically distributed. Let (X', Y') and (X'', Y'') be independent copies of (X, Y). For a metric space $(\mathcal{X}, d_{\mathcal{X}})$, let $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}_1(\mathcal{X})$ denote the set of all finite signed Borel measures on $\mathcal{X}$ and all probability measures on $\mathcal{X}$, respectively. Define $\mathcal{M}^1_{d_{\mathcal{X}}}(\mathcal{X}) := \{v \in \mathcal{M}(\mathcal{X}) : \exists\, x_0 \in \mathcal{X} \text{ s.t. } \int_{\mathcal{X}} d_{\mathcal{X}}(x, x_0)\, d|v|(x) < \infty\}$. For $\theta > 0$, define $\mathcal{M}^{\theta}_{\mathcal{K}}(\mathcal{X}) := \{v \in \mathcal{M}(\mathcal{X}) : \int_{\mathcal{X}} \mathcal{K}^{\theta}(x, x)\, d|v|(x) < \infty\}$, where $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a bivariate kernel function. Define $\mathcal{M}^1_{d_{\mathcal{Y}}}(\mathcal{Y})$ and $\mathcal{M}^{\theta}_{\mathcal{K}}(\mathcal{Y})$ in a similar way. For a matrix $A = (a_{kl})_{k,l=1}^n \in \mathbb{R}^{n\times n}$, define its $\mathcal{U}$-centered version $\tilde{A} = (\tilde{a}_{kl}) \in \mathbb{R}^{n\times n}$ as follows

$$\tilde{a}_{kl} = \begin{cases} a_{kl} - \dfrac{1}{n-2}\displaystyle\sum_{j=1}^n a_{kj} - \dfrac{1}{n-2}\displaystyle\sum_{i=1}^n a_{il} + \dfrac{1}{(n-1)(n-2)}\displaystyle\sum_{i,j=1}^n a_{ij}, & k \neq l, \\ 0, & k = l, \end{cases} \tag{1.1}$$

for $k, l = 1, \ldots, n$. Define

$$(\tilde{A} \cdot \tilde{B}) := \frac{1}{n(n-3)} \sum_{k\neq l} \tilde{a}_{kl}\tilde{b}_{kl}$$

for $\tilde{A} = (\tilde{a}_{kl})$ and $\tilde{B} = (\tilde{b}_{kl}) \in \mathbb{R}^{n\times n}$.

1.2.1 Energy distance and MMD

Energy distance (see Székely et al. (2004, 2005), Baringhaus and Franz (2004)) or the Euclidean energy distance between two random vectors $X, Y \in \mathbb{R}^p$ and $X \perp\!\!\!\perp Y$ with $\mathbb{E}\|X\|_p < \infty$

and $\mathbb{E}\|Y\|_p < \infty$, is defined as

$$ED(X,Y) \;=\; 2\,\mathbb{E}\|X-Y\|_p - \mathbb{E}\|X-X'\|_p - \mathbb{E}\|Y-Y'\|_p\,, \tag{1.2}$$

where (X', Y') is an independent copy of (X, Y). Theorem 1 in Székely et al. (2005) shows that $ED(X, Y) \geq 0$ and the equality holds if and only if $X \overset{d}{=} Y$. In general, for an arbitrary metric space $(\mathcal{X}, d)$, the generalized energy distance between $X \sim P_X$ and $Y \sim P_Y$ where $P_X, P_Y \in \mathcal{M}_1(\mathcal{X}) \cap \mathcal{M}_d^1(\mathcal{X})$ is defined as

$$ED_d(X,Y) \;=\; 2\,\mathbb{E}\,d(X,Y) - \mathbb{E}\,d(X,X') - \mathbb{E}\,d(Y,Y')\,. \tag{1.3}$$

Definition **1.2.1** (Spaces of negative type). *A metric space $(\mathcal{X}, d)$ is said to have negative type if for all $n \geq 1$, $x_1, \ldots, x_n \in \mathcal{X}$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ with $\sum_{i=1}^n \alpha_i = 0$, we have*

$$\sum_{i,j=1}^{n} \alpha_i\,\alpha_j\,d(x_i, x_j) \leq 0\,. \tag{1.4}$$

The metric space $(\mathcal{X}, d)$ is said to be of strong negative type if the equality in (4.5) holds only when $\alpha_i = 0$ for all $i \in \{1, \ldots, n\}$.

If $(\mathcal{X}, d)$ has strong negative type, then $ED_d(X, Y)$ completely characterizes the homogeneity of the distributions of X and Y (see Lyons (2013) and Sejdinovic et al. (2013) for detailed discussions). This quantification of homogeneity of distributions lends itself for reasonable use in one-sample goodness-of-fit testing and two sample testing for equality of distributions.

On the machine learning side, Gretton et al. (2012) proposed a kernel-based metric, namely maximum mean discrepancy (MMD), to conduct two-sample testing for equality of distributions. We provide some background before introducing MMD.

Definition **1.2.2**. *(RKHS) Let $\mathcal{H}$ be a Hilbert space of real valued functions defined on some space $\mathcal{X}$. A bivariate function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if :*

1. $\forall x \in \mathcal{X}, \mathcal{K}(\cdot, x) \in \mathcal{H}$

2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H},\ \langle f, \mathcal{K}(\cdot, x)\rangle_{\mathcal{H}} = f(x)$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product associated with $\mathcal{H}$. If $\mathcal{H}$ has a reproducing kernel, it is said to be a reproducing kernel Hilbert space (RKHS).

By Moore-Aronszajn theorem, for every positive definite function (also called a kernel) $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there is an associated RKHS $\mathcal{H}_{\mathcal{K}}$ with the reproducing kernel $\mathcal{K}$. The map $\Pi : \mathcal{M}_1(\mathcal{X}) \to \mathcal{H}_{\mathcal{K}}$, defined as $\Pi(P) = \int_{\mathcal{X}} \mathcal{K}(\cdot, x)\, dP(x)$ for $P \in \mathcal{M}_1(\mathcal{X})$ is called the mean embedding function associated with $\mathcal{K}$. A kernel $\mathcal{K}$ is said to be characteristic to $\mathcal{M}_1(\mathcal{X})$ if the map Π associated with $\mathcal{K}$ is injective. Suppose $\mathcal{K}$ is a characteristic kernel on $\mathcal{X}$. Then the MMD between $X \sim P_X$ and $Y \sim P_Y$, where $P_X, P_Y \in \mathcal{M}_1(\mathcal{X}) \cap \mathcal{M}_{\mathcal{K}}^{1/2}(\mathcal{X})$ is defined as

$$MMD_{\mathcal{K}}(X, Y) \;=\; \| \, \Pi(P_X) \,-\, \Pi(P_Y) \, \|_{\mathcal{H}_{\mathcal{K}}} \, . \tag{1.5}$$

By virtue of $\mathcal{K}$ being a characteristic kernel, $MMD_{\mathcal{K}}(X, Y) = 0$ if and only if $X \overset{d}{=} Y$. Lemma 6 in Gretton et al. (2012) shows that the squared MMD can be equivalently expressed as

$$MMD_{\mathcal{K}}^2(X, Y) \;=\; \mathbb{E}\,\mathcal{K}(X, X') \,+\, \mathbb{E}\,\mathcal{K}(Y, Y') \,-\, 2\,\mathbb{E}\,\mathcal{K}(X, Y) \, . \tag{1.6}$$

Theorem 22 in Sejdinovic et al. (2013) establishes the equivalence between (generalized) energy distance and MMD. Following is the definition of a kernel induced by a distance metric (refer to Section 4.1 in Sejdinovic et al. (2013) for more details).

DEFINITION **1.2.3**. *(Distance-induced kernel and kernel-induced distance) Let $(\mathcal{X}, d)$ be a metric space of negative type and $x_0 \in \mathcal{X}$. Denote $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as*

$$\mathcal{K}(x, x') \;=\; \frac{1}{2}\, \{ d(x, x_0) + d(x', x_0) - d(x, x') \} \, . \tag{1.7}$$

The kernel $\mathcal{K}$ is positive definite if and only if $(\mathcal{X}, d)$ has negative type, and thus $\mathcal{K}$ is a valid kernel on $\mathcal{X}$ whenever d is a metric of negative type. The kernel $\mathcal{K}$ defined in (1.7) is said to be the *distance-induced kernel induced by d and centered at x_0. One the other hand, the distance d can be generated by the kernel $\mathcal{K}$ through*

$$d(x, x') = \mathcal{K}(x, x) + \mathcal{K}(x', x') - 2\mathcal{K}(x, x'). \tag{1.8}$$

Proposition 29 in Sejdinovic et al. (2013) establishes that the distance-induced kernel $\mathcal{K}$ induced by d is characteristic to $\mathcal{M}_1(\mathcal{X}) \cap \mathcal{M}_{\mathcal{K}}^1(\mathcal{X})$ if and only if $(\mathcal{X}, d)$ has strong negative type. Therefore, MMD can be viewed as a special case of the generalized energy distance in (4.4) with d being the metric induced by a characteristic kernel.

Suppose $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$ are i.i.d samples of X and Y respectively. A U-statistic type estimator of $E_d(X, Y)$ is defined as

$$E_{n,m}(X, Y) = \frac{2}{nm} \sum_{k=1}^{n} \sum_{l=1}^{m} d(X_k, Y_l) - \frac{1}{n(n-1)} \sum_{k \neq l}^{n} d(X_k, X_l) - \frac{1}{m(m-1)} \sum_{k \neq l}^{m} d(Y_k, Y_l)\,. \tag{1.9}$$

1.2.2 Distance covariance and HSIC

Distance covariance (dCov) was first introduced in the seminal paper by Székely et al. (2007) to quantify the dependence between two random vectors of arbitrary (fixed) dimensions. Consider two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with $\mathbb{E}\|X\|_p < \infty$ and $\mathbb{E}\|Y\|_q < \infty$. The Euclidean dCov between X and Y is defined as the positive square root of

$$dCov^2(X, Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2}{\|t\|_p^{1+p} \, \|s\|_q^{1+q}} dtds,$$

where f_X, f_Y and $f_{X,Y}$ are the individual and joint characteristic functions of X and Y respectively, and, $c_p = \pi^{(1+p)/2} / \, \Gamma((1+p)/2)$ is a constant with $\Gamma(\cdot)$ being the complete gamma function.

The key feature of dCov is that it completely characterizes independence between two random vectors of arbitrary dimensions, or in other words $dCov(X, Y) = 0$ if and only if $X \perp\!\!\!\perp Y$.

According to Remark 3 in Székely et al. (2007), dCov can be equivalently expressed as

$$\begin{aligned} dCov^2(X,Y) \;&=\; \mathbb{E}\,\|X-X'\|_p\|Y-Y'\|_q \;+\; \mathbb{E}\,\|X-X'\|_p\,\mathbb{E}\,\|Y-Y'\|_q \\ &\quad -\; 2\,\mathbb{E}\,\|X-X'\|_p\|Y-Y''\|_q. \end{aligned} \tag{1.10}$$

Lyons (2013) extends the notion of dCov from Euclidean spaces to general metric spaces. For arbitrary metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, the generalized dCov between $X \sim P_X \in \mathcal{M}_1(\mathcal{X}) \cap \mathcal{M}^1_{d_{\mathcal{X}}}(\mathcal{X})$ and $Y \sim P_Y \in \mathcal{M}_1(\mathcal{Y}) \cap \mathcal{M}^1_{d_{\mathcal{Y}}}(\mathcal{Y})$ is defined as

$$\begin{aligned} D^2_{d_{\mathcal{X}},d_{\mathcal{Y}}}(X,Y) \;&=\; \mathbb{E}\,d_{\mathcal{X}}(X,X')d_{\mathcal{Y}}(Y,Y') \;+\; \mathbb{E}\,d_{\mathcal{X}}(X,X')\,\mathbb{E}\,d_{\mathcal{Y}}(Y,Y') \\ &\quad -\; 2\,\mathbb{E}\,d_{\mathcal{X}}(X,X')d_{\mathcal{Y}}(Y,Y''). \end{aligned} \tag{1.11}$$

Theorem 3.11 in Lyons (2013) shows that if $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are both metric spaces of strong negative type, then $D_{d_{\mathcal{X}},d_{\mathcal{Y}}}(X,Y) = 0$ if and only if $X \perp\!\!\!\perp Y$. In other words, the complete characterization of independence by dCov holds true for any metric spaces of strong negative type. According to Theorem 3.16 in Lyons (2013), every separable Hilbert space is of strong negative type. As Euclidean spaces are separable Hilbert spaces, the characterization of independence by dCov between two random vectors in $(\mathbb{R}^p, \|\cdot\|_p)$ and $(\mathbb{R}^q, \|\cdot\|_q)$ is just a special case.

Hilbert-Schmidt Independence Criterion (HSIC) was introduced as a kernel-based independence measure by Gretton et al. (2005, 2007). Suppose $\mathcal{X}$ and $\mathcal{Y}$ are arbitrary topological spaces, $\mathcal{K}_{\mathcal{X}}$ and $\mathcal{K}_{\mathcal{Y}}$ are characteristic kernels on $\mathcal{X}$ and $\mathcal{Y}$ with the respective RKHSs $\mathcal{H}_{\mathcal{K}_{\mathcal{X}}}$ and $\mathcal{H}_{\mathcal{K}_{\mathcal{Y}}}$. Let $\mathcal{K} = \mathcal{K}_{\mathcal{X}} \otimes \mathcal{K}_{\mathcal{Y}}$ be the tensor product of the kernels $\mathcal{K}_{\mathcal{X}}$ and $\mathcal{K}_{\mathcal{Y}}$, and, $\mathcal{H}_{\mathcal{K}}$ be the tensor product of the RKHSs $\mathcal{H}_{\mathcal{K}_{\mathcal{X}}}$ and $\mathcal{H}_{\mathcal{K}_{\mathcal{Y}}}$. The HSIC between $X \sim P_X \in \mathcal{M}_1(\mathcal{X}) \cap \mathcal{M}^{1/2}_{\mathcal{K}}(\mathcal{X})$ and $Y \sim P_Y \in \mathcal{M}_1(\mathcal{Y}) \cap \mathcal{M}^{1/2}_{\mathcal{K}}(\mathcal{Y})$ is defined as

$$HSIC_{\mathcal{K}_{\mathcal{X}},\mathcal{K}_{\mathcal{Y}}}(X,Y) \;=\; \|\,\Pi(P_{XY}) \;-\; \Pi(P_X P_Y)\,\|_{\mathcal{H}_{\mathcal{K}}}, \tag{1.12}$$

where P_{XY} denotes the joint probability distribution of X and Y. The HSIC between X and Y is essentially the MMD between the joint distribution P_{XY} and the product of the marginals P_X and

P_Y. Clearly, $HSIC_{\mathcal{K}_\mathcal{X},\mathcal{K}_\mathcal{Y}}(X,Y)=0$ if and only if $X \perp\!\!\!\perp Y$. Gretton et al. (2005) shows that the squared HSIC can be equivalently expressed as

$$\begin{aligned} HSIC^2_{\mathcal{K}_\mathcal{X},\mathcal{K}_\mathcal{Y}}(X,Y) \;&=\; \mathbb{E}\,\mathcal{K}_\mathcal{X}(X,X')\mathcal{K}_\mathcal{Y}(Y,Y') \,+\, \mathbb{E}\,\mathcal{K}_\mathcal{X}(X,X')\,\mathbb{E}\,\mathcal{K}_\mathcal{Y}(Y,Y') \\ &\quad -\, 2\,\mathbb{E}\,\mathcal{K}_\mathcal{X}(X,X')\mathcal{K}_\mathcal{Y}(Y,Y''). \end{aligned} \tag{1.13}$$

Theorem 24 in Sejdinovic et al. (2013) establishes the equivalence between the generalized dCov and HSIC.

For an observed random sample $(X_i,Y_i)_{i=1}^n$ from the joint distribution of X and Y, a U-statistic type estimator of the generalized dCov in (1.11) can be defined as

$$\widetilde{D^2_{n\,;\,d_\mathcal{X},d_\mathcal{Y}}}(X,Y) \;=\; (\tilde{A}\cdot\tilde{B}) \;=\; \frac{1}{n(n-3)}\sum_{k\neq l}\tilde{a}_{kl}\tilde{b}_{kl}\;, \tag{1.14}$$

where $\tilde{A}$, $\tilde{B}$ are the $\mathcal{U}$-centered versions (see (4.1)) of $A=\big(d_\mathcal{X}(X_k,X_l)\big)_{k,l=1}^n$ and $B=\big(d_\mathcal{Y}(Y_k,Y_l)\big)_{k,l=1}^n$, respectively. We denote $\widetilde{D^2_{n\,;\,d_\mathcal{X},d_\mathcal{Y}}}(X,Y)$ by $dCov^2_n(X,Y)$ when $d_\mathcal{X}$ and $d_\mathcal{Y}$ are Euclidean distances.

2. DISTANCE-BASED NONPARAMETRIC TESTS FOR JOINT INDEPENDENCE[*]

2.1 Background and notations

Many statistical applications require the quantification of joint dependence among $d \geq 2$ random variables (or vectors). Examples include model diagnostic checking for directed acyclic graph (DAG) where inferring pairwise independence is not enough in this case (see more details in Section 2.6), and independent component analysis which is a means for finding a suitable representation of multivariate data such that the components of the transformed data are mutually independent. In this work, we shall introduce new metrics which generalize the notion of dCov to quantify joint dependence of $d \geq 2$ random vectors. We first introduce the notion of high order dCov to measure the so-called Lancaster interaction dependence (Lancaster, 1969). We generalize the notion of Brownian covariance (Székely et al., 2009) and show that it coincides with the high order distance covariance. We then define the joint dCov (Jdcov) as a linear combination of pairwise dCov and their high order counterparts. The proposed metric provides a natural decomposition of joint dependence into the sum of lower order and high order effects, where the relative importance of the lower order effect terms and the high order effect terms is determined by a user-chosen number. In the population case, Jdcov is equal to zero if and only if the d random vectors are mutually independent, and thus completely characterizes joint independence. It is also worth mentioning that the proposed metrics are invariant to permutation of the variables and they inherit some nice properties of dCov, see Section 2.2.2.

Following the idea of Streitberg (1990), we introduce the concept of distance cumulant and distance characteristic function, which leads us to an equivalent characterization of independence of the d random vectors. Furthermore, we establish a scale invariant version of Jdcov and discuss the concept of rank-based distance measures, which can be viewed as the counterparts of Spearman's rho to dCov and JdCov.

[*] Reprinted with permission from "Distance Metrics for Measuring Joint Dependence with Application to Causal Inference" by Shubhadeep Chakraborty and Xianyang Zhang, 2019. *Journal of the American Statistical Association*, 114(528), 1638-1650, Copyright[2019] by Taylor & Francis Group.

JdCov and its scale-invariant versions can be conveniently estimated in finite sample using V-statistics or their bias-corrected versions. We study the asymptotic properties of the estimators, and introduce a bootstrap procedure to approximate their sampling distributions. The asymptotic validity of the bootstrap procedure is justified under both the null and alternative hypotheses. The new metrics are employed to perform model selection in a causal inference problem, which is based on the joint independence testing of the residuals from the fitted structural equation models. We compare our tests with the bootstrap version of the d-variate HSIC (dHSIC) test recently introduced in Pfister et al. (2018) and the mutual independence test proposed by Matteson and Tsay (2017). Finally we remark that although we focus on Euclidean space valued random variables, our results can be readily extended to general metric spaces in view of the results in Lyons (2013).

The rest of the work is organized as follows. Section 2.2.1 introduces the high order distance covariance and studies its basic properties. Section 2.2.2 describes the JdCov to quantity joint dependence. Sections 2.2.3-2.2.4 further introduce some related concepts including the distance cumulant, distance characteristic function, and rank-based distance covariance. We study the estimation of the distance metrics in Section 2.3 and present a joint independence test based on the proposed metrics in Section 2.4. Section 4.4 is devoted to numerical studies. The new metrics are employed to perform model selection in causal inference in Section 2.6. Section 2.7 discusses the efficient computation of distance metrics and future research directions. The technical details are gathered in the appendix.

Notations. Consider $d \geq 2$ random vectors $\mathcal{X} = \{X_1, \ldots, X_d\}$, where $X_i \in \mathbb{R}^{p_i}$. Set $p_0 = \sum_{i=1}^{d} p_i$. Let $\{X_1', \ldots, X_d'\}$ be an independent copy of $\mathcal{X}$. Denote by $\imath = \sqrt{-1}$ the imaginary unit. Let $|\cdot|_p$ be the Euclidean norm of $\mathbb{R}^p$ with the subscript omitted later without ambiguity. For $a, b \in \mathbb{R}^p$, let $\langle a, b \rangle = a^\top b$. For a complex number a, denote by $\bar{a}$ its conjugate. Let f_i be the characteristic function of X_i, i.e., $f_i(t) = \mathbb{E}[e^{\imath \langle t, X_i \rangle}]$ with $t \in \mathbb{R}^{p_i}$. Define $w_p(t) = (c_p |t|_p^{1+p})^{-1}$ with $c_p = \pi^{(1+p)/2} / \Gamma((1+p)/2)$. Write $dw = (c_{p_1} c_{p_2} \ldots c_{p_d} |t_1|_{p_1}^{1+p_1} \cdots |t_d|_{p_d}^{1+p_d})^{-1} dt_1 \cdots dt_d$. Let I_k^d be the collection of k-tuples of indices from $\{1, 2, \ldots, d\}$ such that each index occurs exactly once. Denote by $\lfloor a \rfloor$ the integer part of $a \in \mathbb{R}$. Write $X \perp\!\!\!\perp Y$ if X is independent of Y.

2.2 Measuring joint dependence

2.2.1 High order distance covariance

We briefly review the concept of Lancaster interactions first introduced by Lancaster (1969). The Lancaster interaction measure associated with a multidimensional probability distribution of d random variables $\{X_1, \dots, X_d\}$ with the joint distribution $F = F_{1,2,\dots,d}$, is a signed measure ΔF given by

$$\Delta F \;=\; (F_1^* - F_1)(F_2^* - F_2) \cdots (F_d^* - F_d)\,, \tag{2.1}$$

where after expansion, a product of the form $F_i^* F_j^* \cdots F_k^*$ denotes the corresponding joint distribution function $F_{i,j,\dots,k}$ of $\{X_i, X_j, \dots, X_k\}$. For example for $d = 4$, the term $F_1^* F_2^* F_3 F_4$ stands for $F_{12} F_3 F_4$, $F_1^* F_2 F_3 F_4$ stands for $F_1 F_2 F_3 F_4$, etc. In particular for $d = 3$, (2.1) simplifies to

$$\Delta F \;=\; F_{123} - F_1 F_{23} - F_2 F_{13} - F_3 F_{12} + 2 F_1 F_2 F_3\,. \tag{2.2}$$

In light of the Lancaster interaction measure, we introduce the concept of dth order dCov as follows.

DEFINITION **2.2.1**. *The dth order dCov is defined as the positive square root of*

$$dCov^2(X_1, \dots, X_d) = \int_{\mathbb{R}^{p_0}} \left| \mathbb{E}\left[\prod_{i=1}^{d} (f_i(t_i) - e^{\imath \langle t_i, X_i \rangle}) \right] \right|^2 dw, \tag{2.3}$$

When $d = 2$, it reduces to the dCov in Székely et al. (2007).

The term $\mathbb{E}[\prod_{i=1}^d (f_i(t_i) - e^{\imath\langle t_i, X_i\rangle})]$ in the definition of dCov is a counterpart of the Lancaster interaction measure in (2.1) with the joint distribution functions replaced by the joint characteristic functions. When $d = 3$, $dCov^2(X_1, X_2, X_3) > 0$ rules out the possibility of any factorization of the joint distribution. To see this, we note that $X_1 \perp\!\!\!\perp (X_2, X_3)$, $X_2 \perp\!\!\!\perp (X_1, X_3)$ or $X_3 \perp\!\!\!\perp (X_1, X_2)$

all lead to $dCov^2(X_1, X_2, X_3) = 0$. On the other hand, $dCov^2(X_1, X_2, X_3) = 0$ implies that

$$\begin{aligned} & f_{123}(t_1, t_2, t_3) - f_1(t_1)f_2(t_2)f_3(t_3) \\ =& f_1(t_1)f_{23}(t_2, t_3) + f_2(t_2)f_{13}(t_1, t_3) + f_3(t_3)f_{12}(t_1, t_2) - 3f_1(t_1)f_2(t_2)f_3(t_3) \end{aligned}$$

for $t_i \in \mathbb{R}^{p_i}$ almost everywhere. In this case, the "higher order effect" i.e., $f_{123}(t_1, t_2, t_3) - f_1(t_1)f_2(t_2)f_3(t_3)$ can be represented by the "lower order/pairwise effects" $f_{ij}(t_i, t_j) - f_i(t_i)f_j(t_j)$ for $1 \leq i \neq j \leq 3$. However, this does not necessarily imply that X_1, X_2 and X_3 are jointly independent. In other words when $d = 3$ (or more generally when $d \geq 3$), joint independence of X_1, X_2 and X_3 is not a necessary condition for dCov to be zero. To address this issue, we shall introduce a new distance metric to quantify any forms of dependence among $\mathcal{X}$ in Section 2.2.2.

In the following, we present some basic properties of high order dCov. Define the bivariate function $U_i(x, x') = \mathbb{E}|x - X_i'| + \mathbb{E}|X_i - x'| - |x - x'| - \mathbb{E}|X_i - X_i'|$ for $x, x' \in \mathbb{R}^{p_i}$ with $1 \leq i \leq d$. Our definition of dCov is partly motivated by the following lemma.

LEMMA **2.2.1**. *For* $1 \leq i \leq d$,

$$U_i(x, x') = \int_{\mathbb{R}^{p_i}} \left\{ (f_i(t) - e^{\imath\langle t, x\rangle})(f_i(-t) - e^{-\imath\langle t, x'\rangle}) \right\} w_{p_i}(t) dt.$$

By Lemma 2.2.1 and Fubini's theorem, the dth order (squared) dCov admits the following equivalent representation,

$$\begin{aligned} dCov^2(X_1, \dots, X_d) =& \int_{\mathbb{R}^{p_0}} \left| \mathbb{E}\left[\prod_{i=1}^{d} (f_i(t_i) - e^{\imath\langle t_i, X_i\rangle}) \right] \right|^2 dw \\ =& \int_{\mathbb{R}^{p_0}} \mathbb{E}\left[\prod_{i=1}^{d} (f_i(t_i) - e^{\imath\langle t_i, X_i\rangle}) \right] \mathbb{E}\left[\prod_{i=1}^{d} \overline{(f_i(t_i) - e^{\imath\langle t_i, X_i'\rangle})} \right] dw \qquad (2.4) \\ =& \mathbb{E}\left[\prod_{i=1}^{d} U_i(X_i, X_i') \right]. \end{aligned}$$

This suggests that similar to dCov, its high order counterpart has an expression based on the moments of U_is, which results in very simple and applicable empirical formulas, see more details in Section 2.3.

REMARK **2.2.1**. From the definition of dCov in Székely et al. (2007), it might appear that its most natural generalization to the case of $d = 3$ would be to define a measure in the following way

$$\frac{1}{c_p c_q c_r} \int_{\mathbb{R}^{p+q+r}} \frac{|f_{X,Y,Z}(t,s,u) - f_X(t) f_Y(s) f_Z(u)|^2}{|t|_p^{1+p} |s|_q^{1+q} |u|_r^{1+r}} \, dt ds du \,,$$

where $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ and $Z \in \mathbb{R}^r$. Assuming that the integral above exists, one can easily verify that such a measure completely characterizes joint independence among X, Y and Z. However, it does not admit a nice equivalent representation as in (2.4) (unless one considers a different weighting function). We exploit this equivalent representation of the dth order dCov to propose a V-statistic type estimator of the population quantity (see Section 3) which is much simpler to compute rather than evaluating an integral as in the original definition in (2.3).

REMARK **2.2.2**. Székely et al. (2009) introduced the notion of covariance with respect to a stochastic process. Theorem 8 in Székely et al. (2009) shows that population distance covariance coincides with the covariance with respect to Brownian motion (or the so-called *Brownian covariance*). The Brownian covariance of two random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with $\mathbb{E}(|X|^2 + |Y|^2) < \infty$ is defined as the positive square root of

$$\mathcal{W}^2(X,Y) = Cov_W^2(X,Y) = \mathbb{E}[X_W X_W^{'} Y_{W'} Y_{W'}^{'}] \,,$$

where W and $W^{'}$ are independent Brownian motions with zero mean and covariance function $C(t,s) = |s| + |t| - |s-t|$ on $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively, and

$$X_W = W(X) - \mathbb{E}[\, W(X) | W \,] \,.$$

Conditional on W (or $W^{'}$), $X_W^{'}$ (or $Y_{W'}^{'}$) is an i.i.d. copy of X_W (or $Y_{W'}$). Then following Theorem 8 in Székely et al. (2009) and Definition 2.1, we have $dCov^2(X,Y) = \mathcal{W}^2(X,Y)$.

Now for $d \geq 2$ random variables $\{X_1, X_2, \ldots, X_d\}$ where $X_i \in \mathbb{R}^{p_i}, 1 \leq i \leq d$, we can generalize the notion of Brownian covariance as the positive square root of

$$\mathcal{W}^2(X_1, \ldots, X_d) = \mathbb{E}\left[\prod_{i=1}^{d} X_{i_{W_i}} X'_{i_{W_i}}\right],$$

where W_i's are independent Brownian motions on $\mathbb{R}^{p_i}$, $1 \leq i \leq d$. Property (2) in Proposition 2.1 below establishes the connection between the higher order distance covariances and the generalized notion of Brownian covariance.

Similar to $dCov$, our definition of high order $dCov$ possesses the following important properties.

Proposition **2.2.1**. *We have the following properties regarding* $dCov(X_1, X_2, \ldots, X_d)$*:*

(1) For any $a_i \in \mathbb{R}^{p_i}$*,* $c_i \in \mathbb{R}$*, and orthogonal transformations* $A_i \in \mathbb{R}^{p_i \times p_i}$*,* $dCov^2(a_1 + c_1A_1X_1, \ldots, a_d + c_dA_dX_d) = \prod_{i=1}^{d} |c_i|\, dCov^2(X_1, \ldots, X_d)$*. Moreover, dCov is invariant to any permutation of* $\{X_1, X_2,$
$\ldots, X_d\}$*.*

(2) Under Assumption 4.3.2 (see Section 2.3), the dth order $dCov$ *exists and*

$$\mathcal{W}^2(X_1, \ldots, X_d) = dCov^2(X_1, \ldots, X_d)\,.$$

Property (1) shows that dCov is invariant to translation, orthogonal transformation and permutation on X_is. In property (2), the existence of the dth order $dCov$ follows from (2.4) and application of Fubini's Theorem and Hölder's inequality. The equality with Brownian covariance readily follows from the proof of Theorem 7 in Székely et al. (2009).

Theorem 7 in Székely et al. (2007) shows the relationship between distance correlation and the correlation coefficient for bivariate normal distributions. We extend that result in case of multivariate normal random variables with zero mean, unit variance and pairwise correlation ρ. Proposition

2.2.2 below establishes a relationship between the correlation coefficient and higher order distance covariances for multivariate normal random variables.

PROPOSITION **2.2.2**. *Suppose* $(X_1, X_2, \ldots, X_d) \sim N(0, \Sigma)$*, where* $\Sigma = (\sigma_{i,j})_{i,j=1}^d$ *with* $\sigma_{ii} = 1$ *for* $1 \leq i \leq d$ *and* $\sigma_{ij} = \rho$ *for* $1 \leq i \neq j \leq d$*. When* $d = 2k-1$ *or* $d = 2k$*,* $dCov^2(X_1, \ldots, X_d) = O(|\rho|^{2k})$ *for* $k \geq 2$*.*

Proposition A.0.1 in the appendix shows some additional properties of the dth order $dCov$.

2.2.2 Joint distance covariance

In this subsection, we introduce a new joint dependence measure called the joint dCov (Jdcov), which is designed to capture all types of interaction dependence among the d random vectors. To achieve this goal, we define JdCov as the linear combination of all kth order dCov for $1 \leq k \leq d$.

DEFINITION **2.2.2**. *The JdCov among* $\{X_1, \ldots, X_d\}$ *is given by*

$$\begin{aligned} & JdCov^2(X_1, \ldots, X_d; C_2, \ldots, C_d) \\ =& C_2 \sum_{(i_1,i_2)\in I_2^d} dCov^2(X_{i_1}, X_{i_2}) + C_3 \sum_{(i_1,i_2,i_3)\in I_3^d} dCov^2(X_{i_1}, X_{i_2}, X_{i_3}) \\ & + \cdots + C_d \, dCov^2(X_1, \ldots, X_d), \end{aligned} \tag{2.5}$$

for some nonnegative constants $C_i \geq 0$ *with* $2 \leq i \leq d$*.*

Proposition 2.2.3 below states that JdCov completely characterizes joint independence among $\{X_1, \ldots, X_d\}$.

PROPOSITION **2.2.3**. *Suppose* $C_i > 0$ *for* $2 \leq i \leq d$*. Then* $JdCov^2(X_1, \ldots, X_d; C_2, \ldots, C_d) = 0$ *if and only if* $\{X_1, \ldots, X_d\}$ *are mutually independent.*

Next we show that by properly choosing C_is, $JdCov^2(X_1, \ldots, X_d; C_2, \ldots, C_d)$ has a relatively simple expression, which does not require the evaluation of $2^d - d - 1$ dCov terms in its original definition (2.5). Specifically, let $C_i = c^{d-i}$ for $c \geq 0$ in the definition of JdCov and denote

$JdCov^2(X_1,\dots,X_d;c) = JdCov^2(X_1,\dots,X_d;c^{d-2},c^{d-1},\dots,1)$. Then, we have the following result.

PROPOSITION **2.2.4**. *For any* $c \geq 0$*,*

$$JdCov^2(X_1,\dots,X_d;c) = \mathbb{E}\left[\prod_{i=1}^{d}(U_i(X_i,X_i') + c)\right] - c^d.$$

In particular, $JdCov^2(X_1,X_2;c) = E[U_1(X_1,X_1')U_2(X_2,X_2')] = dCov^2(X_1,X_2)$.

By (2.5), the dependence measured by JdCov can be decomposed into the main effect term $\sum_{(i_1,i_2)\in I_2^d} dCov^2(X_{i_1},X_{i_2})$ quantifying the pairwise dependence as well as the higher order effect terms $\sum_{(i_1,i_2,\dots,i_k)\in I_k^d} dCov^2(X_{i_1},X_{i_2},\dots,X_{i_k})$ quantifying the multi-way interaction dependence among any k-tuples. The choice of c reflects the relative importance of the main effect and the higher order effects. For $c \geq 1$, $C_i = c^{d-i}$ is nonincreasing in i. Thus, the larger c we select, the smaller weights we put on the higher order terms. In particular, we have

$$\lim_{c\to+\infty} c^{2-d} JdCov^2(X_1,\dots,X_d;c) = \sum_{(i_1,i_2)\in I_2^d} dCov^2(X_{i_1},X_{i_2}),$$

that is JdCov reduces to the main effect term as $c \to +\infty$. We remark that the main effect term fully characterizes joint dependence in the case of elliptical distribution and it has been recently used in Yao *et al.* (2018) to test mutual independence for high-dimensional data. On the other hand, JdCov becomes the dth order dCov as $c \to 0$, i.e.,

$$\lim_{c\to 0} JdCov^2(X_1,\dots,X_d;c) = dCov^2(X_1,\dots,X_d).$$

The choice of c depends on the types of interaction dependence of interest as well as the specific scientific problem, and thus is left for the user to decide.

It is worth noting that $JdCov^2(X_1,\dots,X_d;c)$ depends on the scale of X_i. To obtain a scale-invariant metric, one can normalize U_i by the corresponding distance variance. Specifically, when

$dCov(X_i) := dCov(X_i, X_i) > 0$, the resulting quantity is given by,

$$JdCov_S^2(X_1, \ldots, X_d; c) = \mathbb{E}\left[\prod_{i=1}^{d}\left(\frac{U_i(X_i, X_i')}{dCov(X_i)} + c\right)\right] - c^d,$$

which is scale-invariant. Another way to obtain a scale-invariant metric is presented in Section 2.2.4 based on the idea of rank transformation.

Below we present some basic properties of JdCov, which follow directly from Proposition 2.2.1.

Proposition **2.2.5**. *We have the following properties regarding JdCov:*

(1) For any $a_i \in \mathbb{R}^{p_i}$, $c_0 \in \mathbb{R}$, and orthogonal transformations $A_i \in \mathbb{R}^{p_i \times p_i}$, $JdCov^2(a_1 + c_0A_1X_1, \ldots, a_d + c_0A_dX_d; |c_0|c) = |c_0|^d JdCov^2(X_1, \ldots, X_d; c)$. Moreover, JdCov is invariant to any permutation of $\{X_1, X_2, \ldots, X_d\}$.

(2) For any $a_i \in \mathbb{R}^{p_i}$, $c_i \neq 0$, and orthogonal transformations $A_i \in \mathbb{R}^{p_i \times p_i}$, $JdCov_S^2(a_1 + c_1A_1X_1, \ldots, a_d + c_dA_dX_d; c) = JdCov_S^2(X_1, \ldots, X_d; c)$.

Remark **2.2.3**. A natural question to ask is what should be a data driven way to choose the tuning parameter c. Although we leave it for future research, here we present a heuristic idea of choosing c. In the discussion below Proposition 2.2.4, we pointed out that choosing $c > 1$ (or < 1) puts lesser (or higher) weightage on the higher order effects. Note that if the data is Gaussian, testing for the mutual independence of $\{X_1, \ldots, X_d\}$ is equivalent to testing for their pairwise independences. In that case, intuitively one should choose a larger (> 1) value of c. If, however, the data is non-Gaussian, it might be of interest to look into higher order dependencies and thus a smaller (< 1) choice of c makes sense.

To summarize, a heuristic way to choose the tuning parameter c could be :

$$\text{Choose c} \begin{cases} > 1, & \text{if } \{X_1, \ldots, X_d\} \text{ are jointly Gaussian} \\ < 1, & \text{if } \{X_1, \ldots, X_d\} \text{ are not jointly Gaussian.} \end{cases} \tag{2.6}$$

There is a huge literature on testing for joint normality of random vectors. It has been shown that the test based on energy distance is consistent against fixed alternatives (Székely and Rizzo, 2004) and shows higher empirical power compared to several competing tests (Székely and Rizzo, 2005; 2013). Suppose p is the p-value of the energy distance based test for joint normality of $\{X_1, \ldots, X_d\}$ at level α. We expect c to increase (or decrease) from 1 as $p >$ (or $<$) α, so one heuristic choice of c can be

$$c = 1 + \text{sign}(p - \alpha) \times |p - \alpha|^{1/4} \ , \tag{2.7}$$

where $\text{sign}(x) = 1, 0 \ or - 1$ depending on whether $x > 0, x = 0 \ or \ x < 0$. For example, $p = (0.001, 0.03, 0.0499, 0.0501, 0.1, 0.3)$ and $\alpha = 0.05$ yields $c = (0.53, 0.62, 0.9, 1.1, 1.47, 1.71)$.

2.2.3 Distance cumulant and distance characteristic function

As noted in Streitberg (1990), for $d \geq 4$, the Lancaster interaction measure fails to capture all possible factorizations of the joint distribution. For example, it may not vanish if $(X_1, X_2) \perp\!\!\!\perp (X_3, X_4)$. Streitberg (1990) corrected the definition of Lancaster interaction measure using a more complicated construction, which essentially corresponds to the cumulant version of dCov in our context. Specifically, Streitberg (1990) proposed a corrected version of Lancaster interaction as follows

$$\widetilde{\Delta} F \ = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! \prod_{D \in \pi} F_D,$$

where π is a partition of the set {1,2,...,d}, $|\pi|$ denotes the number of blocks of the partition π and F_D denotes the joint distribution of $\{X_i : i \in D\}$. It has been shown in Streitberg (1990) that $\widetilde{\Delta} F = 0$ whenever F is decomposable. Our definition of joint distance cumulant of $\{X_1, \ldots, X_d\}$ below can be viewed as the dCov version of Streitberg's correction.

DEFINITION **2.2.3**. *The joint distance cumulant among* $\{X_1, \ldots, X_d\}$ *is defined as*

$$cum(X_1, \ldots, X_d) = \sum_{\pi} (-1)^{|\pi|-1}(|\pi| - 1)! \prod_{D\in\pi} \mathbb{E}\left(\prod_{i\in D} U_i(X_i, X_i')\right), \tag{2.8}$$

where π *runs through all partitions of* $\{1, 2, \ldots, d\}$.

It is not hard to verify that $\mathrm{cum}(X_1, \ldots, X_d) = 0$ if $\{X_1, \ldots, X_d\}$ can be decomposed into two mutually independent groups say $(X_i)_{i\in\pi_1}$ and $(X_j)_{j\in\pi_2}$ with π_1 and π_2 being a partition of $\{1, 2, \ldots, d\}$. We further define the distance characteristic function.

DEFINITION **2.2.4**. *The joint distance characteristic function among* $\{X_1, \ldots, X_d\}$ *is defined as*

$$dcf(t_1, \ldots, t_d) = \mathbb{E}\left[\exp\left(\imath \sum_{i=1}^{d} t_i U_i(X_i, X_i')\right)\right], \tag{2.9}$$

for $t_1, \ldots, t_d \in \mathbb{R}$.

The following result shows that distance cumulant can be interpreted as the coefficient of the Taylor expansion of the log distance characteristic function.

PROPOSITION **2.2.6**. *The joint distance cumulant* $cum(X_{i_1}, \ldots, X_{i_s})$ *is given by the coefficient of* $\imath^s \prod_{k=1}^{s} t_{i_k}$ *in the Taylor expansion of* $\log\{dcf(t_1, \ldots, t_d)\}$, *where* $\{i_1, \ldots, i_s\}$ *is any subset of* $\{1, 2, \ldots, d\}$ *with* $s \le d$.

Our next result indicates that the mutual independence among $\{X_1, \ldots, X_d\}$ is equivalent to the mutual independence among $\{U_1(X_1, X_1'), \ldots, U_d(X_d, X_d')\}$.

PROPOSITION **2.2.7**. *The random variables* $\{X_1, \ldots, X_d\}$ *are mutually independent if and only if* $dcf(t_1, \ldots, t_d) = \prod_{i=1}^{d} dcf(t_i)$ *for* t_i *almost everywhere, where* $dcf(t_i) = \mathbb{E}[\exp\{\imath t_i U_i(X_i, X_i')\}]$.

2.2.4 Rank-based metrics

In this subsection, we briefly discuss the concept of rank-based distance measures. For simplicity, we assume that X_is are all univariate and remark that our definition can be generalized to the case where X_is are random vectors without essential difficulty. The basic idea here is to apply the monotonic transformation based on the marginal distribution functions to each X_j, and then use the dCov or JdCov to quantify the interaction and joint dependence of the coordinates after transformation. Therefore it can be viewed as the counterpart of Spearman's rho to dCov or JdCov.

Let F_j be the marginal distribution function for X_j. The squared rank dCov and JdCov among $\{X_1, \ldots, X_d\}$ are defined respectively as

$$dCov_R^2(X_1, \ldots, X_d) = dCov^2(F_1(X_1), \ldots, F_d(X_d)),$$
$$JdCov_R^2(X_1, \ldots, X_d; c) = JdCov^2(F_1(X_1), \ldots, F_d(X_d); c).$$

The rank-based dependence metrics enjoy a few appealing features: (1) they are invariant to monotonic component wise transformations; (2) they are more robust to outliers and heavy tail of the distribution; (3) their existence require very weak moment assumption on the components of $\mathcal{X}$. In Section 4.4, we shall compare the finite sample performance of JdCov$_R^2$ with that of JdCov and JdCov$_S$.

Table 2.1: Comparison of various distance metrics for measuring joint dependence of $d \geq 2$ random vectors of arbitrary dimensions :

Distance metrics	Complete characterization of joint independence	Permutation invariance	Scale invariance
dHSIC	✓	✓	× (for fixed bandwidth)
T_{MT}	✓	×	×
High order $dCov$	× (Captures Lancaster interactions)	✓	×
$JdCov$	✓	✓	×
$JdCov_S$	✓	✓	✓
$JdCov_R$	✓	✓	✓

2.3 Estimation

We now turn to the estimation of the joint dependence metrics. Given n samples $\{\mathbf{X}_j\}_{j=1}^n$ with $\mathbf{X}_j = (X_{j1}, \ldots, X_{jd})$, we consider the plug-in estimators based on the V-statistics as well as their bias-corrected versions to be described below. Denote by $\hat{f}_i(t_i) = n^{-1} \sum_{j=1}^n e^{\imath\langle t_i, X_{ji}\rangle}$ the empirical characteristic function for X_i.

2.3.1 Plug-in estimators

For $1 \leq k, l \leq n$, let $\widehat{U}_i(k,l) = n^{-1}\sum_{v=1}^n |X_{ki} - X_{vi}| + n^{-1}\sum_{u=1}^n |X_{ui} - X_{li}| - |X_{ki} - X_{li}| - n^{-2}\sum_{u,v=1}^n |X_{ui} - X_{vi}|$ be the sample estimate of $U_i(X_{ki}, X_{li})$. The V-statistic type estimators for dCov, JdCov and its scale-invariant version are defined respectively as,

$$\widehat{dCov^2}(X_1, \ldots, X_d) = \frac{1}{n^2}\sum_{k,l=1}^n \prod_{i=1}^d \widehat{U}_i(k,l)^2, \tag{2.10}$$

$$\widehat{JdCov^2}(X_1, \ldots, X_d; c)) = \frac{1}{n^2}\sum_{k,l=1}^n \prod_{i=1}^d \left(\widehat{U}_i(k,l) + c\right) - c^d, \tag{2.11}$$

$$\widehat{JdCov^2_S}(X_1, \ldots, X_d; c) = \frac{1}{n^2}\sum_{k,l=1}^n \prod_{i=1}^d \left(\frac{\widehat{U}_i(k,l)}{\widehat{dCov}(X_i)} + c\right) - c^d, \tag{2.12}$$

where $\widehat{dCov^2}(X_i) = n^{-2}\sum_{k,l=1}^n \widehat{U}_i(k,l)^2$ is the sample (squared) dCov. The following lemma shows that the V-statistic type estimators are equivalent to the plug-in estimators by replacing the characteristic functions and the expectation in the definitions of dCov and JdCov with their sample counterparts.

LEMMA **2.3.1**. *The sample (squared) dCov can be rewritten as,*

$$\widehat{dCov^2}(X_1, \ldots, X_d) = \int_{\mathbb{R}^{p_0}} \left| \frac{1}{n}\sum_{k=1}^n \left[\prod_{i=1}^d (\hat{f}_i(t_i) - e^{\imath\langle t_i, X_{ki}\rangle})\right]\right|^2 dw. \tag{2.13}$$

Moreover, we have

$$\begin{aligned}
&\widehat{JdCov}^2(X_1,\dots,X_d;c)\\
=&c^{d-2}\sum_{(i_1,i_2)\in I_2^d}\widehat{dCov^2}(X_{i_1},X_{i_2})+c^{d-3}\sum_{(i_1,i_2,i_3)\in I_3^d}\widehat{dCov^2}(X_{i_1},X_{i_2},X_{i_3})\\
&+\cdots+\widehat{dCov^2}(X_1,\dots,X_d).
\end{aligned} \tag{2.14}$$

REMARK **2.3.1**. Consider the univariate case where $p_i = 1$ for all $1 \le i \le d$. Let $\widehat{F}_i$ be the empirical distribution based on $\{X_{ji}\}_{j=1}^n$ and define $Z_{ji} = \widehat{F}_i(X_{ji})$. Then, the rank-based metrics defined in Section 2.2.4 can be estimated in a similar way by replacing X_{ji} with Z_{ji} in the definitions of the above estimators.

REMARK **2.3.2**. The distance cumulant can be estimated by

$$\widehat{\text{cum}}(X_1,\dots,X_d)=\sum_{\pi}(-1)^{|\pi|-1}(|\pi|-1)!\prod_{D\in\pi}\left\{\frac{1}{n^2}\sum_{k,l=1}^n\left(\prod_{i\in D}\widehat{U}_i(k,l)\right)\right\}.$$

However, the combinatorial nature of distance cumulant implies that detecting interactions of higher order requires significantly more costly computation.

We study the asymptotic properties of the V-statistic type estimators under suitable moment assumptions.

ASSUMPTION **2.3.1**. *Suppose for any subset S of $\{1,2,\dots,d\}$ with $|S| \ge 2$, there exists a partition $S = S_1 \cup S_2$ such that $\mathbb{E}\prod_{i\in S_1}|X_i| < \infty$ and $\mathbb{E}\prod_{i\in S_2}|X_i| < \infty$.*

PROPOSITION **2.3.1**. *Under Assumption 4.3.2 , we have as $n \to \infty$,*

$$\begin{aligned}
&\widehat{dCov^2}(X_1,\cdots,X_d)\xrightarrow{a.s}dCov^2(X_1,\cdots,X_d),\\
&\widehat{JdCov}^2(X_1,\cdots,X_d;c)\xrightarrow{a.s}JdCov^2(X_1,\dots,X_d;c),\\
&\widehat{JdCov}^2_S(X_1,\cdots,X_d;c)\xrightarrow{a.s}JdCov^2_S(X_1,\dots,X_d;c),
\end{aligned}$$

where “ $\xrightarrow{a.s}$ ” denotes the almost sure convergence.

When $d = 2$, Assumption 2.3.1 reduces to the condition that $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}|X_2| < \infty$ in Theorem 2 of Székely et al. (2007). Suppose X_is are mutually independent. Then Assumption 2.3.1 is fulfilled provided that $\mathbb{E}|X_i| < \infty$ for all i. More generally, if $E|X_i|^{\lfloor (d+1)/2 \rfloor} < \infty$ for $1 \leq i \leq d$, then Assumption 2.3.1 is satisfied.

Let $\Gamma(\cdot)$ denote a complex-valued zero mean Gaussian random process with the covariance function $R(t, t') = \prod_{i=1}^{d} \big(f_i(t_i - t_i') - f_i(t_i) f_i(-t_i')\big)$, where $t = (t_1, t_2, \ldots, t_d), t' = (t_1', t_2', \ldots, t_d') \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \times \cdots \times \mathbb{R}^{p_d}$.

PROPOSITION **2.3.2**. *Suppose $X_1, X_2, \ldots, X_d$ are mutually independent, and $\mathbb{E}|X_i| < \infty$ for $1 \leq i \leq d$. Then we have*

$$n\widehat{dcov}^2(X_1, X_2, \cdots, X_d) \xrightarrow{d} ||\Gamma||^2 = \sum_{j=1}^{+\infty} \lambda_j Z_j^2,$$

where $||\Gamma||^2 = \int \Gamma(t_1, t_2, \ldots, t_d)^2 dw$, $Z_j \overset{i.i.d}{\sim} N(0,1)$ and $\lambda_j > 0$ depends on the distribution of $\mathcal{X}$. As a consequence, we have

$$n\widehat{Jdcov}^2(X_1, X_2, \cdots, X_d; c) \xrightarrow{d} \sum_{j=1}^{+\infty} \lambda_j' Z_j^2,$$

with $\lambda_j' > 0$ and $Z_j \overset{i.i.d}{\sim} N(0,1)$.

Proposition 2.3.2 shows that both $\widehat{dcov}^2$ and $\widehat{Jdcov}^2$ converge to weighted sum of chi-squared random variables, where the weights depend on the marginal characteristic functions in a complicated way. Since the limiting distribution is non-pivotal, we will introduce a bootstrap procedure to approximate their sampling distributions in the next section.

It has been pointed out in the literature that the computational complexity of dCov is $O(n^2)$ if it is implemented directly according to its definition. The computational cost of the V-statistic type estimators and the bias-corrected estimators for JdCov are both of the order $O(n^2 p_0)$.

2.3.2 Bias-corrected estimators

It is well known that V-statistic leads to biased estimation. To remove the bias, one can construct an estimator for the dth order dCov based on a dth order U-statistic. However, the computational complexity for the dth order U-statistic is of the order $O(dn^d)$, which is computationally prohibitive when n and d are both large. Adopting the $\mathcal{U}$-centering idea in Székely and Rizzo (2014), we propose bias-corrected estimators which do not bring extra computational cost as compared to the plug-in estimators. Specifically, for $1 \leq i \leq d$, we define the $\mathcal{U}$-centered version of $|X_{ki} - X_{li}|$ as

$$\begin{aligned}\widetilde{U}_i(k,l) =& \frac{1}{n-2}\sum_{u=1}^{n}|X_{ui} - X_{li}| + \frac{1}{n-2}\sum_{v=1}^{n}|X_{ki} - X_{vi}| - |X_{ki} - X_{li}| \\ &- \frac{1}{(n-1)(n-2)}\sum_{u,v=1}^{n}|X_{ui} - X_{vi}|\end{aligned}$$

when $k \neq l$, and $\widetilde{U}_i(k,l) = 0$ when $k = l$. One can verify that $\sum_{v\neq k}\widetilde{U}_i(k,v) = \sum_{u\neq l}\widetilde{U}_i(u,l) = 0$, which mimics the double-centered property $\mathbb{E}[U_i(X_i, X_i')|X_i] = \mathbb{E}[U_i(X_i, X_i')|X_i'] = 0$ for its population counterpart. Let $\widetilde{dCov}^2(X_i, X_j) = \sum_{k\neq l}\widetilde{U}_i(k,l)\widetilde{U}_j(k,l)/\{n(n-3)\}$ and write $\widetilde{dCov}(X_i) = \widetilde{dCov}(X_i, X_i)$. We define the bias-corrected estimators as,

$$\begin{aligned}\widetilde{JdCov}^2(X_1,\dots,X_d;c) &= \frac{1}{n(n-3)}\sum_{k,l=1}^{n}\prod_{i=1}^{d}\left(\widetilde{U}_i(k,l)+c\right) - \frac{n}{n-3}c^d, \\ \widetilde{JdCov}^2_S(X_1,\dots,X_d;c) &= \frac{1}{n(n-3)}\sum_{k,l=1}^{n}\prod_{i=1}^{d}\left(\frac{\widetilde{U}_i(k,l)}{\widetilde{dCov}(X_i)}+c\right) - \frac{n}{n-3}c^d.\end{aligned}$$

Direct calculation yields that

$$\widetilde{JdCov}^2(X_1,\dots,X_n;c) = c^{d-2}\sum_{(i,j)\in I_2^d}\widetilde{dCov}^2(X_i, X_j) + \text{higher order terms}. \tag{2.15}$$

It has been shown in Proposition 1 of Székely and Rizzo (2014) that $\widetilde{dCov}^2(X_i, X_j)$ is an unbiased estimator for $dCov^2(X_i, X_j)$. In the supplementary material, we provide an alternative proof which simplifies the arguments in Székely and Rizzo (2014). Our argument relies on a new decomposition of $\widetilde{U}_i(k,l)$, which provides some insights on the $\mathcal{U}$-centering idea. See Lemma A.0.1 and Proposition A.0.2 in the supplementary material. In view of (2.15) and Proposition A.0.2, the main effect in $JdCov^2(X_1,\ldots,X_n;c)$ can be unbiasedly estimated by the main effect of $\widetilde{JdCov}^2(X_1,\ldots,X_n;c)$. However, it seems very challenging to study the impact of $\mathcal{U}$-centering on the bias of the high order effect terms. We shall leave this problem to our future research.

2.4 Testing for joint independence

In this section, we consider the problem of testing the null hypothesis

$$H_0 : X_1,\ldots,X_d \text{ are mutually independent} \tag{2.16}$$

against the alternative H_A : negation of H_0. For the purpose of illustration, we use $n\widehat{JdCov}^2$ as our test statistic and set

$$\phi_n(\mathbf{X}_1,\ldots,\mathbf{X}_n) := \begin{cases} 1 \quad if \quad n\widehat{JdCov}^2(X_1,\ldots,X_d) > c_n\,, \\ 0 \quad if \quad n\widehat{JdCov}^2(X_1,\ldots,X_d) \le c_n\,, \end{cases} \tag{2.17}$$

where the threshold c_n remains to be chosen. Consequently, we define a decision rule as follows: reject H_0 if $\phi_n = 1$ and fail to reject H_0 if $\phi_n = 0$.

Below we introduce a bootstrap procedure to approximate the sampling distribution of $n\widehat{JdCov}$ under H_0. Let $\widehat{F}_i$ be the empirical distribution function based on the data points $\{X_{ji}\}_{j=1}^n$. Conditional on the original sample, we define $\mathbf{X}_j^* = (X_{j1}^*,\ldots,X_{jd}^*)$, where X_{ji}^* are generated independently from $\widehat{F}_i$ for $1 \le i \le d$. Let $\{\mathbf{X}_j^*\}_{j=1}^n$ be n bootstrap samples. Then we can compute the bootstrap statistics $\widehat{dCov^2}^*$ and $\widehat{JdCov^2}^*$ in the same way as $\widehat{dCov}^2$ and $\widehat{JdCov}^2$ based on $\{\mathbf{X}_j^*\}_{j=1}^n$. In particular, we note that the bootstrap version of the dth order dCov is given by

$$n\widehat{dCov^2}^*(X_1,\ldots,X_d) = \|\Gamma_n^*\|^2 = \int \Gamma_n^*(t_1,\ldots,t_d)^2 dw,$$

where

$$\Gamma_n^*(t) \;=\; n^{-1/2}\sum_{j=1}^{n}\prod_{i=1}^{d}(\hat{f}_i^*(t_i) - e^{\imath\langle t_i, X_{ji}^*\rangle}).$$

Denote by " $\xrightarrow{d^*}$ " the weak convergence in the bootstrap world conditional on the original sample $\{\mathbf{X}_j\}_{j=1}^n$.

PROPOSITION **2.4.1**. *Suppose* $\mathbb{E}|X_i| < \infty$ *for* $1 \le i \le d$*. Then*

$$n\widehat{dCov^2}^*(X_1, \dots, X_d) \xrightarrow{d^*} \sum_{j=1}^{+\infty}\lambda_j Z_j^2,$$

$$n\widehat{JdCov^2}^*(X_1, \dots, X_d) \xrightarrow{d^*} \sum_{j=1}^{+\infty}\lambda_j' Z_j^2,$$

almost surely as $n \to \infty$*.*

Proposition 2.4.1 shows that the bootstrap statistic is able to imitate the limiting distribution of the test statistic. Thus, we shall choose c_n to be the $1-\alpha$ quantile of the distribution of $n\widehat{JdCov^2}^*$ conditional on the sample $\{\mathbf{X}_j\}_{j=1}^n$. The validity of the bootstrap-assisted test can be justified as follows.

PROPOSITION **2.4.2**. *For all* $\alpha \in (0,1)$*, the* α*-level bootstrap-assisted test has asymptotic level* α *when testing* H_0 *against* H_A*. In other words, under* H_0*,* $\limsup_{n\to\infty} P\,(\,\phi_n(\mathbf{X}_1, \dots, \mathbf{X}_n) = 1\,) = \alpha$.

PROPOSITION **2.4.3**. *For all* $\alpha \in (0,1)$*, the* α*-level bootstrap-assisted test is consistent when testing* H_0 *against* H_A*. In other words, under* H_A*,* $\lim_{n\to\infty} P\,(\,\phi_n(\mathbf{X}_1, \dots, \mathbf{X}_n) = 1\,) = 1$.

2.5 Numerical studies

We investigate the finite sample performance of the proposed methods. Our first goal is to test the joint independence among the variables $\{X_1, \dots, X_d\}$ using the new dependence metrics, and compare the performance with some existing alternatives in the literature in terms of size and power. Throughout the simulation, we set $c = 0.5, 1, 2$ in JdCov and implement the bootstrap-assisted test based on the bias-corrected estimators. We compare our tests with the dHSIC-based test in Pfister et al. (2018) and the mutual independence test proposed in Matteson and Tsay (2017), which is defined as

$$T_{MT} := \sum_{i=1}^{d-1} dCov^2(X_i, X_{(i+1):d}), \tag{2.18}$$

where $X_{(i+1):d} = \{X_{i+1}, X_{i+2}, \ldots, X_d\}$. We consider both Gaussian and non-Gaussian distributions and study the following models, motivated from Sejdinovic et al. (2013) and Yao et al. (2018).

EXAMPLE **2.5.1**. *[Gaussian copula model]* The data $\mathbf{X} = (X_1, \ldots, X_d)$ are generated as follows:

1. $\mathbf{X} \sim N(0, I_d)$;
2. $\mathbf{X} = Z^{1/3}$ and $Z \sim N(0, I_d)$;
3. $\mathbf{X} = Z^3$ and $Z \sim N(0, I_d)$.

EXAMPLE **2.5.2**. *[Multivariate Gaussian model]* The data $\mathbf{X} = (X_1, \ldots, X_d)$ are generated from the multivariate normal distribution with the following three covariance matrices $\Sigma = (\sigma_{ij}(\rho))_{i,j=1}^d$ with $\rho = 0.25$:

1. AR(1): $\sigma_{ij} = \rho^{|i-j|}$ for all $i, j \in \{1, \ldots, d\}$;
2. Banded: $\sigma_{ii} = 1$ for $i = 1, \ldots, d$; $\sigma_{ij} = \rho$ if $1 \leq |i-j| \leq 2$ and $\sigma_{ij} = 0$ otherwise;
3. Block: Define $\Sigma_{\text{block}} = (\sigma_{ij})_{i,j=1}^5$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho$ if $i \neq j$. Let $\Sigma = I_{\lfloor d/5 \rfloor} \otimes \Sigma_{\text{block}}$, where $\otimes$ denotes the Kronecker product.

EXAMPLE **2.5.3**. The data $\mathbf{X} = (X, Y, Z)$ are generated as follows:

1. $X, Y \overset{i.i.d}{\sim} N(0, 1)$, $Z = \text{sign}(XY)\, W$, where W follows an exponential distribution with mean $\sqrt{2}$;
2. X, Y are independent Bernoulli random variables with the success probability 0.5, and $Z = \mathbf{1}\{X = Y\}$.

EXAMPLE **2.5.4**. In this example, we consider a triplet of random vectors (X, Y, Z) on $\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p$, with $X, Y \overset{i.i.d}{\sim} N(0, I_p)$. We focus on the following cases :

1. $Z_1 = \text{sign}(X_1 Y_1)\, W$ and $Z_{2:p} \sim N(0, I_{p-1})$, where W follows an exponential distribution with mean $\sqrt{2}$;

2. $Z_{2:p} \sim N(0, I_{p-1})$ and

$$Z_1 = \begin{cases} X_1^2 + \epsilon, & \text{with probability } 1/3, \\ Y_1^2 + \epsilon, & \text{with probability } 1/3, \\ X_1 Y_1 + \epsilon, & \text{with probability } 1/3, \end{cases}$$

   where $\epsilon \sim U(-1, 1)$.

We conduct tests for joint independence among the random variables described in the above examples. For each example, we draw 1000 simulated datasets and perform tests of joint independence with 500 bootstrap resamples. We try small and moderate sample sizes, i.e., $n = 50, 100$ or 200. Figure 2.1 and Figure 2.2 display the proportion of rejections (out of 1000 simulation runs) for the five different tests, based on the statistics $\widetilde{JdCov}^2$, $\widetilde{JdCov}_S^2$, $\widetilde{JdCov}_R^2$, dHSIC and T_{MT}. The detailed figures are reported in Tables A.1 and $A.2$ in the appendix.

In Example 2.5.1, the data generating scheme suggests that the variables are jointly independent. The plots in Figure 2.1 show that all the five tests perform more or less equally well in examples 2.5.1.1 and 2.5.1.2, and the rejection probabilities are quite close to the 10% or 5% nominal level. In Example 2.5.1.3, the tests based on our proposed statistics show greater conformation of the empirical size to the actual size of the test than T_{MT}. In Example $2.5.2$, the tests based on $\widetilde{JdCov}^2$, $\widetilde{JdCov}_S^2$ and $\widetilde{JdCov}_R^2$ as well as T_{MT} significantly outperform the dHSIC-based test. Note that the empirical power becomes higher when c increases to 2. From Figure 2.2, we observe that in Example $2.5.3$ all the tests perform very well in the second case. However, in the first case, our tests and the dHSIC-based test deliver higher power as compared to T_{MT}. Finally, in Example

2.5.4, we allow X, Y, Z to be random vectors with dimension $p = 5, 10$. The rejection probabilities for each of the five tests increase with n, and the proposed tests provide better performances in comparison with the other two competitors. In particular, the test based on $\widetilde{JdCov}_S^2$ outperforms all the others in a majority of the cases. In Examples 2.5.3 and 2.5.4, the power becomes higher when c decreases to 0.5. These results are consistent with our statistical intuition and the discussions in Section 2.2.2. For the Gaussian copula model, only the main effect term matters, so a larger c is preferable. For non-Gaussian models, the high order terms kick in and hence a smaller c may lead to higher power.

REMARK **2.5.1**. We have considered U-statistic type estimators of $JdCov^2$, $JdCov_S^2$ and $JdCov_R^2$ so far in all the above computations, as they remove the bias due to the main effects (see Section 2.3.2). However it might be interesting to see if the bias correction has any empirical impact. We conduct tests for joint independence of the random variables in some of the above examples, this time using the V-statistic type estimators (described in Section 2.3.1). Table A.3 (in the appendix) shows the proportion of rejections (out of 1000 simulation runs) for the tests based on $\widehat{JdCov}^2$, $\widehat{JdCov}_S^2$ and $\widehat{JdCov}_R^2$, setting $c = 1$. The results indicate that use of the bias corrected estimators lead to greater conformation of the empirical size to the actual size of the test (in Example 2.5.1), and slightly better power in Example 2.5.3.

REMARK **2.5.2**. In connection to the heuristic idea discussed in Remark 2.2.3 about choosing the tuning parameter c, we conduct tests for joint independence of the random variables in all the above examples, choosing c in that way. Table A.4 (in the appendix) presents the proportion of rejections for the proposed tests and the values of c for each example, averaged over the 1000 simulated datasets. The plots in Figure 2.1 and Figure 2.2 reveal some interesting features. In Example 2.5.2 we have Gaussian data, so a larger c is preferable. Clearly the proportion of rejections are a little higher (or lower) in most of the cases when we choose c in the data-driven way (c turns out to be around 1.6 or 1.7), than when c is subjectively chosen to be 0.5 (or 2). On the contrary, in Example 2.5.3, the data is non-Gaussian and a smaller c is preferable. Evidently choosing c in the data-driven way leads to nearly equally good power compared to when $c = 0.5$, and higher power compared to when $c = 2$.

2.6 Application to causal inference

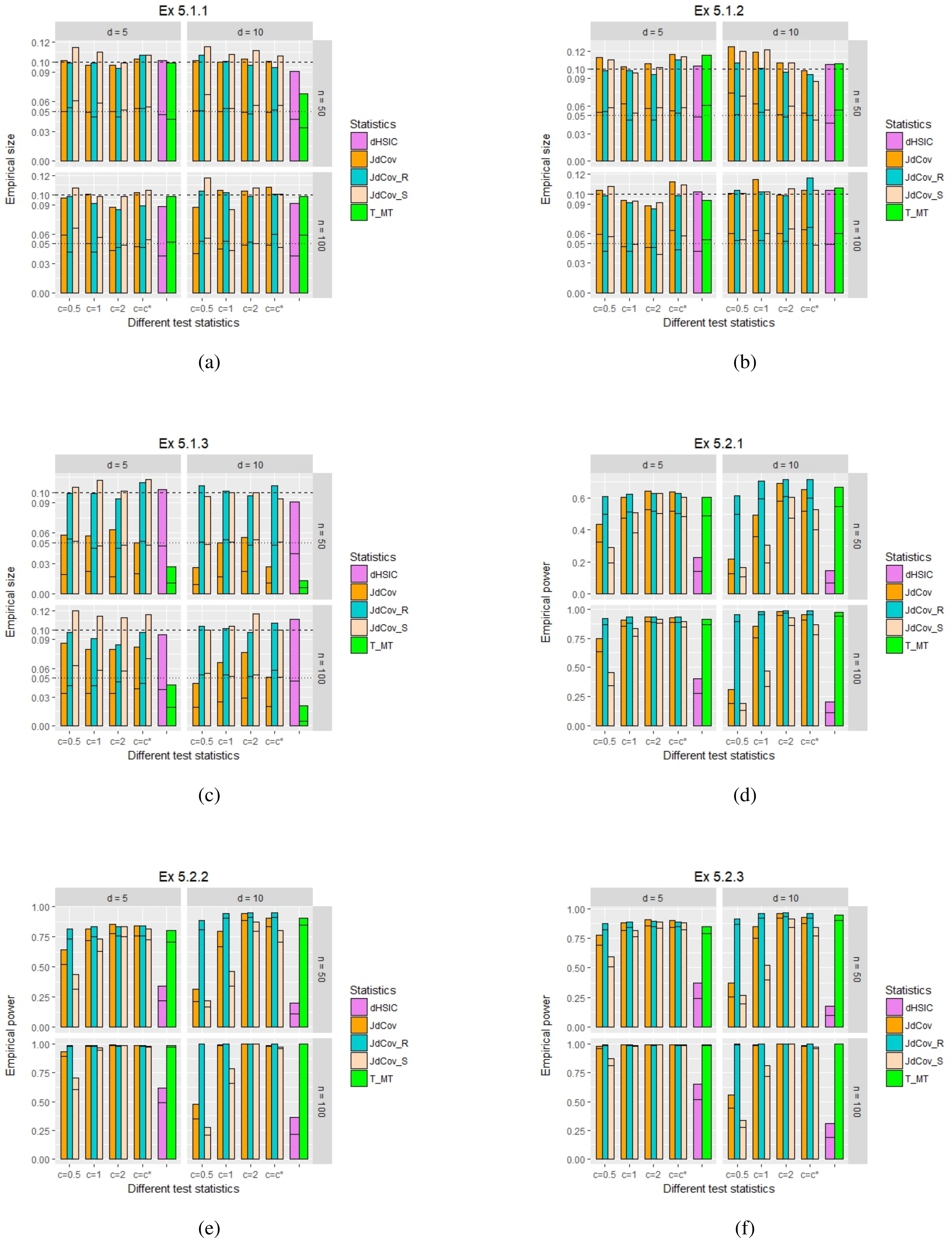
2.6.1 Model diagnostic checking for Directed Acyclic Graph (DAG)

We employ the proposed metrics to perform model selection in causal inference which is based on the joint independence testing of the residuals from the fitted structural equation models. Specifically, given a candidate DAG $\mathcal{G}$, we let Par(j) denote the index associated with the parents of the jth node. Following Peters et al. (2014) and Bühlmann et al. (2014), we consider the structural equation models with additive components

$$X_j = \sum_{k \in \text{Par}(j)} f_{j,k}(X_k) + \epsilon_j \,,\ j = 1, 2, \ldots, d, \tag{2.19}$$

where the noise variables $\epsilon_1, \ldots, \epsilon_d$ are jointly independent variables. Given n observations $\{\mathbf{X}_i\}_{i=1}^{n}$ with $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$, we use generalized additive regression (Wood and Augustin, 2002) to regress X_j on all its parents $\{X_k, k \in \text{Par}(j)\}$ and denote the resulting residuals by

$$\hat{\epsilon}_{ij} = X_{ij} - \sum_{k \in \text{Par}(j)} \hat{f}_{j,k}(X_{ik}), \quad 1 \leq j \leq d, \quad 1 \leq i \leq n,$$



(a)

(b)

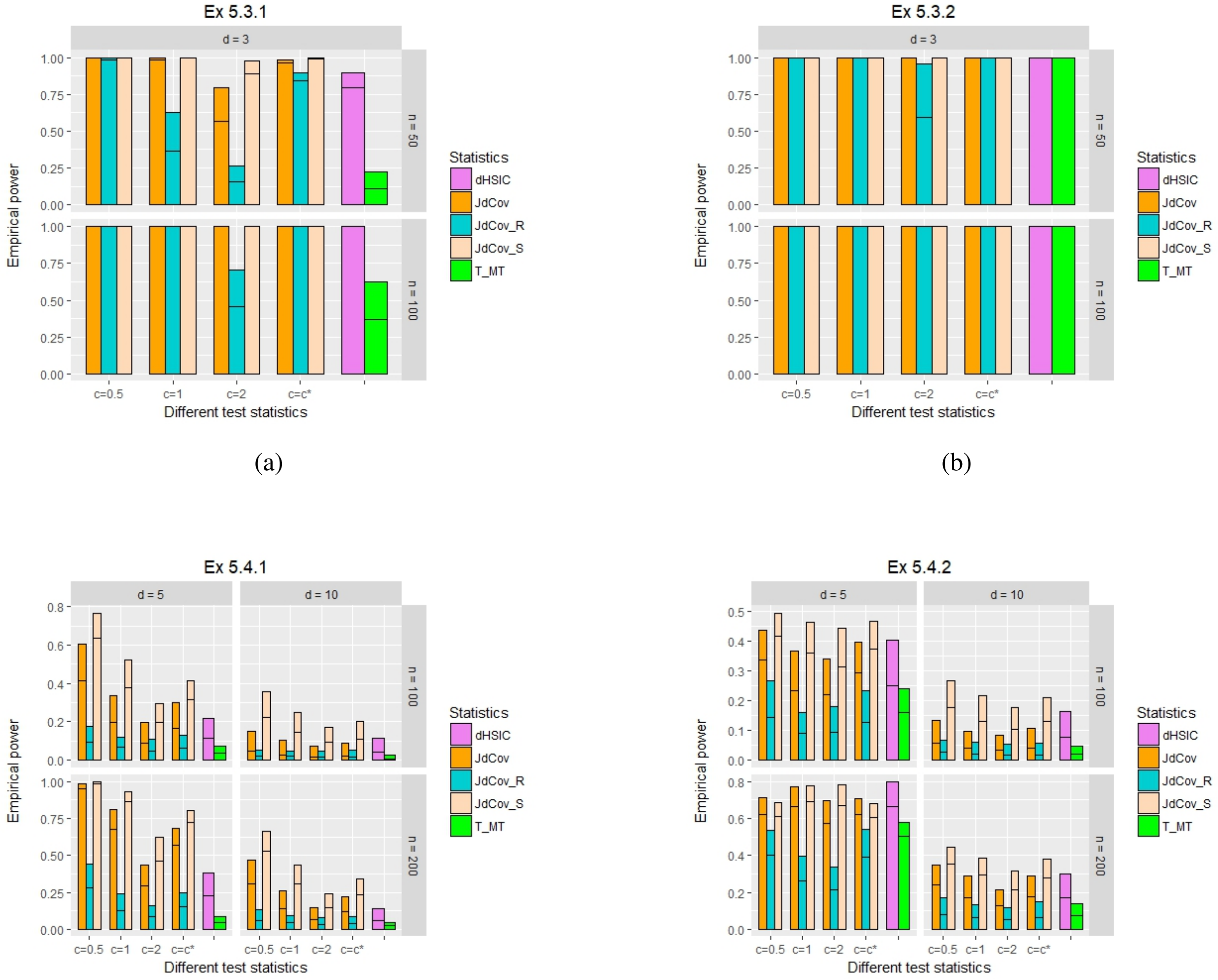
(c)

(d)

(e)

(f)

Figure 2.1: Figures showing the empirical size and power for the different tests statistics in Examples 2.5.1 and 2.5.2. c^* denotes the data-driven choice of c. The vertical height of a bar and a line on a bar stand for the empirical size or power at levels $\alpha = 0.1$ or $\alpha = 0.05$, respectively.



(a)

(b)

(c)

(d)

Figure 2.2: Figures showing the empirical power for the different tests statistics in Examples 2.5.3 and 2.5.4. c^* denotes the data-driven choice of c. The vertical height of a bar and a line on a bar stand for the empirical power at levels $\alpha = 0.1$ or $\alpha = 0.05$, respectively.

where $\hat{f}_{j,k}$ is the B-spline estimator for $f_{j,k}$. To check the goodness of fit of $\mathcal{G}$, we test the joint independence of the residuals. Let T_n be the statistic (e.g. $\widetilde{JdCov}^2$, $\widetilde{JdCov}_S^2$ or $\widetilde{JdCov}_R^2$) to test the joint dependence of $(\epsilon_1, \ldots, \epsilon_d)$ constructed based on the fitted residuals $\hat{\epsilon}_i = (\hat{\epsilon}_{i1}, \ldots, \hat{\epsilon}_{id})$ for $1 \leq i \leq n$. Following the idea presented in Sen and Sen (2014), it seems that T_n might have a limiting distribution different from the one mentioned in Proposition 2.3.2. So to approximate the sampling distribution of T_n, we introduce the following residual bootstrap procedure.

1. Randomly sample $\epsilon_j^* = (\epsilon_{1j}^*, \dots, \epsilon_{nj}^*)$ with replacement from the residuals $\{\hat{\epsilon}_{1j}, \dots, \hat{\epsilon}_{nj}\}$, $1 \leq j \leq d$. Construct the bootstrap sample $X_{ij}^* = \sum_{k \in \text{Par}(j)} \hat{f}_{j,k}(X_{ik}) + \epsilon_{ij}^*$.

2. Based on the bootstrap sample $\{\mathbf{X}_i^*\}_{i=1}^n$ with $\mathbf{X}_i^* = (X_{i1}^*, \dots, X_{id}^*)$, estimate $f_{j,k}$ for $k \in \text{Par}(j)$, and denote the corresponding residuals by $\hat{\epsilon}_{ij}^*$.

3. Calculate the bootstrap statistic T_n^* based on $\{\hat{\epsilon}_{ij}^*\}$.

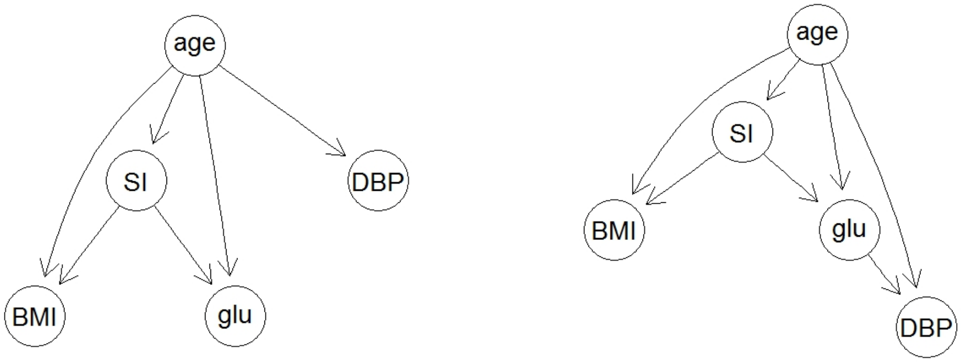
4. Repeat the above steps B times and let $\{T_{b,n}^*\}_{b=1}^B$ be the corresponding values of the bootstrap statistics. The p-value is given by $B^{-1} \sum_{b=1}^B \{T_{b,n}^* > T_n\}$.

Pfister et al. (2018) proposed to bootstrap the residuals directly and used the bootstrapped residuals to construct the test statistic. In contrast, we suggest the use of the above residual bootstrap to capture the estimation effect caused by replacing $f_{j,k}$ with the estimate $\hat{f}_{j,k}$.

2.6.2 Real data example

We now apply the model diagnostic checking procedure for DAG to one real world dataset. A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. We downloaded the data from `https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes`. We focus only on the following five variables : Age, Body Mass Index (BMI), 2-Hour Serum Insulin (SI), Plasma Glucose Concentration (glu) and Diastolic Blood Pressure (DBP). Further, we only selected the instances with non-zero values, as it seems that zero values encode missing data. This yields $n = 392$ samples.

Now, age is likely to affect all the other variables (but of course not the other way round). Moreover, serum insulin also has plausible causal effects on BMI and plasma glucose concentration. We try to determine the correct causal structure out of 48 candidate DAG models and perform model diagnostic checking for each of the 48 models, as illustrated in Section 6.1. We first center each of the variables and scale them so that l_2 norm of each of the variables is $\sqrt{n}$. We perform the mutual independence test of residuals based on the statistics $\widetilde{JdCov}^2$, $\widetilde{JdCov}^2_S$ and $\widetilde{JdCov}^2_R$ with $c = 1$, and compare with the bootstrap-assisted version of the dHSIC-based test proposed in Pfister et al. (2018) and T_{MT}. For each of the tests, we implement the residual bootstrap to obtain the p-value with $B = 1000$. Figure 3.2 shows the selected DAG models corresponding to the largest p-values from each of the five tests.



(a) $\widetilde{JdCov}^2$, $\widetilde{JdCov}^2_S$, $\widetilde{JdCov}^2_R$ and T_{MT}

(b) dHSIC

Figure 2.3: The DAG models corresponding to the largest p-values from the five tests.

Figure 2.3a shows the model with the maximum p-value among all the 48 candidate DAG models, when the test for joint independence of the residuals is conducted based on $\widetilde{JdCov}^2$, $\widetilde{JdCov}^2_S$ and $\widetilde{JdCov}^2_R$ and T_{MT}. This graphical structure goes in tune with the biological evidences of causal relationships among these five variables. Figure 2.3b stands for the model with the maximum p-value when the test is based on dHSIC. Its only difference with Figure 2.3a is that, it has an additional edge from glu to DBP, indicating a causal effect of Plasma Glucose Concentration on Diastolic Blood Pressure. We are unsure of any biological evidence that supports such a causal relationship in reality.

REMARK **2.6.1**. In view of Remark 2.2.3, it might be intriguing to take into account the heuristic data-driven way of determining c in the above example, instead of setting c at a default value of

1. Our findings indicate that choosing c in the data-driven way leads to a slightly different result. The tests based on dHSIC and $\widetilde{JdCov}^2_S$ select the DAG model shown in Figure 2.3b (considering the maximum p-value among all the 48 candidate DAG models), whereas Figure 2.3a is the DAG model selected when the test is based on $\widetilde{JdCov}^2$, $\widetilde{JdCov}^2_R$ and T_{MT}. The proposed tests (based on $\widetilde{JdCov}^2$ and $\widetilde{JdCov}^2_R$) still perform well.

2.6.3 A simulation study

We conduct a simulation study based on our findings in the previous real data example. To save the computational cost, we focus our attention on three of the five variables, viz. Age, glu and DBP. In the correct causal structure among these three variables, there are directed edges from Age to glu and Age to DBP. We consider the additive structural equation models

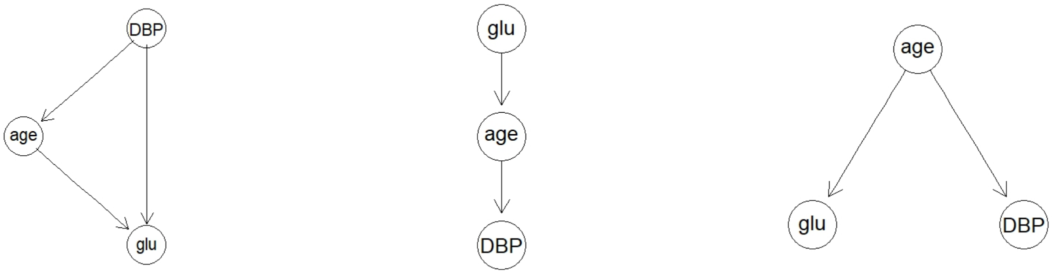
$$X_j = \sum_{k \in \text{Par}(j)} \hat{f}_{j,k}(X_k) + e_j \ , \ j = 1, 2, 3, \tag{2.20}$$

where X_1, X_2, X_3 correspond to Age, glu and DBP (after centering and scaling) respectively, and $\hat{f}_{j,k}$ denotes the estimated function from the real data. Note that X_1 is the only variable without any parent. In Section 2.6.2, we get from our numerical studies that the standard deviation of X_1 is 1.001, and the standard deviations of the residuals when X_2 and X_3 are regressed on X_1 (according to the structural equation models in (2.19), are 0.918 and 0.95, respectively. In this simulation study, we simulate X_1 from a zero mean Gaussian distribution with standard deviation 1. For X_2 and X_3, we simulate the noise variables from zero mean Gaussian distributions with standard deviations 0.918 and 0.95, respectively. The same $n = 392$ is considered for the number of generated observations, and based on this simulated dataset we perform the model diagnostic checking for 27 candidate DAG models. The number of bootstrap replications is set to be $B = 100$ (to save the computational cost). This procedure is repeated 100 times to note how many times out of 100 that the five tests select the correct model, based on the largest p-value. The results in Table 2.2 below indicate that the proposed tests with $c = 1$ and the dHSIC-based test outperform T_{MT}.

Table 2.2: The number of times (out of 100) that the true model is being selected.

$\widetilde{JdCov^2}$	$\widetilde{JdCov_S^2}$	$\widetilde{JdCov_R^2}$	dHSIC	T_{MT}
45	61	54	52	32

REMARK **2.6.2**. A natural question to raise is why do we bootstrap the residuals and not test for the joint independence of the estimated residuals directly, to check for the goodness of fit of the DAG model. From the idea in Sen and Sen (2014), it appears that the joint distance covariance of the estimated residuals might have a limiting distribution different from the one stated in Proposition 2.3.2. We leave the formulation of a rigorous theory in support of that for future research. We present below the models selected most frequently (out of 100 times) by the different test statistics if we repeat the simulation study done above in Section 2.6.3 without using residual bootstrap to re-estimate $f_{j,k}$. We immediately see that joint independence tests of the estimated residuals based on all of the five statistics we consider, select a DAG model that is meaningless and far away from the correct one.



(a) $\widetilde{JdCov^2}$, $\widetilde{JdCov_S^2}$, $\widetilde{JdCov_R^2}$, dHSIC

(b) T_{MT}

(c) Correct model

Figure 2.4: The DAG models selected (most frequently out of 100 times) by the five tests, without doing residual bootstrap to re-estimate $f_{j,k}$.

REMARK **2.6.3**. In view of Remark 2.2.3, it might be intriguing to take into account the heuristic data-driven way of choosing c in the simulation study in Section 2.6.3, instead of setting c at a default value of 1. Our findings indicate that our proposed tests and the dHSIC-based test still outperform T_{MT}. In the context of Remark 2.6.2, if we repeat the simulation study done in Section 2.6.3 (choosing c in the heuristic way), we still reach the same conclusion presented in Remark 2.6.2.

2.7 Discussions

Huo and Székely (2016) proposed an $O(n \log n)$ algorithm to compute dCov of univariate random variables. In a more recent work, Huang and Huo (2017) introduced a fast method for multivariate cases which is based on random projection and has computational complexity $O(nK \log n)$, where K is the number of random projections. One of the possible directions for future research is to come up with a fast algorithm to compute JdCov. When $p_i = 1$, we can indeed use the method in Huo and Székely (2016) to compute JdCov. But their method may be inefficient when d is large and it is not applicable to the case where $p_i > 1$. Another direction is, to introduce the notion of Conditional JdCov in light of Wang et al. (2015), to test if the variables $(X_1, \ldots, X_d)$ are jointly independent given another variable Z.

3. NONPARAMETRIC TESTS FOR INDEPENDENCE AND EQUALITY OF DISTRIBUTIONS IN HIGH DIMENSIONS

3.1 Background and notations

The behavior of the classical distance and kernel-based tests for independence and equality of distributions in the high dimensional setup is still a pretty unexplored area. In a very recent work, Zhu et al. (2020) showed that in the high dimension low sample size (HDLSS) setting, i.e., when the dimensions grow while the sample size is held fixed, the sample distance covariance can only measure the component-wise *linear dependence* between the two vectors. As a consequence, the distance correlation based t-test proposed by Székely et al. (2013) for independence between two high dimensional random vectors has trivial power when the two random vectors are nonlinearly dependent but component-wise uncorrelated. As a remedy, Zhu et al. (2020) proposed a test by aggregating the pairwise squared sample distance covariances and studied its asymptotic behavior under the HDLSS setup.

This work presents a new class of metrics to quantify the homogeneity of distributions and independence between two high-dimensional random vectors. The core of our methodology is a new way of defining the distance between sample points (interpoint distance) in the high-dimensional Euclidean spaces. In the first part of this work, we show that the energy distance based on the usual Euclidean distance cannot completely characterize the homogeneity of two high-dimensional distributions in the sense that it only detects the *equality of means and the traces of covariance matrices* in the high-dimensional setup. To overcome such a limitation, we propose a new class of metrics based on the new distance which inherits the nice properties of energy distance and maximum mean discrepancy in the low-dimensional setting and is capable of detecting the *pairwise homogeneity of the low-dimensional marginal distributions* in the HDLSS setup. We construct a high-dimensional two sample t-test based on the U-statistic type estimator of the proposed metric, which can be viewed as a generalization of the classical two-sample t-test with equal variances. We show under the HDLSS setting that the new two sample t-test converges to a central t-distribution under the null and it has nontrivial power for a broader class of alternatives compared to the energy distance. We further show that the two sample t-test converges to a standard normal limit under the null when the dimension and sample size both grow to infinity with the dimension growing more rapidly. It is worth mentioning that we develop an approach to unify the analysis for the usual energy distance and the proposed metrics. Compared to existing works, we make the following contribution.

- We derive the asymptotic variance of the generalized energy distance under the HDLSS setting and propose a computationally efficient variance estimator (whose computational cost is linear in the dimension). Our analysis is based on a pivotal t-statistic which does not require permutation or resampling-based inference and allows an asymptotic exact power analysis.

In the second part, we propose a new framework to construct dependence metrics to quantify the dependence between two high-dimensional random vectors X and Y of possibly different dimensions. The new metric, denoted by $\mathcal{D}^2(X, Y)$, generalizes both the distance covariance and HSIC. It completely characterizes independence between X and Y and inherits all other desirable properties of the distance covariance and HSIC for fixed dimensions. In the HDLSS setting, we show that the proposed population dependence metric behaves as an aggregation of group-wise (generalized) distance covariances. We construct an unbiased U-statistic type estimator of $\mathcal{D}^2(X, Y)$ and show that with growing dimensions, the unbiased estimator is asymptotically equivalent to the sum of group-wise squared sample (generalized) distance covariances. Thus it can quantify *group-wise non-linear dependence* between two high-dimensional random vectors, going beyond the scope of the distance covariance based on the usual Euclidean distance and HSIC which have been recently shown only to capture the componentwise linear dependence in high dimension, see Zhu et al. (2020). We further propose a t-test based on the new metrics to perform high-dimensional independence testing and study its asymptotic size and power behaviors under both the HDLSS and high dimension medium sample size (HDMSS) setups. In particular, under the HDLSS setting, we prove that the proposed t-test converges to a central t-distribution under the null and a noncentral t-distribution with a random noncentrality parameter under the alternative. Through extensive numerical studies, we demonstrate that the newly proposed t-test can capture group-wise nonlinear dependence which cannot be detected by the usual distance covariance and HSIC in the high dimensional regime. Compared to the marginal aggregation approach in Zhu et al. (2020), our new method enjoys two major advantages.

- Our approach provides a neater way of generalizing the notion of distance and kernel-based dependence metrics. The newly proposed metrics completely characterize dependence in the low-dimensional case and capture group-wise nonlinear dependence in the high-dimensional case. In this sense, our metric can detect a wider range of dependence compared to the marginal aggregation approach.
- The computational complexity of the t-tests only grows linearly with the dimension and thus is scalable to very high dimensional data.

Notation. Let $X = (X_1, \dots X_p) \in \mathbb{R}^p$ and $Y = (Y_1, \dots, Y_q) \in \mathbb{R}^q$ be two random vectors of dimensions p and q respectively. Denote by $\| \cdot \|_p$ the Euclidean norm of $\mathbb{R}^p$ (we shall use it interchangeably with $\| \cdot \|$ when there is no confusion). Let 0_p be the origin of $\mathbb{R}^p$. We use $X \perp\!\!\!\perp Y$ to denote that X is independent of Y, and use "$X \overset{d}{=} Y$" to indicate that X and Y are identically distributed. Let (X', Y'), (X'', Y'') and (X''', Y''') be independent copies of (X, Y). We utilize the order in probability notations such as stochastic boundedness O_p (big O in probability), convergence in probability o_p (small o in probability) and equivalent order $\asymp$, which is defined as follows: for a sequence of random variables $\{Z_n\}_{n=1}^{\infty}$ and a sequence of real numbers $\{a_n\}_{n=1}^{\infty}$, $Z_n \asymp_p a_n$ if and only if $Z_n/a_n = O_p(1)$ and $a_n/Z_n = O_p(1)$ as $n \to \infty$. For a metric space $(\mathcal{X}, d_{\mathcal{X}})$, let $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}_1(\mathcal{X})$ denote the set of all finite signed Borel measures on $\mathcal{X}$ and all probability measures on $\mathcal{X}$, respectively. Define $\mathcal{M}^1_{d_{\mathcal{X}}}(\mathcal{X}) := \{v \in \mathcal{M}(\mathcal{X}) \, : \, \exists\, x_0 \in \mathcal{X} \text{ s.t. } \int_{\mathcal{X}} d_{\mathcal{X}}(x, x_0)\, d|v|(x) < \infty\}$. For $\theta > 0$, define $\mathcal{M}^{\theta}_{\mathcal{K}}(\mathcal{X}) := \{v \in \mathcal{M}(\mathcal{X}) \, : \, \int_{\mathcal{X}} \mathcal{K}^{\theta}(x, x)\, d|v|(x) < \infty\}$, where $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a bivariate kernel function. Define $\mathcal{M}^1_{d_{\mathcal{Y}}}(\mathcal{Y})$ and $\mathcal{M}^{\theta}_{\mathcal{K}}(\mathcal{Y})$ in a similar way. For a matrix

$A = (a_{kl})_{k,l=1}^n \in \mathbb{R}^{n\times n}$, define its $\mathcal{U}$-centered version $\tilde{A} = (\tilde{a}_{kl}) \in \mathbb{R}^{n\times n}$ as follows

$$\tilde{a}_{kl} = \begin{cases} a_{kl} - \dfrac{1}{n-2}\displaystyle\sum_{j=1}^n a_{kj} - \dfrac{1}{n-2}\displaystyle\sum_{i=1}^n a_{il} + \dfrac{1}{(n-1)(n-2)}\displaystyle\sum_{i,j=1}^n a_{ij}, & k \neq l, \\ 0, & k = l, \end{cases} \tag{3.1}$$

for $k, l = 1, \ldots, n$. Define

$$(\tilde{A}\cdot\tilde{B}) := \frac{1}{n(n-3)}\sum_{k\neq l}\tilde{a}_{kl}\tilde{b}_{kl}$$

for $\tilde{A} = (\tilde{a}_{kl})$ and $\tilde{B} = (\tilde{b}_{kl}) \in \mathbb{R}^{n\times n}$. Denote by $\mathrm{tr}(A)$ the trace of a square matrix A. $A \otimes B$ denotes the kronecker product of two matrices A and B. Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution. Denote by $t_{a,b}$ the noncentral t-distribution with a degrees of freedom and noncentrality parameter b. Write $t_a = t_{a,0}$. Denote by $q_{\alpha,a}$ and Z_α the upper α quantile of the distribution of t_a and the standard normal distribution, respectively, for $\alpha \in (0,1)$. Also denote by χ^2_a the chi-square distribution with a degrees of freedom. Denote $U \sim \text{Rademacher}\,(0.5)$ if $P(U=1) = P(U=-1) = 0.5$. Let $\mathbb{1}_A$ denote the indicator function associated with a set A. Finally, denote by $\lfloor a \rfloor$ the integer part of $a \in \mathbb{R}$.

3.2 New distance for Euclidean space

We introduce a family of distances for Euclidean space, which shall play a central role in the subsequent developments. For $x \in \mathbb{R}^{\tilde{p}}$, we partition x into p sub-vectors or groups, namely $x = (x_{(1)}, \ldots, x_{(p)})$, where $x_{(i)} \in \mathbb{R}^{d_i}$ with $\sum_{i=1}^p d_i = \tilde{p}$. Let ρ_i be a metric or semimetric (see for example Definition 1 in Sejdinovic et al. (2013)) defined on $\mathbb{R}^{d_i}$ for $1 \leq i \leq p$. We define a family of distances for $\mathbb{R}^{\tilde{p}}$ as

$$K_{\mathbf{d}}(x, x') := \sqrt{\rho_1(x_{(1)}, x'_{(1)}) + \ldots + \rho_p(x_{(p)}, x'_{(p)})}\,, \tag{3.2}$$

where $x, x' \in \mathbb{R}^{\tilde{p}}$ with $x = (x_{(1)}, \ldots, x_{(p)})$ and $x' = (x'_{(1)}, \ldots, x'_{(p)})$, and $\mathbf{d} = (d_1, d_2, \ldots, d_p)$ with $d_i \in \mathbb{Z}_+$ and $\sum_{i=1}^p d_i = \tilde{p}$.

Proposition **3.2.1**. *Suppose each ρ_i is a metric of strong negative type on $\mathbb{R}^{d_i}$. Then $\big(\mathbb{R}^{\tilde{p}}, K_{\mathbf{d}}\big)$ satisfies the following two properties:*

1. *$K_{\mathbf{d}} : \mathbb{R}^{\tilde{p}} \times \mathbb{R}^{\tilde{p}} \to [0, \infty)$ is a valid metric on $\mathbb{R}^{\tilde{p}}$;*
2. *$\big(\mathbb{R}^{\tilde{p}}, K_{\mathbf{d}}\big)$ has strong negative type.*

In a special case, suppose ρ_i is the Euclidean distance on $\mathbb{R}^{d_i}$. By Theorem 3.16 in Lyons (2013), $(\mathbb{R}^{d_i}, \rho_i)$ is a separable Hilbert space, and hence has strong negative type. Then the Euclidean space equipped with the metric

$$K_{\mathbf{d}}(x, x') \;=\; \sqrt{\|x_{(1)} - x'_{(1)}\| \,+\, \ldots \,+\, \|x_{(p)} - x'_{(p)}\|}\;. \tag{3.3}$$

is of strong negative type. Further, if all the components $x_{(i)}$ are unidimensional, i.e., $d_i = 1$ for $1 \le i \le p$, then the metric boils down to

$$K_{\mathbf{d}}(x, x') \;=\; \|x - x'\|_1^{1/2} \;=\; \sqrt{\sum_{j=1}^{p} |x_j - x'_j|}\;, \tag{3.4}$$

where $\|x\|_1 = \sum_{j=1}^{p} |x_j|$ is the l_1 or the absolute norm on $\mathbb{R}^p$. If

$$\rho_i(x_{(i)}, x'_{(i)}) = \|x_{(i)} - x'_{(i)}\|^2, \quad 1 \le i \le p, \tag{3.5}$$

then $K_{\mathbf{d}}$ reduces to the usual Euclidean distance. We shall unify the analysis of our new metrics with the classical metrics by considering $K_{\mathbf{d}}$ which is defined in (3.2) with

S1 each ρ_i being a metric of strong negative type on $\mathbb{R}^{d_i}$;

S2 each ρ_i being a semimetric defined in (3.5).

The first case corresponds to the newly proposed metrics while the second case leads to the classical metrics based on the usual Euclidean distance. Remarks 3.2.1 and 3.2.2 provide two different ways of generalizing the class in (3.2). To be focused, our analysis below shall only concern about the distances defined in (3.2). In the numerical studies in Section 4.4, we consider ρ_i to be the Euclidean distance and the distances induced by the Laplace and Gaussian kernels (see Definition 1.2.3) which are of strong negative type on $\mathbb{R}^{d_i}$ for $1 \leq i \leq p$.

Remark **3.2.1**. *A more general family of distances can be defined as*

$$K_{\mathbf{d},r}(x,x') = \Big(\rho_1(x_{(1)},x'_{(1)}) + \cdots + \rho_p(x_{(p)},x'_{(p)})\Big)^r, \quad 0 < r < 1.$$

According to Remark 3.19 of Lyons (2013), the space $(\mathbb{R}^{\tilde{p}}, K_{\mathbf{d},r})$ *is of strong negative type. The proposed distance is a special case with* $r = 1/2$.

Remark **3.2.2**. *Based on the proposed distance, one can construct the generalized Gaussian and Laplacian kernels as*

$$f(K_{\mathbf{d}}(x,x')/\gamma) = \begin{cases} \exp(-K_{\mathbf{d}}^2(x,x')/\gamma^2), & f(x) = \exp(-x^2) \textit{ for Gaussian kernel}, \\ \exp(-K_{\mathbf{d}}(x,x')/\gamma), & f(x) = \exp(-x) \textit{ for Laplacian kernel}. \end{cases}$$

If $K_{\mathbf{d}}$ *is translation invariant, then by Theorem 9 in Sriperumbudur et al. (2010) it can be verified that* $f(K_{\mathbf{d}}(x,x')/\gamma)$ *is a characteristic kernel on* $\mathbb{R}^{\tilde{p}}$. *As a consequence, the Euclidean space equipped with the distance*

$$K_{\mathbf{d},f}(x,x') = f(K_{\mathbf{d}}(x,x)/\gamma) + f(K_{\mathbf{d}}(x',x')/\gamma) - 2f(K_{\mathbf{d}}(x,x')/\gamma)$$

is of strong negative type.

Remark **3.2.3**. *In Sections 3.3 and 3.4 we develop new classes of homogeneity and dependence metrics to quantify the pairwise homogeneity of distributions or the pairwise non-linear dependence of the low-dimensional groups. A natural question to arise in this regard is how to partition the random vectors optimally in practice. We present some real data examples in Section 3.5.3 where all the group sizes have been considered to be one (as a special case of the general theory proposed in this work), and an additional real data example in Section B.3 of the appendix where* *the data admits some natural grouping. We believe this partitioning can be very much problem specific and may require subject knowledge. We leave it for future research to develop an algorithm to find the optimal groups using the data and perhaps some auxiliary information.*

3.3 Homogeneity metrics

Consider $X, Y \in \mathbb{R}^{\tilde{p}}$. Suppose X and Y can be partitioned into p sub-vectors or groups, viz. $X = \big(X_{(1)}, X_{(2)}, \ldots, X_{(p)}\big)$ and $Y = \big(Y_{(1)}, Y_{(2)}, \ldots, Y_{(p)}\big)$, where the groups $X_{(i)}$ and $Y_{(i)}$ are d_i dimensional, $1 \leq i \leq p$, and p might be fixed or growing. We assume that $X_{(i)}$ and $Y_{(i)}$'s are finite (low) dimensional vectors, i.e., $\{d_i\}_{i=1}^{p}$ is a bounded sequence. Clearly $\tilde{p} = \sum_{i=1}^{p} d_i = O(p)$. Denote the mean vectors and the covariance matrices of X and Y by μ_X and μ_Y, and, Σ_X and Σ_Y, respectively. We propose the following class of metrics $\mathcal{E}$ to quantify the homogeneity of the distributions of X and Y:

$$\mathcal{E}(X,Y) \;=\; 2\,\mathbb{E}\,K_{\mathbf{d}}(X,Y) \;-\; \mathbb{E}\,K_{\mathbf{d}}(X,X') \;-\; \mathbb{E}\,K_{\mathbf{d}}(Y,Y')\;, \tag{3.6}$$

with $\mathbf{d} = (d_1, \ldots, d_p)$. We shall drop the subscript $\mathbf{d}$ below for the ease of notation.

Assumption **3.3.1**. *Assume that* $\sup_{1\leq i\leq p} \mathbb{E}\,\rho_i^{1/2}(X_{(i)}, 0_{d_i}) < \infty$ *and* $\sup_{1\leq i\leq p} \mathbb{E}\,\rho_i^{1/2}(Y_{(i)}, 0_{d_i})$ $< \infty$.

Under Assumption 4.3.1, $\mathcal{E}$ is finite. In Section B.1.1 of the appendix we illustrate that in the low-dimensional setting, $\mathcal{E}(X,Y)$ completely characterizes the homogeneity of the distributions of X and Y.

Consider i.i.d. samples $\{X_k\}_{k=1}^{n}$ and $\{Y_l\}_{l=1}^{m}$ from the respective distributions of X and $Y \in \mathbb{R}^{\tilde{p}}$, where $X_k = (X_{k(1)}, \ldots, X_{k(p)})$, $Y_l = (Y_{l(1)}, \ldots, Y_{l(p)})$ for $1 \leq k \leq n$, $1 \leq l \leq m$ and $X_{k(i)}, Y_{l(i)} \in \mathbb{R}^{d_i}$. We propose an unbiased U-statistic type estimator $\mathcal{E}_{n,m}(X,Y)$ of $\mathcal{E}(X,Y)$ as in equation (4.6) with d being the new metric K. We refer the reader to Section B.1.1 of the appendix, where we show that $\mathcal{E}_{n,m}(X,Y)$ essentially inherits all the nice properties of the U-statistic type estimator of generalized energy distance and MMD.

We define the following quantities which will play an important role in our subsequent analysis:

$$\tau_X^2 = \mathbb{E}\, K(X, X')^2, \quad \tau_Y^2 = \mathbb{E}\, K(Y, Y')^2, \quad \tau^2 = \mathbb{E}\, K(X, Y)^2. \tag{3.7}$$

In Case S2 (i.e., when K is the Euclidean distance), we have

$$\tau_X^2 = 2\mathrm{tr}\Sigma_X, \quad \tau_Y^2 = 2\mathrm{tr}\Sigma_Y, \quad \tau^2 = \mathrm{tr}\Sigma_X + \mathrm{tr}\Sigma_Y + \|\mu_X - \mu_Y\|^2. \tag{3.8}$$

Under the null hypothesis $H_0 : X \overset{d}{=} Y$, it is clear that $\tau_X^2 = \tau_Y^2 = \tau^2$.

In the subsequent discussion we study the asymptotic behavior of $\mathcal{E}$ in the high-dimensional framework, i.e., when p grows to ∞ with fixed n and m (discussed in Subsection 3.3.1) and when n and m grow to ∞ as well (discussed in Subsection B.2.1 in the appendix). We point out some limitations of the test for homogeneity of distributions in the high-dimensional setup based on the usual Euclidean energy distance. Consequently we propose a test based on the proposed metric and justify its consistency for growing dimension.

3.3.1 High dimension low sample size (HDLSS)

In this subsection, we study the asymptotic behavior of the Euclidean energy distance and our proposed metric $\mathcal{E}$ when the dimension grows to infinity while the sample sizes n and m are held fixed. We make the following moment assumption.

ASSUMPTION **3.3.2**. *There exist constants a, a', a'', A, A', A'' such that uniformly over p,*

$$0 < a \leq \inf_{1\leq i\leq p} \mathbb{E}\, \rho_i(X_{(i)}, X'_{(i)}) \leq \sup_{1\leq i\leq p} \mathbb{E}\, \rho_i(X_{(i)}, X'_{(i)}) \leq A < \infty,$$

$$0 < a' \leq \inf_{1\leq i\leq p} \mathbb{E}\, \rho_i(Y_{(i)}, Y'_{(i)}) \leq \sup_{1\leq i\leq p} \mathbb{E}\, \rho_i(Y_{(i)}, Y'_{(i)}) \leq A' < \infty,$$

$$0 < a'' \leq \inf_{1\leq i\leq p} \mathbb{E}\, \rho_i(X_{(i)}, Y_{(i)}) \leq \sup_{1\leq i\leq p} \mathbb{E}\, \rho_i(X_{(i)}, Y_{(i)}) \leq A'' < \infty.$$

Under Assumption 3.3.2, it is not hard to see that $\tau_X, \tau_Y, \tau \asymp p^{1/2}$. The proposition below provides an expansion for K evaluated at random samples.

PROPOSITION **3.3.1**. *Under Assumption 3.3.2, we have*

$$\frac{K(X,X')}{\tau_X} = 1 + \frac{1}{2}L_X(X,X') + R_X(X,X'), \tag{3.9}$$

$$\frac{K(Y,Y')}{\tau_Y} = 1 + \frac{1}{2}L_Y(Y,Y') + R_Y(Y,Y'), \tag{3.10}$$

and

$$\frac{K(X,Y)}{\tau} = 1 + \frac{1}{2}L(X,Y) + R(X,Y), \tag{3.11}$$

where

$$L_X(X,X') := \frac{K^2(X,X') - \tau_X^2}{\tau_X^2}, \quad L_Y(Y,Y') := \frac{K^2(Y,Y') - \tau_Y^2}{\tau_Y^2}, \quad L(X,Y) := \frac{K^2(X,Y) - \tau^2}{\tau^2},$$

and $R_X(X,X'), R_Y(Y,Y'), R(X,Y)$ are the remainder terms. In addition, if $L_X(X,X'), L_Y(Y,Y')$ and $L(X,Y)$ are $o_p(1)$ random variables as $p \to \infty$, then $R_X(X,X') = O_p\left(L_X^2(X,X')\right)$, $R_Y(Y,Y') = O_p\left(L_Y^2(Y,Y')\right)$ and $R(X,Y) = O_p\left(L^2(X,Y)\right)$.

Henceforth we will drop the subscripts X and Y from L_X, L_Y, R_X and R_Y for notational convenience. Theorem 1 and Lemma 3.3.1 below provide insights into the behavior of $\mathcal{E}(X,Y)$ in the high-dimensional framework.

ASSUMPTION **3.3.3**. *Assume that $L(X,Y) = O_p(a_p)$, $L(X,X') = O_p(b_p)$ and $L(Y,Y') = O_p(c_p)$, where a_p, b_p, c_p are positive real sequences satisfying $a_p = o(1)$, $b_p = o(1)$, $c_p = o(1)$ and $\tau a_p^2 + \tau_X b_p^2 + \tau_Y c_p^2 = o(1)$.*

REMARK **3.3.1**. *To illustrate Assumption 3.3.3, we observe that under assumption 3.3.2 we can write*

$$var\left(L(X,X')\right) = O\Big(\frac{1}{p^2}\Big)\sum_{i,j=1}^{p} cov\left(\rho_i(X_{(i)},X'_{(i)}), \rho_j(X_{(j)},X'_{(j)})\right) = O\Big(\frac{1}{p^2}\Big)\sum_{i,j=1}^{p} cov\left(Z_i, Z_j\right),$$

where $Z_i := \rho_i(X_{(i)}, X'_{(i)})$ for $1 \leq i \leq p$. Assume that $\sup_{1\leq i\leq p} \mathbb{E}\, \rho_i^2(X_{(i)}, 0_{d_i}) < \infty$, which implies $\sup_{1\leq i\leq p} \mathbb{E}\, Z_i^2 < \infty$. Under certain strong mixing conditions or in general certain weak *dependence assumptions, it is not hard to see that* $\sum_{i,j=1}^{p} cov\,(Z_i, Z_j) = O(p)$ *as* $p \to \infty$ *(see for example Theorem 1.2 in Rio (1993) or Theorem 1 in Doukhan et al. (1999)). Therefore we have* $var\,(L(X, X')) = O(\frac{1}{p})$ *and hence by Chebyshev's inequality, we have* $L(X, X') = O_p(\frac{1}{\sqrt{p}})$*. We refer the reader to Remark 2.1.1 in Zhu et al. (2020) for illustrations when each* ρ_i *is the squared Euclidean distance.*

THEOREM **1**. *Suppose Assumptions 3.3.2 and 3.3.3 hold. Further assume that the following three sequences*

$$\left\{\frac{\sqrt{p}L^2(X,Y)}{1+L(X,Y)}\right\}, \ \left\{\frac{\sqrt{p}L^2(X,X')}{1+L(X,X')}\right\}, \ \left\{\frac{\sqrt{p}L^2(Y,Y')}{1+L(Y,Y')}\right\}$$

indexed by p *are all uniformly integrable. Then we have*

$$\mathcal{E}(X,Y) \ = \ 2\tau - \tau_X - \tau_Y \ + \ o(1). \tag{3.12}$$

REMARK **3.3.2**. *Remark B.4.1 in the appendix provides some illustrations on certain sufficient conditions under which* $\{\sqrt{p}L^2(X,Y)/(1 + L(X,Y))\}$*,* $\{\sqrt{p}L^2(X,X')/(1 + L(X,X'))\}$ *and* $\{\sqrt{p}L^2(Y,Y')/(1 + L(Y,Y'))\}$ *are uniformly integrable.*

REMARK **3.3.3**. *To illustrate that the leading term in equation (3.12) indeed gives a close approximation of the population* $\mathcal{E}(X,Y)$*, we consider the special case when* K *is the Euclidean distance. Suppose* $X \sim N_p(0, I_p)$ *and* $Y = X + N$ *where* $N \sim N_p(0, I_p)$ *with* $N \perp\!\!\!\perp X$*. Clearly from (3.8) we have* $\tau_X^2 = 2p$*,* $\tau_Y^2 = 4p$ *and* $\tau^2 = 3p$*. We simulate large samples of sizes* $m = n = 5000$ *from the distributions of* X *and* Y *for* $p = 20, 40, 60, 80$ *and* 100*. The large sample sizes are to ensure that the U-statistic type estimator of* $\mathcal{E}(X,Y)$ *gives a very close approximation of the population* $\mathcal{E}(X,Y)$*. In Table 3.1 we list the ratio between* $\mathcal{E}(X,Y)$ *and the leading term in (3.12) for the different values of* p*, which turn out to be very close to* 1*, demonstrating that the leading term in (3.12) indeed approximates* $\mathcal{E}(X,Y)$ *reasonably well.*

LEMMA **3.3.1**. *Assume* $\tau, \tau_X, \tau_Y < \infty$*. We have*

1. *In Case S1,* $2\tau - \tau_X - \tau_Y = 0$ *if and only if* $X_{(i)} \overset{d}{=} Y_{(i)}$ *for* $i \in \{1, \ldots, p\}$*;*

Table 3.1: Ratio of $\mathcal{E}(X,Y)$ and the leading term in (3.12) for different values of p.

$p=20$	$p=40$	$p=60$	$p=80$	$p=100$
0.995	0.987	0.992	0.997	0.983

2. *In Case S2,* $2\tau - \tau_X - \tau_Y = 0$ *if and only if* $\mu_X = \mu_Y$ *and* $tr\,\Sigma_X = tr\,\Sigma_Y$.

It is to be noted that assuming $\tau, \tau_X, \tau_Y < \infty$ does not contradict with the growth rate $\tau, \tau_X, \tau_Y = O(p^{1/2})$. Clearly under H_0, $2\tau - \tau_X - \tau_Y = 0$ irrespective of the choice of K. In view of Lemma 3.3.1 and Theorem 1, in Case S2, the leading term of $\mathcal{E}(X,Y)$ becomes zero if and only if $\mu_X = \mu_Y$ and $\mathrm{tr}\,\Sigma_X = \mathrm{tr}\,\Sigma_Y$. In other words, when dimension grows high, the Euclidean energy distance can only capture the equality of the means and the first spectral means, whereas our proposed metric captures the pairwise homogeneity of the low dimensional marginal distributions of $X_{(i)}$ and $Y_{(i)}$. Clearly $X_{(i)} \overset{d}{=} Y_{(i)}$ for $1 \leq i \leq p$ implies $\mu_X = \mu_Y$ and $\mathrm{tr}\,\Sigma_X = \mathrm{tr}\,\Sigma_Y$. Thus the proposed metric can capture a wider range of inhomogeneity of distributions than the Euclidean energy distance.

Define

$$\begin{aligned} d_{kl}(i) :&= \rho_i(X_{k(i)}, Y_{l(i)}) \,-\, \mathbb{E}\left[\rho_i(X_{k(i)}, Y_{l(i)})|X_{k(i)}\right] \,-\, \mathbb{E}\left[\rho_i(X_{k(i)}, Y_{l(i)})|Y_{l(i)}\right] \\ &\quad + \mathbb{E}\left[\rho_i(X_{k(i)}, Y_{l(i)})\right], \end{aligned}$$

as the double-centered distance between $X_{k(i)}$ and $Y_{l(i)}$ for $1 \leq i \leq p$, $1 \leq k \leq n$ and $1 \leq l \leq m$. Similarly define $d^X_{kl}(i)$ and $d^Y_{kl}(i)$ as the double-centered distances between $X_{k(i)}$ and $X_{l(i)}$ for $1 \leq k \neq l \leq n$, and, $Y_{k(i)}$ and $Y_{l(i)}$ for $1 \leq k \neq l \leq m$, respectively. Further define $H(X_k, Y_l) := \frac{1}{\tau}\sum_{i=1}^p d_{kl}(i)$ for $1 \leq k \leq n$, $1 \leq l \leq m$, $H(X_k, X_l) := \frac{1}{\tau_X}\sum_{i=1}^p d^X_{kl}(i)$ for $1 \leq k \neq l \leq n$ and $H(Y_k, Y_l)$ in a similar way.

We impose the following conditions to study the asymptotic behavior of the (unbiased) U-statistic type estimator of $\mathcal{E}(X,Y)$ in the HDLSS setup.

Assumption **3.3.4**. *For fixed n and m, as $p \to \infty$,*

$$\begin{pmatrix} H(X_k, Y_l) \\ H(X_s, X_t) \\ H(Y_u, Y_v) \end{pmatrix}_{k,l,\, s<t,\, u<v} \xrightarrow{d} \begin{pmatrix} a_{kl} \\ b_{st} \\ c_{uv} \end{pmatrix}_{k,l,\, s<t,\, u<v},$$

where $\{a_{kl}, b_{st}, c_{uv}\}_{k,l,\, s<t,\, u<v}$ are jointly Gaussian with zero mean. Further we assume that

$$\begin{aligned} var(a_{kl}) &:= \sigma^2 = \lim_{p\to\infty} \mathbb{E}\left[H^2(X_k, Y_l)\right], \\ var(b_{st}) &:= \sigma_X^2 = \lim_{p\to\infty} \mathbb{E}\left[H^2(X_s, X_t)\right], \\ var(c_{uv}) &:= \sigma_Y^2 = \lim_{p\to\infty} \mathbb{E}\left[H^2(Y_u, Y_v)\right]. \end{aligned}$$

$\{a_{kl}, b_{st}, c_{uv}\}_{k,l,\, s<t,\, u<v}$ are all independent with each other.

Due to the double-centering property and the independence between the two samples, it is straightforward to verify that $\{H(X_k, Y_l), H(X_s, X_t), H(Y_u, Y_v)\}_{k,l,s<t,u<t}$ are uncorrelated with each other. So it is natural to expect that the limit $\{a_{kl}, b_{st}, c_{uv}\}_{k,l,\, s<t,\, u<v}$ are all independent with each other.

Remark **3.3.4**. *The above multi-dimensional central limit theorem is classic and can be derived under suitable moment and weak dependence assumptions on the components of X and Y, such as mixing or near epoch dependent conditions. We refer the reader to Doukhan and Neumann (2008) for a review on central limit theorem results under weak dependence assumptions.*

We describe a new two-sample t-test for testing the null hypothesis $H_0 : X \overset{d}{=} Y$. The t statistic can be constructed based on either the Euclidean energy distance or the new homogeneity metrics. We show that the t-tests based on different metrics can have strikingly different power behaviors under the HDLSS setup. The major difficulty here is to introduce a consistent and computationally efficient variance estimator. Towards this end, we define a quantity called Cross Distance Covariance (cdCov) between X and Y, which plays an important role in the construction of the t-test statistic:

$$cdCov^2_{n,m}(X,Y) := \frac{1}{(n-1)(m-1)} \sum_{k=1}^{n} \sum_{l=1}^{m} \widehat{K}(X_k, Y_l)^2,$$

where

$$\widehat{K}(X_k, Y_l) \;=\; K(X_k, Y_l) - \frac{1}{n}\sum_{i=1}^{n} K(X_i, Y_l) - \frac{1}{m}\sum_{j=1}^{m} K(X_k, Y_j) + \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} K(X_i, Y_j).$$

Let $v_s := s(s-3)/2$ for $s = m, n$. We introduce the following quantities

$$\begin{aligned} m_0 &:= \frac{\sigma^2\,(n-1)(m-1) + \sigma_X^2\,v_n + \sigma_Y^2\,v_m}{(n-1)(m-1) + v_n + v_m}\,, \\ \sigma_{nm} &:= \sqrt{\frac{\sigma^2}{nm} + \frac{\sigma_X^2}{2n(n-1)} + \frac{\sigma_Y^2}{2m(m-1)}}\,, \\ a_{nm} &:= \sqrt{\frac{1}{nm} + \frac{1}{2n(n-1)} + \frac{1}{2m(m-1)}}\,, \\ \Delta &:= \lim_{p\to\infty} 2\tau - \tau_X - \tau_Y, \end{aligned} \tag{3.13}$$

where σ^2, σ_X^2 and σ_Y^2 are defined in Assumption 3.3.4. Under Assumption 3.3.5, further define

$$\begin{aligned} m_0^* := \lim_{m,n\to\infty} m_0 &= \frac{2\alpha_0\,\sigma^2 + \sigma_X^2 + \sigma_Y^2\,\alpha_0^2}{2\alpha_0 + 1 + \alpha_0^2}, \\ a_0^* := \lim_{m,n\to\infty} \frac{a_{nm}}{\sigma_{nm}} &= \Big(\frac{2\alpha_0 + \alpha_0^2 + 1}{2\alpha_0\,\sigma^2 + \alpha_0^2\,\sigma_X^2 + \sigma_Y^2}\Big)^{1/2}. \end{aligned}$$

We are now ready to introduce the two-sample t-test

$$T_{n,m} \;:=\; \frac{\mathcal{E}_{n,m}(X,Y)}{a_{nm}\,\sqrt{S_{n,m}}},$$

where

$$S_{n,m} \;:=\; \frac{4(n-1)(m-1)\,cdCov^2_{n,m}(X,Y) \;+\; 4v_n\,\widetilde{\mathcal{D}^2_n}(X,X) \;+\; 4v_m\,\widetilde{\mathcal{D}^2_n}(Y,Y)}{(n-1)(m-1) + v_n + v_m}$$

is the pool variance estimator with $\widetilde{\mathcal{D}^2_n}(X,X)$ and $\widetilde{\mathcal{D}^2_m}(Y,Y)$ being the unbiased estimators of the (squared) distance variances defined in equation (1.14). It is interesting to note that the variability of the sample generalized energy distance depends on the distance variances as well as the cdCov. It is also worth mentioning that the computational complexity of the pool variance estimator and thus the t-statistic is linear in p.

To study the asymptotic behavior of the test, we consider the following class of distributions on (X,Y):

$$\begin{aligned}\mathcal{P} =& \Big\{(P_X,P_Y):\ X \sim P_X,\ Y \sim P_Y,\ E[\tau L(X,Y) - \tau_X L(X,X')|X] = o_p(1),\\ & E[\tau L(X,Y) - \tau_Y L(Y,Y')|Y] = o_p(1)\Big\}.\end{aligned}$$

If $P_X = P_Y$ (i.e., under the H_0), it is clear that $(P_X,P_Y) \in \mathcal{P}$ irrespective of the metrics in the definition of L. Suppose $\|X-\mu_X\|^2 - \text{tr}(\Sigma_X) = O_p(\sqrt{p})$ and $\|Y-\mu_Y\|^2 - \text{tr}(\Sigma_Y) = O_p(\sqrt{p})$, which hold under weak dependence assumptions on the components of X and Y. Then in Case S2 (i.e., K is the Euclidean distance), a set of sufficient conditions for $(P_X,P_Y) \in \mathcal{P}$ is given by

$$(\mu_X-\mu_Y)^\top(\Sigma_X+\Sigma_Y)(\mu_X-\mu_Y) = o(p), \quad \tau-\tau_X = o(\sqrt{p}), \quad \tau-\tau_Y = o(\sqrt{p}), \tag{3.14}$$

which suggests that the first two moments of P_X and P_Y are not too far away from each other. In this sense, $\mathcal{P}$ defines a class of local alternative distributions (with respect to the null $H_0 : P_X = P_Y$). We now state the main result of this subsection.

THEOREM **2**. *In both Cases S1 and S2, under Assumptions 3.3.2, 3.3.3 and 3.3.4 as $p \to \infty$ with n and m remaining fixed, and further assuming that $(P_X,P_Y) \in \mathcal{P}$, we have*

$$\frac{\mathcal{E}_{n,m}(X,Y) - (2\tau-\tau_X-\tau_Y)}{a_{nm}\sqrt{S_{n,m}}} \xrightarrow{d} \frac{\sigma_{nm}\, Z}{a_{nm}\sqrt{M}},$$

where

$$M \overset{d}{=} \frac{\sigma^2\,\chi^2_{(n-1)(m-1)} + \sigma^2_X\chi^2_{v_n} + \sigma^2_Y\chi^2_{v_m}}{(n-1)(m-1)+v_n+v_m},$$

$\chi^2_{(n-1)(m-1)}$, $\chi^2_{v_n}$, $\chi^2_{v_m}$ *are independent chi-squared random variables, and* $Z \sim N(0,1)$. *In other words,*

$$T_{n,m} \xrightarrow{d} \frac{\sigma_{nm}\, N(\Delta/\sigma_{nm}, 1)}{a_{nm}\,\sqrt{M}}\,,$$

where σ_{nm} *and* a_{nm} *are defined in equation (3.13). In particular, under* H_0*, we have*

$$T_{n,m} \xrightarrow{d} t_{(n-1)(m-1)+v_n+v_m}.$$

Based on the asymptotic behavior of $T_{n,m}$ for growing dimensions, we propose a test for H_0 as follows: at level $\alpha \in (0,1)$, reject H_0 if $T_{n,m} > q_{\alpha,(n-1)(m-1)+v_n+v_m}$ and fail to reject H_0 otherwise, where $P(t_{(n-1)(m-1)+v_n+v_m} > q_{\alpha,(n-1)(m-1)+v_n+v_m}) = \alpha$. For a fixed real number t, define

$$\begin{aligned}\phi_{n,m}(t) \;:=\; \lim_{p\to\infty} P(T_{n,m} \le t) \;=\; \mathbb{E}\left[P\left(\frac{\sigma_{nm}\, N(\Delta/\sigma_{nm}, 1)}{a_{nm}\,\sqrt{M}} \le t \;\middle|\; M\right)\right] \\ = \mathbb{E}\left[\Phi\left(\frac{a_{nm}\,\sqrt{M}\, t - \Delta}{\sigma_{nm}}\right)\right].\end{aligned} \tag{3.15}$$

The asymptotic power curve for testing H_0 based on $T_{n,m}$ is given by $1 - \phi_{m,n}(t)$. The following proposition gives a large sample approximation of the power curve.

ASSUMPTION **3.3.5**. *As* $m, n \to \infty$, $m/n \to \alpha_0$ *where* $\alpha_0 > 0$.

PROPOSITION **3.3.2**. *Suppose* $\Delta = \Delta_0/\sqrt{nm}$ *where* Δ_0 *is a constant with respect to* n, m. *Then for any bounded real number* t *as* $n, m \to \infty$ *and under Assumption 3.3.5, we have*

$$\lim_{m,n\to\infty} \phi_{n,m}(t) \;=\; \Phi\left(a_0^*\sqrt{m_0^*}\; t \;-\; \Delta_0^*\right)\,,$$

where

$$\Delta_0^* = \Delta_0 \lim_{m,n\to\infty} \frac{1}{\sigma_{nm}\sqrt{nm}} = \Delta_0\left(\frac{2\alpha_0}{2\sigma^2\,\alpha_0 + \sigma_X^2\,\alpha_0^2 + \sigma_Y^2}\right)^{1/2}.$$

Under the alternative, if $\Delta_0 \to \infty$ as $n, m \to \infty$, we have

$$\lim_{m,n\to\infty} \left\{1 - \phi_{n,m}\big(q_{\alpha,(n-1)(m-1)+v_n+v_m}\big)\right\} \;=\; 1,$$

thereby justifying the consistency of the test.

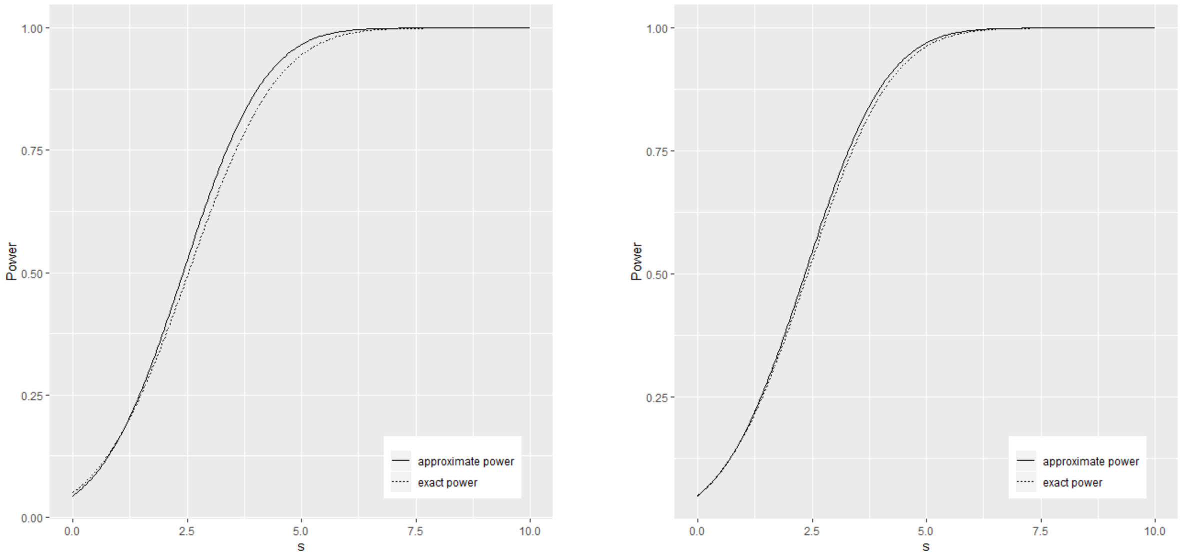
REMARK **3.3.5**. *We first derive the power function $1 - \phi_{n,m}(t)$ under the assumption that n and m are fixed. The main idea behind Proposition 3.3.2 where we let $n, m \to \infty$ is to see whether we get a reasonably good approximation of power when n, m are large. In a sense we are doing sequential asymptotics, first letting $p \to \infty$ and deriving the power function, and then deriving the leading term by letting $n, m \to \infty$. This is a quite common practice in Econometrics (see for example Phillips and Moon (1999)). The aim is to derive a leading term for the power when n, m are fixed but large. Consider $\Delta = s/\sqrt{nm}$ (as in Proposition 3.3.2) and set $\sigma^2 = \sigma_X^2 = \sigma_Y^2 = 1$. In Figure 3.1 below, we plot the exact power (computed from (3.15) with $50,000$ Monte Carlo samples from the distribution of M) with $n = m = 5$ and 10, $t = q_{\alpha,(n-1)(m-1)+v_n+v_m}$ and $\alpha = 0.05$, over different values of s. We overlay the large sample approximation of the power function (given in Proposition 3.3.2) and observe that the approximation works reasonably well even for small sample sizes. Clearly larger s results in better power and $s = 0$ corresponds to trivial power.*

We now discuss the power behavior of $T_{n,m}$ based on the Euclidean energy distance. In Case S2, it can be seen that

$$\sigma_X^2 = \lim_{p\to\infty} \frac{1}{\tau_X^2} \sum_{i,i'=1}^{p} 4\,\mathrm{tr}\,\Sigma_X^2(i,i'), \tag{3.16}$$

where $\Sigma_X^2(i,i')$ is the covariance matrix between $X_{(i)}$ and $X_{(i')}$, and similar expressions for σ_Y^2. In case S2 (i.e., when K is the Euclidean distance), if we further assume $\mu_X = \mu_Y$, it can be verified that

$$\sigma^2 \;=\; \lim_{p\to\infty} \frac{1}{\tau^2} \sum_{i,i'=1}^{p} 4\,\mathrm{tr}\big(\Sigma_X(i,i')\,\Sigma_Y(i,i')\big)\,. \tag{3.17}$$



(a) Power comparison when $m = n = 5$ (b) Power comparison when $m = n = 10$

Figure 3.1: Comparison of exact and approximate power.

Hence in Case S2, under the assumptions that $\mu_X = \mu_Y$, $\operatorname{tr}\Sigma_X = \operatorname{tr}\Sigma_Y$ and $\operatorname{tr}\Sigma_X^2 = \operatorname{tr}\Sigma_Y^2 = \operatorname{tr}\Sigma_X\Sigma_Y$, it can be easily seen from equations (3.8), (3.16) and (3.17) that

$$\tau_X^2 = \tau_Y^2 = \tau^2, \quad \sigma_X^2 = \sigma_Y^2 = \sigma^2, \tag{3.18}$$

which implies that $\Delta_0^* = 0$ in Proposition 3.3.2. Consider the following class of alternative distributions

$$H_A = \{(P_X, P_Y) : P_X \neq P_Y,\ \mu_X = \mu_Y,\ \operatorname{tr}\Sigma_X = \operatorname{tr}\Sigma_Y,\ \operatorname{tr}\Sigma_X^2 = \operatorname{tr}\Sigma_Y^2 = \operatorname{tr}\Sigma_X\Sigma_Y\}.$$

According to Theorem 2, the t-test $T_{n,m}$ based on Euclidean energy distance has trivial power against H_A. In contrast, the t-test based on the proposed metrics has non-trivial power against H_A as long as $\Delta_0^* > 0$.

To summarize our contributions :

- We show that the Euclidean energy distance can only detect the equality of means and the traces of covariance matrices in the high-dimensional setup. To the best of our knowledge, such a limitation of the Euclidean energy distance has not been pointed out in the literature before.

- We propose a new class of homogeneity metrics which completely characterizes homogeneity of two distributions in the low-dimensional setup and has nontrivial power against a broader range of alternatives, or in other words, can detect a wider range of inhomogeneity of two distributions in the high-dimensional setup.

- Grouping allows us to detect homogeneity beyond univariate marginal distributions, as the difference between two univariate marginal distributions is automatically captured by the difference between the marginal distributions of the groups that contain these two univariate components.

- Consequently we construct a high-dimensional two-sample t-test whose computational cost is linear in p. Owing to the pivotal nature of the limiting distribution of the test statistic, no resampling-based inference is needed.

REMARK **3.3.6**. *Although the test based on our proposed statistic is asymptotically powerful against the alternative H_A unlike the Euclidean energy distance, it can be verified that it has trivial power against the alternative $H_{A'} = \{(X,Y) : X_{(i)} \overset{d}{=} Y_{(i)}, 1 \leq i \leq p\}$. Thus although it can detect differences between two high-dimensional distributions beyond the first two moments (as a significant improvement to the Euclidean energy distance), it cannot capture differences beyond the equality of the low-dimensional marginal distributions. We conjecture that there might be some intrinsic difficulties for distance and kernel-based metrics to completely characterize the discrepancy between two high-dimensional distributions.*

3.4 Dependence metrics

In this section, we focus on dependence testing of two random vectors $X \in \mathbb{R}^{\tilde{p}}$ and $Y \in \mathbb{R}^{\tilde{q}}$. Suppose X and Y can be partitioned into p and q groups, viz. $X = \big(X_{(1)}, X_{(2)}, \ldots, X_{(p)}\big)$ and $Y =$

$\big(Y_{(1)}, Y_{(2)}, \ldots, Y_{(q)}\big)$, where the components $X_{(i)}$ and $Y_{(j)}$ are d_i and g_j dimensional, respectively, for $1 \leq i \leq p, 1 \leq j \leq q$. Here p, q might be fixed or growing. We assume that $X_{(i)}$ and $Y_{(j)}$'s are finite (low) dimensional vectors, i.e., $\{d_i\}_{i=1}^p$ and $\{g_j\}_{j=1}^q$ are bounded sequences. Clearly, $\tilde{p} = \sum_{i=1}^p d_i = O(p)$ and $\tilde{q} = \sum_{j=1}^q g_j = O(q)$. We define a class of dependence metrics $\mathcal{D}$ between X and Y as the positive square root of

$$\mathcal{D}^2(X,Y) \;:=\; \mathbb{E}\, K_{\mathbf{d}}(X,X')\, K_{\mathbf{g}}(Y,Y') + \mathbb{E}\, K_{\mathbf{d}}(X,X')\, \mathbb{E}\, K_{\mathbf{g}}(Y,Y') - 2\, \mathbb{E}\, K_{\mathbf{d}}(X,X')\, K_{\mathbf{g}}(Y,Y'')\,, \tag{3.19}$$

where $\mathbf{d} = (d_1, \ldots, d_p)$ and $\mathbf{g} = (g_1, \ldots, g_q)$. We drop the subscripts $\mathbf{d}, \mathbf{g}$ of K for notational convenience.

To ensure the existence of $\mathcal{D}$, we make the following assumption.

ASSUMPTION **3.4.1**. *Assume that* $\sup_{1 \leq i \leq p} \mathbb{E}\rho_i^{1/2}(X_{(i)}, 0_{d_i}) < \infty$ *and* $\sup_{1 \leq i \leq q} \mathbb{E}\rho_i^{1/2}(Y_{(i)}, 0_{g_i}) < \infty$.

In Section B.1.2 of the appendix we demonstrate that in the low-dimensional setting, $\mathcal{D}(X,Y)$ completely characterizes independence between X and Y. For an observed random sample $(X_k, Y_k)_{k=1}^n$ from the joint distribution of X and Y, define $D^X := (d_{kl}^X) \in \mathbb{R}^{n \times n}$ with $d_{kl}^X := K(X_k, X_l)$ and $k, l \in \{1, \ldots, n\}$. Define d_{kl}^Y and D^Y in a similar way. With some abuse of notation, we consider the U-statistic type estimator $\widetilde{\mathcal{D}_n^2}(X,Y)$ of $\mathcal{D}^2$ as defined in (1.14) with $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ being $K_{\mathbf{d}}$ and $K_{\mathbf{g}}$ respectively. In Section B.1.2 of the appendix, we illustrate that $\widetilde{\mathcal{D}_n^2}(X,Y)$ essentially inherits all the nice properties of the U-statistic type estimator of generalized dCov and HSIC.

In the subsequent discussion we study the asymptotic behavior of $\mathcal{D}$ in the high-dimensional framework, i.e., when p and q grow to ∞ with fixed n (discussed in Subsection 3.4.1) and when n grows to ∞ as well (discussed in Subsection B.2.2 in the appendix).

3.4.1 High dimension low sample size (HDLSS)

In this subsection, our goal is to explore the behavior of $\mathcal{D}^2(X,Y)$ and its unbiased U-statistic type estimator in the HDLSS setting where p and q grow to ∞ while the sample size n is held fixed. Denote $\tau_{XY}^2 = \tau_X^2 \tau_Y^2 = \mathbb{E}\, K^2(X,X')\, \mathbb{E}\, K^2(Y,Y')$. We impose the following conditions.

Assumption **3.4.2**. *$\mathbb{E}\,[L^2(X,X')] = O(a_p'^2)$ and $\mathbb{E}\,[L^2(Y,Y')] = O(b_q'^2)$, where a_p' and b_q' are positive real sequences satisfying $a_p' = o(1)$, $b_q' = o(1)$, $\tau_{XY}\, a_p'^2 b_q' = o(1)$ and $\tau_{XY}\, a_p' b_q'^2 = o(1)$. Further assume that $\mathbb{E}\,[R^2(X,X')] = O(a_p'^4)$ and $\mathbb{E}\,[R^2(Y,Y')] = O(b_q'^4)$.*

Remark **3.4.1**. *We refer the reader to Remark 3.3.1 in Section 3.3 for illustrations about some sufficient conditions under which we have var $(L(X,X')) = \mathbb{E}\, L^2(X,X') = O(\frac{1}{p})$, and similarly for $L(Y,Y')$. Remark B.4.1 in the appendix illustrates certain sufficient conditions under which $\mathbb{E}\,[R^2(X,X')] = O(\frac{1}{p^2})$, and similarly for $R(Y,Y')$.*

Theorem **3**. *Under Assumptions 3.3.2 and 3.4.2, we have*

$$\mathcal{D}^2(X,Y) = \frac{1}{4\tau_{XY}} \sum_{i=1}^{p} \sum_{j=1}^{q} D^2_{\rho_i,\rho_j}(X_{(i)}, Y_{(j)}) \,+\, \mathcal{R}\,, \tag{3.20}$$

where $\mathcal{R}$ is the remainder term such that $\mathcal{R} = O(\tau_{XY}\, a_p'^2 b_q' + \tau_{XY}\, a_p' b_q'^2) = o(1)$.

Theorem 3 shows that when dimensions grow high, the population $\mathcal{D}^2(X,Y)$ behaves as an aggregation of group-wise generalized dCov and thus essentially captures group-wise non-linear dependencies between X and Y.

Remark **3.4.2**. *Consider a special case where $d_i = 1$ and $g_j = 1$, and ρ_i and ρ_j are Euclidean distances for all $1 \le i \le p$ and $1 \le j \le q$. Then Theorem 3 essentially boils down to*

$$\mathcal{D}^2(X,Y) = \frac{1}{4\tau_{XY}} \sum_{i=1}^{p} \sum_{j=1}^{q} dCov^2(X_i, Y_j) \,+\, \mathcal{R}\,, \tag{3.21}$$

where $\mathcal{R} = o(1)$. This shows that in a special case (when we have unit group sizes), $\mathcal{D}^2(X,Y)$ essentially behaves as an aggregation of cross-component dCov between X and Y. If $K_{\boldsymbol{d}}$ and $K_{\boldsymbol{g}}$ are Euclidean distances, or in other words if each ρ_i and ρ_j are squared Euclidean distances, then using equation (1.10) it is straightforward to verify that $D^2_{\rho_i,\rho_j}(X_i, Y_j) = 4\, cov^2(X_i, Y_j)$ for all

$1 \leq i \leq p$ and $1 \leq j \leq q$. Consequently we have

$$\mathcal{D}^2(X,Y) = dCov^2(X,Y) = \frac{1}{\tau_{XY}} \sum_{i=1}^{p} \sum_{j=1}^{q} cov^2(X_i, Y_j) \, + \, \mathcal{R}_1 \, , \tag{3.22}$$

where $\mathcal{R}_1 = o(1)$, which essentially presents a population version of Theorem 2.1.1 in Zhu et al. (2020) as a special case of Theorem 3.

REMARK **3.4.3**. *To illustrate that the leading term in equation (3.20) indeed gives a close approximation of the population $\mathcal{D}^2(X,Y)$, we consider the special case when $K_{\boldsymbol{d}}$ and $K_{\mathbf{g}}$ are Euclidean distances and $p = q$. Suppose $X \sim N_p(0, I_p)$ and $Y = X + N$ where $N \sim N_p(0, I_p)$ with $N \perp\!\!\!\perp X$. Clearly we have $\tau_X^2 = 2p$, $\tau_Y^2 = 4p$, $D^2_{\rho_i,\rho_j}(X_i, Y_j) = 4\, cov^2(X_i, Y_j) = 4$ for all $1 \leq i = j \leq p$ and $D^2_{\rho_i,\rho_j}(X_i, Y_j) = 0$ for all $1 \leq i \neq j \leq p$. From Remark 3.4.2, it is clear that in this case we essentially have $\mathcal{D}^2(X,Y) = dCov^2(X,Y)$. We simulate a large sample of size $n = 5000$ from the distribution of (X,Y) for $p = 20, 40, 60, 80$ and 100. The large sample size is to ensure that the U-statistic type estimator of $\mathcal{D}^2(X,Y)$ (given in (1.14)) gives a very close approximation of the population $\mathcal{D}^2(X,Y)$. We list the ratio between $\mathcal{D}^2(X,Y)$ and the leading term in (3.20) for the different values of p, which turn out to be very close to 1, demonstrating that the leading term in (3.20) indeed approximates $\mathcal{D}^2(X,Y)$ reasonably well.*

Table 3.2: Ratio of $\mathcal{D}^2(X,Y)$ and the leading term in (3.20) for different values of p.

$p = 20$	$p = 40$	$p = 60$	$p = 80$	$p = 100$
0.980	0.993	0.994	0.989	0.997

The following theorem explores the behavior of the population $\mathcal{D}^2(X,Y)$ when p is fixed and q grows to infinity, while the sample size is held fixed. As far as we know, this asymptotic regime has not been previously considered in the literature. In this case, the Euclidean distance covariance behaves as an aggregation of martingale difference divergences proposed in Shao and Zhang

(2014) which measures conditional mean dependence. Figure 3.2 below summarizes the curse of dimensionality for the Euclidean distance covariance under different asymptotic regimes.

THEOREM **4**. *Under Assumption 3.3.2 and the assumption that* $\mathbb{E}\,[R^2(Y,Y')] = O(b_q'^4)$ *with* $\tau_Y\, b_q'^2 = o(1)$*, as* $q \to \infty$ *with* p *and* n *remaining fixed, we have*

$$\mathcal{D}^2(X,Y) \;=\; \frac{1}{2\tau_Y} \sum_{j=1}^{q} D^2_{K_{\boldsymbol{d}}\,,\rho_j}(X, Y_{(j)}) \;+\; \mathcal{R},$$

where $\mathcal{R}$ *is the remainder term such that* $\mathcal{R} = O(\tau_Y\, b_q'^2) = o(1)$.

REMARK **3.4.4**. *In particular, when both* $K_{\boldsymbol{d}}$ *and* $K_{\boldsymbol{g}}$ *are Euclidean distances, we have*

$$\mathcal{D}^2(X,Y) \;=\; dCov^2(X,Y) \;=\; \frac{1}{\tau_Y} \sum_{j=1}^{\tilde{q}} MDD^2(Y_j|X) \;+\; \mathcal{R},$$

where $MDD^2(Y_j|X) = -\mathbb{E}[(Y_j - \mathbb{E}Y_j)(Y_j' - \mathbb{E}Y_j)\|X - X'\|]$ *is the martingale difference divergence which completely characterizes the conditional mean dependence of* Y_j *given* X *in the sense that* $E[Y_j|X] = E[Y_j]$ *almost surely if and only if* $MDD^2(Y_j|X) = 0$.

Next we study the asymptotic behavior of the sample version $\widetilde{\mathcal{D}}_n^2(X,Y)$.

ASSUMPTION **3.4.3**. *Assume that* $L(X,X') = O_p(a_p)$ *and* $L(Y,Y') = O_p(b_q)$*, where* a_p *and* b_q *are positive real sequences satisfying* $a_p = o(1)$*,* $b_q = o(1)$*,* $\tau_{XY}\, a_p^2 b_q = o(1)$ *and* $\tau_{XY}\, a_p b_q^2 = o(1)$.

REMARK **3.4.5**. *We refer the reader to Remark 3.3.1 in Section 3.3 for illustrations about Assumption 3.4.3.*

THEOREM **5**. *Under Assumptions 3.3.2 and 3.4.3, it can be shown that*

$$\widetilde{\mathcal{D}}_n^2(X,Y) = \frac{1}{4\tau_{XY}} \sum_{i=1}^{p} \sum_{j=1}^{q} \widetilde{D}^2_{n\,;\,\rho_i,\rho_j}(X_{(i)}, Y_{(j)}) \;+\; \mathcal{R}_n\;, \tag{3.23}$$

where $X_{(i)}, Y_{(j)}$ *are the* i^{th} *and* j^{th} *groups of* X *and* Y*, respectively,* $1 \le i \le p$*,* $1 \le j \le q$ *, and* $\mathcal{R}_n$ *is the remainder term. Moreover* $\mathcal{R}_n = O_p(\tau_{XY}\, a_p^2 b_q + \tau_{XY}\, a_p b_q^2) = o_p(1)$*, i.e.,* $\mathcal{R}_n$ *is of smaller order compared to the leading term and hence is asymptotically negligible.*

Figure 3.2: Curse of dimensionality for the Euclidean distance covariance under different asymptotic regimes

$$\begin{tikzcd}
\textit{Arbitrary dependence} & \textit{Conditional mean dependence} \\
dCov^2(X,Y) \arrow[r, "\tilde{q} \to +\infty"] \arrow[d, "\tilde{p} \to +\infty"'] \arrow[dr, "\tilde{p}, \tilde{q} \to +\infty"] & \sum_{j=1}^{\tilde{q}} MDD^2(Y_j|X) \arrow[d, "\tilde{p} \to +\infty"] \\
\sum_{i=1}^{\tilde{p}} MDD^2(X_i|Y) \arrow[r, "\tilde{q} \to +\infty"'] & \sum_{i=1}^{\tilde{p}} \sum_{j=1}^{\tilde{q}} cov^2(X_i, Y_j) \\
\textit{Conditional mean dependence} & \textit{Linear dependence}
\end{tikzcd}$$

The above theorem generalizes Theorem 2.1.1 in Zhu et al. (2020) by showing that the leading term of $\widetilde{\mathcal{D}_n^2}(X,Y)$ is the sum of all the group-wise (unbiased) squared sample generalized dCov scaled by τ_{XY} . In other words, in the HDLSS setting, $\widetilde{\mathcal{D}_n^2}(X,Y)$ is asymptotically equivalent to the aggregation of group-wise squared sample generalized dCov. Thus $\widetilde{\mathcal{D}_n^2}(X,Y)$ can quantify group-wise non-linear dependencies between X and Y, going beyond the scope of the usual Euclidean dCov.

Remark **3.4.6**. *Consider a special case where $d_i = 1$ and $g_j = 1$, and ρ_i and ρ_j are Euclidean distances for all $1 \leq i \leq p$ and $1 \leq j \leq q$. Then Theorem 5 essentially states that*

$$\widetilde{\mathcal{D}_n^2}(X,Y) = \frac{1}{4\tau_{XY}} \sum_{i=1}^{p} \sum_{j=1}^{q} dCov_n^2(X_i, Y_j) \, + \, \mathcal{R}_n \, , \tag{3.24}$$

where $\mathcal{R}_n = o_p(1)$. This demonstrates that in a special case (when we have unit group sizes), $\widetilde{\mathcal{D}_n^2}(X,Y)$ is asymptotically equivalent to the marginal aggregation of cross-component distance covariances proposed by Zhu et al. (2020) as dimensions grow high. If $K_{\boldsymbol{d}}$ and $K_{\boldsymbol{g}}$ are Euclidean distances, then Theorem 5 essentially boils down to Theorem 2.1.1 in Zhu et al. (2020) as a special *case.*

REMARK **3.4.7**. *To illustrate the approximation of $\widetilde{\mathcal{D}_n^2}(X,Y)$ by the aggregation of group-wise squared sample generalized dCov given by Theorem 5, we simulated the datasets in Examples 4.4.2.1, 4.4.2.2, 4.4.3.1 and 4.4.3.2* 100 *times each with $n = 50$ and $p = q = 50$. For each of the datasets, the difference between $\widetilde{\mathcal{D}_n^2}(X,Y)$ and the leading term in the RHS of equation (3.23) is smaller than* 0.01 100% *of the times, which illustrates that the approximation works reasonably well.*

The following theorem illustrates the asymptotic behavior of $\widetilde{\mathcal{D}_n^2}(X,Y)$ when p is fixed and q grows to infinity while the sample size is held fixed. Under this setup, if both $K_{\mathbf{d}}$ and $K_{\mathbf{g}}$ are Euclidean distances, the leading term of $\widetilde{\mathcal{D}_n^2}(X,Y)$ is the sum of the group-wise unbiased U-statistic type estimators of $MDD^2(Y_j|X)$ for $1 \leq j \leq q$, scaled by τ_Y. In other words, the sample Euclidean distance covariance behaves as an aggregation of sample martingale difference divergences.

THEOREM **6**. *Under Assumption 3.3.2 and the assumption that $L(Y,Y') = O_p(b_q)$ with $b_q = o(1)$ and $\tau_Y\, b_q^2 = o(1)$, as $q \to \infty$ with p and n remaining fixed, we have*

$$\widetilde{\mathcal{D}_n^2}(X,Y) \;=\; \frac{1}{2\tau_Y}\sum_{j=1}^{q}\widetilde{\mathcal{D}^2_{n\,;\,K_{\boldsymbol{d}},\rho_j}}(X,Y_{(j)}) \,+\, \mathcal{R}_n,$$

where $\mathcal{R}_n$ is the remainder term such that $\mathcal{R}_n = O_p(\tau_Y\, b_q^2) = o_p(1)$.

REMARK **3.4.8**. *In particular, when both $K_{\boldsymbol{d}}$ and $K_{\boldsymbol{g}}$ are Euclidean distances, we have*

$$\widetilde{\mathcal{D}_n^2}(X,Y) \;=\; dCov_n^2(X,Y) \;=\; \frac{1}{\tau_Y}\sum_{j=1}^{\tilde{q}} MDD_n^2(Y_j|X) \,+\, \mathcal{R}_n,$$

where $MDD_n^2(Y_j|X)$ is the unbiased U-statistic type estimator of $MDD^2(Y_j|X)$ defined as in (1.14) with $d_{\mathcal{X}}(x,x') = \|x - x'\|$ for $x,x' \in \mathbb{R}^{\tilde{p}}$ and $d_{\mathcal{Y}}(y,y') = |y - y'|^2/2$ for $y,y' \in \mathbb{R}$, respectively.

Now denote $X_k = (X_{k(1)}, \ldots, X_{k(p)})$ and $Y_k = (Y_{k(1)}, \ldots, Y_{k(q)})$ for $1 \leq k \leq n$. Define the leading term of $\widetilde{\mathcal{D}_n^2}(X, Y)$ in equation (3.23) as

$$L := \frac{1}{4\tau_{XY}} \sum_{i=1}^{p} \sum_{j=1}^{q} \widetilde{D_n^2}_{\,;\,\rho_i,\rho_j}(X_{(i)}, Y_{(j)})\,.$$

It can be verified that

$$L \;=\; \frac{1}{4\tau_{XY}} \sum_{i=1}^{p} \sum_{j=1}^{q} \left(\tilde{D}^X(i) \cdot \tilde{D}^Y(j) \right)\,,$$

where $\tilde{D}^X(i), \tilde{D}^Y(j)$ are the $\mathcal{U}$-centered versions of $D^X(i) = \big(d_{kl}^X(i)\big)_{k,l=1}^n$ and $D^Y(j) = \big(d_{kl}^Y(j)\big)_{k,l=1}^n$, respectively. As an advantage of using the double-centered distances, we have for all $1 \leq i, i' \leq p$, $1 \leq j, j' \leq q$ and $\{k, l\} \neq \{u, v\}$,

$$\mathbb{E}\left[d_{kl}^X(i)\, d_{uv}^X(i')\right] \;=\; \mathbb{E}\left[d_{kl}^Y(j)\, d_{uv}^Y(j')\right] \;=\; \mathbb{E}\left[d_{kl}^X(i)\, d_{uv}^Y(j)\right] \;=\; 0. \qquad (3.25)$$

See for example the proof of Proposition 2.2.1 in Zhu et al. (2020) for a detailed explanation.

ASSUMPTION **3.4.4**. *For fixed n, as $p, q \to \infty$,*

$$\begin{pmatrix} \frac{1}{2\,\tau_X} \sum_{i=1}^{p} d_{kl}^X(i) \\ \frac{1}{2\,\tau_Y} \sum_{j=1}^{q} d_{uv}^Y(j) \end{pmatrix}_{k<l,\,u<v} \xrightarrow{d} \begin{pmatrix} d_{kl}^1 \\ \\ d_{uv}^2 \end{pmatrix}_{k<l,\,u<v},$$

where $\{d_{kl}^1,\ d_{uv}^2\}_{k<l,\,u<v}$ are jointly Gaussian. Further we assume that

$$var(d_{kl}^1) := \sigma_X^2 = \lim_{p\to\infty} \frac{1}{4\tau_X^2} \sum_{i,i'=1}^{p} D_{\rho_i,\rho_{i'}}^2\left(X_{(i)}, X_{(i')}\right),$$

$$var(d_{kl}^2) := \sigma_Y^2 = \lim_{q\to\infty} \frac{1}{4\tau_Y^2} \sum_{j,j'=1}^{q} D_{\rho_j,\rho_{j'}}^2\left(Y_{(j)}, Y_{(j')}\right),$$

$$cov\,(d_{kl}^1, d_{kl}^2) := \sigma_{XY}^2 = \lim_{p,q\to\infty} \frac{1}{4\tau_{XY}} \sum_{i=1}^{p} \sum_{j=1}^{q} D_{\rho_i,\rho_j}^2\left(X_{(i)}, Y_{(j)}\right).$$

In view of (3.25), we have $\operatorname{cov}(d^1_{kl}, d^1_{uv}) = \operatorname{cov}(d^2_{kl}, d^2_{uv}) = \operatorname{cov}(d^1_{kl}, d^2_{uv}) = 0$ for $\{k,l\} \neq \{u,v\}$. Theorem 5 states that for growing p and q and fixed n, $\widetilde{\mathcal{D}}^2_n(X,Y)$ and L are asymptotically equivalent. By studying the leading term, we obtain the limiting distribution of $\widetilde{\mathcal{D}}^2_n(X,Y)$ as follows.

THEOREM **7**. *Under Assumptions 3.3.2, 3.4.3 and 3.4.4, for fixed n and $p, q \to \infty$,*

$$\begin{aligned}
\widetilde{\mathcal{D}}^2_n(X,Y) &\xrightarrow{d} \frac{1}{\nu} d^{1\top} M d^2 \,, \\
\widetilde{\mathcal{D}}^2_n(X,X) &\xrightarrow{d} \frac{1}{\nu} d^{1\top} M d^1 \overset{d}{=} \frac{\sigma_X^2}{\nu}\chi^2_\nu \,, \\
\widetilde{\mathcal{D}}^2_n(Y,Y) &\xrightarrow{d} \frac{1}{\nu} d^{2\top} M d^2 \overset{d}{=} \frac{\sigma_Y^2}{\nu}\chi^2_\nu \,,
\end{aligned}$$

where M is a projection matrix of rank $\nu = \frac{n(n-3)}{2}$, and

$$\begin{pmatrix} d^1 \\ \\ d^2 \end{pmatrix} \sim N\left(0\,, \begin{pmatrix} \sigma_X^2\, I_{\frac{n(n-1)}{2}} & \sigma_{XY}^2\, I_{\frac{n(n-1)}{2}} \\ \\ \sigma_{XY}^2\, I_{\frac{n(n-1)}{2}} & \sigma_Y^2\, I_{\frac{n(n-1)}{2}} \end{pmatrix}\right) .$$

To perform independence testing, in the spirit of Székely and Rizzo (2014), we define the studentized test statistic

$$\mathcal{T}_n := \sqrt{\nu - 1}\, \frac{\widetilde{\mathcal{DC}}^2_n(X,Y)}{\sqrt{1 - \left(\widetilde{\mathcal{DC}}^2_n(X,Y)\right)^2}} \,, \tag{3.26}$$

where

$$\widetilde{\mathcal{DC}}^2_n(X,Y) = \frac{\widetilde{\mathcal{D}}^2_n(X,Y)}{\sqrt{\widetilde{\mathcal{D}}^2_n(X,X)\, \widetilde{\mathcal{D}}^2_n(Y,Y)}} \,.$$

Define $\psi = \sigma_{XY}^2 / \sqrt{\sigma_X^2 \sigma_Y^2}$. The following theorem states the asymptotic distributions of the test statistic $\mathcal{T}_n$ under the null hypothesis $\tilde{H}_0 : X \perp\!\!\!\perp Y$ and the alternative hypothesis $\tilde{H}_A : X \not\!\perp\!\!\!\perp Y$.

THEOREM **8**. *Under Assumptions 3.3.2, 3.4.3 and 3.4.4, for fixed n and $p, q \to \infty$,*

$$\begin{aligned} P_{\tilde{H}_0}\left(\mathcal{T}_n \leq t\right) &\longrightarrow P\left(t_{\nu-1} \leq t\right), \\ P_{\tilde{H}_A}\left(\mathcal{T}_n \leq t\right) &\longrightarrow \mathbb{E}\left[P\left(t_{\nu-1,W} \leq t | W\right)\right], \end{aligned}$$

where t is any fixed real number and $W \sim \sqrt{\frac{\psi^2}{1-\psi^2}\,\chi^2_\nu}$.

For an explicit form of $\mathbb{E}\left[P\left(t_{\nu-1,W} \leq t|W\right)\right]$, we refer the reader to Lemma 3 in the appendix of Zhu et al. (2020). Now consider the local alternative hypothesis $\tilde{H}_A^*$: $X \not\perp\!\!\!\perp Y$ with $\psi = \psi_0/\sqrt{\nu}$, where ψ_0 is a constant with respect to n. The following proposition gives an approximation of $\mathbb{E}\left[P\left(t_{\nu-1,W} \leq t|W\right)\right]$ under the local alternative hypothesis $\tilde{H}_A^*$ when n is allowed to grow.

PROPOSITION **3.4.1**. *Under $\tilde{H}_A^*$, as $n \to \infty$ and $t = O(1)$,*

$$\mathbb{E}\left[P\left(t_{\nu-1,W} \leq t|W\right)\right] \;=\; P\left(t_{\nu-1,\,\psi_0} \leq t\right) \;+\; O\Big(\frac{1}{\nu}\Big)\,.$$

The following summarizes our key findings in this section.

- **Advantages of our proposed metrics over the Euclidean dCov and HSIC :**

  i) Our proposed dependence metrics completely characterize independence between X and Y in the low-dimensional setup, and can detect group-wise non-linear dependencies between X and Y in the high-dimensional setup as opposed to merely detecting component-wise linear dependencies by the Euclidean dCov and HSIC (in light of Theorem 2.1.1 in Zhu et al. (2020)).

  ii) We also showed that with p remaining fixed and q growing high, the Euclidean dCov can only quantify conditional mean independence of the components of Y given X (which is weaker than independence). To the best of our knowledge, this has not been pointed out in the literature before.

- **Advantages over the marginal aggregation approach by Zhu et al. (2020) :**

  i) In the low-dimensional setup, our proposed dependence metrics can completely characterize independence between X and Y, whereas the metric proposed by Zhu et al. (2020) can only capture pairwise dependencies between the components of X and Y.

  ii) We provide a neater way of generalizing dCov and HSIC between X and Y which is shown to be asymptotically equivalent to the marginal aggregation of cross-component distance covariances proposed by Zhu et al. (2020) as dimensions grow high. Also grouping or partitioning the two high-dimensional random vectors (which again may be problem specific) allows us to detect a wider range of alternatives compared to only detecting component-wise non-linear dependencies, as independence of two univariate marginals is implied from independence of two higher dimensional marginals containing the two univariate marginals.

  iii) The computational complexity of the (unbiased) squared sample $\mathcal{D}(X, Y)$ is $O(n^2(p+q))$. Thus the computational cost of our proposed two-sample t-test only grows linearly with the dimension and therefore is scalable to very high-dimensional data. Although a naive aggregation of marginal distance covariances has a computational complexity of $O(n^2pq)$, the approach of Zhu et al. (2020) essentially corresponds to the use of an additive kernel and the computational cost of their proposed estimator can also be made linear in the dimensions if properly implemented.

3.5 Numerical studies

3.5.1 Testing for homogeneity of distributions

We investigate the empirical size and power of the tests for homogeneity of two high dimensional distributions. For comparison, we consider the t-tests based on the following metrics:

I. $\mathcal{E}$ with ρ_i as the Euclidean distance for $1 \leq i \leq p$;

Table 3.3: Summary of the behaviors of the proposed homogeneity/dependence metrics for different choices of $\rho_i(x, x')$ in high dimension.

Choice of $\rho_i(x, x')$	Asymptotic behavior of the proposed homogeneity metric	Asymptotic behavior of the proposed dependence metric
the semi-metric $\|x - x'\|^2$	Behaves as a sum of squared Euclidean distances	Behaves as a sum of squared Pearson correlations
metric of strong negative type on $\mathbb{R}^{d_i}$	Behaves as a sum of groupwise energy distances with the metric ρ_i	Behaves as a sum of groupwise dCov with the metric ρ_i
$k_i(x, x) + k_i(x', x') - 2k_i(x, x')$, where k_i is a characteristic kernel on $\mathbb{R}^{d_i} \times \mathbb{R}^{d_i}$	Behaves as a sum of groupwise MMD with the kernel k_i	Behaves as a sum of groupwise HSIC with the kernel k_i

II. $\mathcal{E}$ with ρ_i as the distance induced by the Laplace kernel for $1 \le i \le p$;

III. $\mathcal{E}$ with ρ_i as the distance induced by the Gaussian kernel for $1 \le i \le p$;

IV. the usual Euclidean energy distance;

V. MMD with the Laplace kernel;

VI. MMD with the Gaussian kernel.

We set $d_i = 1$ in Examples 3.5.1 and 3.5.2, and $d_i = 2$ in Example 3.5.3 for $1 \le i \le p$.

EXAMPLE **3.5.1**. *Consider $X_k = (X_{k1}, \ldots, X_{kp})$ and $Y_l = (Y_{l1}, \ldots, Y_{lp})$ with $k = 1, \ldots, n$ and $l = 1, \ldots, m$. We generate i.i.d. samples from the following models:*

1. *$X_k \sim N(0, I_p)$ and $Y_l \sim N(0, I_p)$.*
2. *$X_k \sim N(0, \Sigma)$ and $Y_l \sim N(0, \Sigma)$, where $\Sigma = (\sigma_{ij})_{i,j=1}^p$ with $\sigma_{ii} = 1$ for $i = 1, \ldots, p$, $\sigma_{ij} = 0.25$ if $1 \le |i - j| \le 2$ and $\sigma_{ij} = 0$ otherwise.*
3. *$X_k \sim N(0, \Sigma)$ and $Y_l \sim N(0, \Sigma)$, where $\Sigma = (\sigma_{ij})_{i,j=1}^p$ with $\sigma_{ij} = 0.7^{|i-j|}$.*

EXAMPLE **3.5.2**. *Consider $X_k = (X_{k1}, \ldots, X_{kp})$ and $Y_l = (Y_{l1}, \ldots, Y_{lp})$ with $k = 1, \ldots, n$ and $l = 1, \ldots, m$. We generate i.i.d. samples from the following models:*

1. $X_k \sim N(\mu, I_p)$ *with* $\mu = (1, \ldots, 1) \in \mathbb{R}^p$ *and* $Y_{li} \overset{ind}{\sim}$ *Poisson* (1) *for* $i = 1, \ldots, p$.

2. $X_k \sim N(\mu, I_p)$ *with* $\mu = (1, \ldots, 1) \in \mathbb{R}^p$ *and* $Y_{li} \overset{ind}{\sim}$ *Exponential* (1) *for* $i = 1, \ldots, p$.

3. $X_k \sim N(0, I_p)$ *and* $Y_l = (Y_{l1}, \ldots, Y_{l\lfloor \beta p \rfloor}, Y_{l(\lfloor \beta p \rfloor + 1)}, \ldots, Y_{lp})$, *where* $Y_{l1}, \ldots, Y_{l\lfloor \beta p \rfloor} \overset{i.i.d.}{\sim}$ *Rademacher* (0.5) *and* $Y_{l(\lfloor \beta p \rfloor + 1)}, \ldots, Y_{lp} \overset{i.i.d.}{\sim} N(0, 1)$.

4. $X_k \sim N(0, I_p)$ *and* $Y_l = (Y_{l1}, \ldots, Y_{l\lfloor \beta p \rfloor}, Y_{l(\lfloor \beta p \rfloor + 1)}, \ldots, Y_{lp})$, *where* $Y_{l1}, \ldots, Y_{l\lfloor \beta p \rfloor} \overset{i.i.d.}{\sim}$ *Uniform* $(-\sqrt{3}, \sqrt{3})$ *and* $Y_{l(\lfloor \beta p \rfloor + 1)}, \ldots, Y_{lp} \overset{i.i.d.}{\sim} N(0, 1)$.

5. $X_k = R^{1/2} Z_{1k}$ *and* $Y_l = R^{1/2} Z_{2l}$, *where* $R = (r_{ij})_{i,j=1}^p$ *with* $r_{ii} = 1$ *for* $i = 1, \ldots, p$, $r_{ij} = 0.25$ *if* $1 \le |i - j| \le 2$ *and* $r_{ij} = 0$ *otherwise,* $Z_{1k} \sim N(0, I_p)$ *and* $Z_{2l} = \underbrace{(Z_{2l1}, \ldots, Z_{2lp})}_{\overset{i.i.d.}{\sim} Exponential(1)} - 1$.

EXAMPLE **3.5.3**. *Consider* $X_k = (X_{k(1)}, \ldots, X_{k(p)})$ *and* $Y_l = (Y_{l(1)}, \ldots, Y_{l(p)})$ *with* $k = 1, \ldots, n$ *and* $l = 1, \ldots, m$ *and* $d_i = 2$ *for* $1 \le i \le p$. *We generate i.i.d. samples from the following models:*

1. $X_{k(i)} \sim N(\mu, \Sigma_1)$ *and* $Y_{l(i)} \sim N(\mu, \Sigma_2)$ *with* $X_{k(i)} \perp\!\!\!\perp X_{k(j)}$ *and* $Y_{l(i)} \perp\!\!\!\perp Y_{l(j)}$ *for* $1 \le i \ne j \le p$, *where* $\mu = (1, 1)^\top$, $\Sigma_1 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$ *and* $\Sigma_2 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$.

2. $X_{k(i)} \sim N(\mu, \Sigma)$ *with* $X_{k(i)} \perp\!\!\!\perp X_{k(j)}$ *for* $1 \le i \ne j \le p$, *where* $\mu = (1, 1)^\top$, $\Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$. *The components of* Y_l *are i.i.d. Exponential* (1).

Note that for Examples 3.5.1 and 3.5.2, the metric defined in equation (3.2) essentially boils down to the special case in equation (4.7). We try small sample sizes $n = m = 50$, dimensions $p = q = 50, 100$ and 200, and $\beta = 1/2$. Table 3.4 reports the proportion of rejections out of 1000 simulation runs for the different tests. For the tests V and VI, we chose the bandwidth parameter heuristically as the median distance between the aggregated sample observations. For tests II and III, the bandwidth parameters are chosen using the median heuristic separately for each group.

In Example 3.5.1, the data generating scheme suggests that the variables X and Y are identically distributed. The results in Table 3.4 show that the tests based on both the proposed homogeneity metrics and the usual Euclidean energy distance and MMD perform more or less equally good, and the rejection probabilities are quite close to the 10% or 5% nominal level. In Example 3.5.2, clearly X and Y have different distributions but $\mu_X = \mu_Y$ and $\Sigma_X = \Sigma_Y$. The results in Table 3.4 indicate that the tests based on the proposed homogeneity metrics are able to detect the differences between the two high-dimensional distributions beyond the first two moments unlike the tests based on the usual Euclidean energy distance and MMD, and thereby outperform the latter in terms of empirical power. In Example 3.5.3, clearly $\mu_X = \mu_Y$ and $\operatorname{tr}\Sigma_X = \operatorname{tr}\Sigma_Y$ and the results show that the tests based on the proposed homogeneity metrics are able to detect the in-homogeneity of the low-dimensional marginal distributions unlike the tests based on the usual Euclidean energy distance and MMD.

Table 3.4: Empirical size and power for the different tests of homogeneity of distributions.

			I		II		III		IV		V		VI	
		p	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
Ex 3.5.1	(1)	50	0.109	0.062	0.109	0.058	0.106	0.063	0.109	0.068	0.110	0.069	0.109	0.070
	(1)	100	0.124	0.073	0.119	0.053	0.121	0.063	0.116	0.067	0.114	0.068	0.117	0.068
	(1)	200	0.086	0.043	0.099	0.048	0.088	0.035	0.090	0.045	0.086	0.043	0.090	0.045
	(2)	50	0.114	0.069	0.108	0.054	0.118	0.068	0.116	0.077	0.115	0.073	0.116	0.078
	(2)	100	0.130	0.069	0.133	0.073	0.124	0.070	0.126	0.067	0.123	0.068	0.124	0.067
	(2)	200	0.099	0.048	0.103	0.041	0.092	0.047	0.097	0.040	0.095	0.039	0.097	0.040
	(3)	50	0.100	0.064	0.107	0.057	0.099	0.060	0.112	0.072	0.105	0.067	0.110	0.073
	(3)	100	0.103	0.062	0.113	0.061	0.113	0.063	0.097	0.060	0.100	0.057	0.098	0.059
	(3)	200	0.108	0.062	0.115	0.062	0.117	0.064	0.091	0.055	0.093	0.056	0.090	0.055
Ex 3.5.2	(1)	50	1	1	1	1	0.995	0.994	0.102	0.067	0.111	0.069	0.103	0.066
	(1)	100	1	1	1	1	1	1	0.120	0.066	0.120	0.071	0.119	0.066
	(1)	200	1	1	1	1	1	1	0.111	0.057	0.111	0.057	0.111	0.057
	(2)	50	1	1	1	1	1	1	0.126	0.085	0.154	0.105	0.119	0.073
	(2)	100	1	1	1	1	1	1	0.098	0.058	0.108	0.066	0.094	0.055
	(2)	200	1	1	1	1	1	1	0.111	0.055	0.114	0.056	0.108	0.054
	(3)	50	1	1	1	1	1	0.999	0.118	0.069	0.117	0.072	0.120	0.070
	(3)	100	1	1	1	1	1	1	0.102	0.067	0.106	0.065	0.103	0.067
	(3)	200	1	1	1	1	1	1	0.103	0.046	0.103	0.049	0.102	0.046
	(4)	50	0.452	0.328	0.863	0.771	0.552	0.421	0.114	0.061	0.111	0.061	0.114	0.061
	(4)	100	0.640	0.491	0.990	0.967	0.761	0.637	0.098	0.063	0.104	0.063	0.098	0.062
	(4)	200	0.840	0.733	1	0.999	0.933	0.876	0.105	0.042	0.108	0.042	0.105	0.043
	(5)	50	1	1	1	1	1	1	0.128	0.078	0.163	0.098	0.115	0.077
	(5)	100	1	1	1	1	1	1	0.098	0.053	0.115	0.063	0.091	0.051
	(5)	200	1	1	1	1	1	1	0.100	0.050	0.103	0.054	0.098	0.050
Ex 3.5.3	(1)	50	1	1	1	1	1	1	0.157	0.098	0.223	0.137	0.156	0.098
	(1)	100	1	1	1	1	1	1	0.158	0.089	0.188	0.124	0.157	0.090
	(1)	200	1	1	1	1	1	1	0.122	0.074	0.161	0.091	0.121	0.074
	(2)	50	1	1	1	1	1	1	0.140	0.078	0.190	0.118	0.137	0.075
	(2)	100	1	1	1	1	1	1	0.139	0.080	0.171	0.105	0.136	0.080
	(2)	200	1	1	1	1	1	1	0.109	0.053	0.127	0.069	0.108	0.053

REMARK **3.5.1**. *In Example 3.5.3.1, marginally the p-many two-dimensional groups of X and Y are not identically distributed, but each of the $2p$ unidimensional components of X and Y have* *identical distributions. Consequently, choosing $d_i = 1$ for $1 \leq i \leq p$ leads to trivial power of even our proposed tests, as is evident from Table 3.5 below. This demonstrates that grouping allows us to detect a wider range of alternatives.*

Table 3.5: Empirical power in Example 3.5.3.1 if we choose $d_i = 1$ for $1 \leq i \leq p$.

			I		II		III		IV		V		VI	
		p	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
Ex 3.5.3	(1)	50	0.144	0.087	0.133	0.076	0.143	0.086	0.174	0.107	0.266	0.170	0.175	0.105
	(1)	100	0.145	0.085	0.134	0.070	0.142	0.085	0.157	0.098	0.223	0.137	0.156	0.098
	(1)	200	0.126	0.063	0.101	0.058	0.111	0.065	0.158	0.089	0.188	0.124	0.157	0.090

3.5.2 Testing for independence

We study the empirical size and power of tests for independence between two high dimensional random vectors. We consider the t-tests based on the following metrics:

I. $\mathcal{D}$ with $d_i = 1$ and ρ_i be the Euclidean distance for $1 \leq i \leq p$;

II. $\mathcal{D}$ with $d_i = 1$ and ρ_i be the distance induced by the Laplace kernel for $1 \leq i \leq p$;

III. $\mathcal{D}$ with $d_i = 1$ and ρ_i be the distance induced by the Gaussian kernel for $1 \leq i \leq p$;

IV. the usual Euclidean distance covariance;

V. HSIC with the Laplace kernel;

VI. HSIC with the Gaussian kernel.

We also compare the empirical size and power of the above tests with the

VII. projection correlation based test for independence proposed by Zhu et al. (2017), which is shown to have higher empirical power compared to the usual Euclidean distance covariance when the dimensions are relatively large. The numerical examples we consider are motivated from Zhu et al. (2020).

EXAMPLE **3.5.4**. *Consider* $X_k = (X_{k1}, \ldots, X_{kp})$ *and* $Y_k = (Y_{k1}, \ldots, Y_{kp})$ *for* $k = 1, \ldots, n$. *We generate i.i.d. samples from the following models :*

1. $X_k \sim N(0, I_p)$ *and* $Y_k \sim N(0, I_p)$.

2. $X_k \sim AR(1), \phi = 0.5$, $Y_k \sim AR(1), \phi = -0.5$, *where* $AR(1)$ *denotes the autoregressive model of order* 1 *with parameter* ϕ.

3. $X_k \sim N(0, \Sigma)$ *and* $Y_k \sim N(0, \Sigma)$, *where* $\Sigma = (\sigma_{ij})_{i,j=1}^{p}$ *with* $\sigma_{ij} = 0.7^{|i-j|}$.

EXAMPLE **3.5.5**. *Consider* $X_k = (X_{k1}, \ldots, X_{kp})$ *and* $Y_k = (Y_{k1}, \ldots, Y_{kp})$, $k = 1, \ldots, n$. *We generate i.i.d. samples from the following models :*

1. $X_k \sim N(0, I_p)$ *and* $Y_{kj} = X_{kj}^2$ *for* $j = 1, \ldots, p$.

2. $X_k \sim N(0, I_p)$ *and* $Y_{kj} = \log |X_{kj}|$ *for* $j = 1, \ldots, p$.

3. $X_k \sim N(0, \Sigma)$ *and* $Y_{kj} = X_{kj}^2$ *for* $j = 1, \ldots, p$, *where* $\Sigma = (\sigma_{ij})_{i,j=1}^{p}$ *with* $\sigma_{ij} = 0.7^{|i-j|}$.

EXAMPLE **3.5.6**. *Consider* $X_k = (X_{k1}, \ldots, X_{kp})$ *and* $Y_k = (Y_{k1}, \ldots, Y_{kp})$, $k = 1, \ldots, n$. *Let* $\circ$ *denote the Hadamard product of matrices. We generate i.i.d. samples from the following models:*

1. $X_{kj} \sim U(-1, 1)$ *for* $j = 1, \ldots, p$, *and* $Y_k = X_k \circ X_k$.

2. $X_{kj} \sim U(0, 1)$ *for* $j = 1, \ldots, p$, *and* $Y_k = 4X_k \circ X_k - 4X_k + 2$.

3. $X_{kj} = \sin(Z_{kj})$ *and* $Y_{kj} = \cos(Z_{kj})$ *with* $Z_{kj} \sim U(0, 2\pi)$ *and* $j = 1, \ldots, p$.

For each example, we draw 1000 simulated datasets and perform tests for independence between the two variables based on the proposed dependence metrics, and the usual Euclidean dCov and HSIC. We try a small sample size $n = 50$ and dimensions $p = 50, 100$ and 200. For the tests

II, III, V and VI, we chose the bandwidth parameter heuristically as the median distance between the sample observations. Table 3.6 reports the proportion of rejections out of the 1000 simulation runs for the different tests. For VII, we conduct a permutation based test with 500 replicates.

In Example 4.4.2, the data generating scheme suggests that the variables X and Y are independent. The results in Table 3.6 show that the tests based on the proposed dependence metrics perform almost equally good as the other competitors, and the rejection probabilities are quite close to the 10% or 5% nominal level. In Examples 4.4.3 and 4.4.4, the variables are clearly (componentwise non-linearly) dependent by virtue of the data generating scheme. The results indicate that the tests based on the proposed dependence metrics are able to detect the componentwise non-linear dependence between the two high-dimensional random vectors unlike the tests based on the usual Euclidean dCov and HSIC, and thereby outperform the latter in terms of empirical power. Also, our proposed tests clearly perform far better compared to the projection correlation based test.

Table 3.6: Empirical size and power for the different tests of independence.

			I		II		III		IV		V		VI		VII	
		p	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
Ex 4.4.2	(1)	50	0.115	0.053	0.109	0.055	0.106	0.053	0.112	0.060	0.112	0.053	0.111	0.061	0.119	0.059
	(1)	100	0.106	0.057	0.090	0.046	0.095	0.048	0.111	0.060	0.112	0.059	0.113	0.060	0.116	0.062
	(1)	200	0.076	0.031	0.084	0.046	0.084	0.042	0.096	0.035	0.090	0.038	0.095	0.035	0.091	0.038
	(2)	50	0.101	0.052	0.096	0.061	0.094	0.053	0.096	0.050	0.103	0.054	0.096	0.052	0.094	0.050
	(2)	100	0.080	0.036	0.083	0.035	0.086	0.042	0.081	0.041	0.088	0.044	0.083	0.041	0.081	0.037
	(2)	200	0.117	0.051	0.098	0.056	0.103	0.052	0.104	0.048	0.103	0.052	0.106	0.048	0.101	0.050
	(3)	50	0.093	0.056	0.098	0.052	0.097	0.056	0.091	0.052	0.080	0.050	0.087	0.052	0.094	0.044
	(3)	100	0.104	0.052	0.085	0.046	0.091	0.054	0.104	0.048	0.105	0.051	0.102	0.048	0.098	0.045
	(3)	200	0.105	0.059	0.110	0.057	0.103	0.051	0.106	0.055	0.099	0.052	0.105	0.056	0.109	0.058
Ex 4.4.3	(1)	50	1	1	1	1	1	1	0.267	0.172	0.534	0.398	0.277	0.182	0.388	0.280
	(1)	100	1	1	1	1	1	1	0.171	0.102	0.284	0.180	0.167	0.102	0.323	0.204
	(1)	200	1	1	1	1	1	1	0.130	0.075	0.194	0.108	0.128	0.073	0.302	0.188
	(2)	50	1	1	1	1	1	1	0.154	0.092	0.199	0.130	0.154	0.091	0.147	0.077
	(2)	100	1	1	1	1	1	1	0.109	0.050	0.128	0.064	0.108	0.049	0.108	0.048
	(2)	200	1	1	1	1	1	1	0.099	0.057	0.107	0.060	0.097	0.057	0.101	0.048
	(3)	50	1	1	1	1	1	1	0.654	0.546	0.981	0.959	0.708	0.631	0.661	0.545
	(3)	100	1	1	1	1	1	1	0.418	0.309	0.790	0.700	0.455	0.343	0.535	0.419
	(3)	200	1	1	1	1	1	1	0.277	0.188	0.504	0.391	0.284	0.193	0.454	0.345
Ex 4.4.4	(1)	50	1	1	1	1	1	1	0.129	0.072	0.193	0.105	0.130	0.071	0.141	0.076
	(1)	100	1	1	1	1	1	1	0.145	0.069	0.158	0.091	0.145	0.069	0.155	0.084
	(1)	200	1	1	1	1	1	1	0.113	0.065	0.123	0.067	0.113	0.065	0.130	0.068
	(2)	50	1	1	1	1	1	1	0.129	0.072	0.193	0.105	0.130	0.071	0.141	0.076
	(2)	100	1	1	1	1	1	1	0.145	0.069	0.158	0.091	0.145	0.069	0.155	0.084
	(2)	200	1	1	1	1	1	1	0.113	0.065	0.123	0.067	0.113	0.065	0.130	0.068
	(3)	50	0.540	0.388	1	1	0.859	0.760	0.110	0.057	0.108	0.063	0.111	0.056	0.092	0.049
	(3)	100	0.550	0.416	1	1	0.857	0.761	0.108	0.063	0.112	0.063	0.108	0.062	0.097	0.051
	(3)	200	0.542	0.388	1	1	0.872	0.765	0.106	0.049	0.111	0.051	0.106	0.050	0.089	0.044

3.5.3 Real data analysis

3.5.3.1 Testing for homogeneity of distributions

We consider the two sample testing problem of homogeneity of two high-dimensional distributions on Earthquakes data. The dataset has been downloaded from UCR Time Series Classification Archive (`https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`). The data are taken from Northern California Earthquake Data Center. There are 368 negative and 93 positive earthquake events and each data point is of length 512.

Table 3.7 shows the p-values corresponding to the different tests for the homogeneity of distributions between the two classes. Here we set $d_i = 1$ for tests I-III. Clearly the tests based on the proposed homogeneity metrics reject the null hypothesis of equality of distributions at 5% level. However the tests based on the usual Euclidean energy distance and MMD fail to reject the null at 5% level, thereby indicating no significant difference between the distributions of the two classes.

Table 3.7: p-values corresponding to the different tests for homogeneity of distributions for Earthquakes data.

I	II	III	IV	V	VI
2.27×10^{-93}	3.19×10^{-86}	9.74×10^{-110}	0.070	0.068	0.070

3.5.3.2 Testing for independence

We consider the daily closed stock prices of $p = 126$ companies under the finance sector and $q = 122$ companies under the healthcare sector on the first dates of each month during the time period between January 1, 2017 and December 31, 2018. The data has been downloaded from Yahoo Finance via the R package 'quantmod'. At each time t, denote the closed stock prices of these companies from the two different sectors by $X_t = (X_{1t}, \ldots, X_{pt})$ and $Y_t = (Y_{1t}, \ldots, Y_{qt})$ for $1 \leq t \leq 24$. We consider the stock returns $S_t^X = (S_{1t}^X, \ldots, S_{pt}^X)$ and $S_t^Y = (S_{1t}^Y, \ldots, S_{qt}^Y)$ for $1 \leq t \leq 23$, where $S_{it}^X = \log \frac{X_{i,t+1}}{X_{it}}$ and $S_{jt}^Y = \log \frac{Y_{j,t+1}}{Y_{jt}}$ for $1 \leq i \leq p$ and $1 \leq j \leq q$. It seems intuitive that the stock returns for the companies under two different sectors are not totally independent, especially when a large number of companies are being considered. Table 3.8 shows the p-values corresponding to the different tests for independence between $\{S_t^X\}_{t=1}^{23}$ and $\{S_t^Y\}_{t=1}^{23}$, where we set $d_i = g_i = 1$ for the proposed tests. The tests based on the proposed dependence metrics deliver much smaller p-values compared to the tests based on traditional metrics. We note that the tests based on the usual dCov and HSIC as well as projection correlation fail to reject the null at 5% level, thereby indicating cross-sector independence of stock return values. These results are consistent with the fact that the dependence among financial asset returns is usually nonlinear and thus cannot be fully characterized by traditional metrics in the high dimensional setup.

Table 3.8: p-values corresponding to the different tests for cross-sector independence of stock returns data.

I	II	III	IV	V	VI	VII
4.91×10^{-12}	4.29×10^{-11}	1.12×10^{-11}	0.093	0.084	0.099	0.154

We present an additional real data example on testing for independence in high dimensions in Section B.3 of the appendix. There the data admits a natural grouping, and our results indicate that our proposed tests for independence exhibit better power when we consider the natural grouping than when we consider unit group sizes. It is to be noted that considering unit group sizes makes our proposed statistics essentially equivalent to the marginal aggregation approach proposed by Zhu et al. (2020). This indicates that grouping or clustering might improve the power of testing as they are capable of detecting a wider range of dependencies.

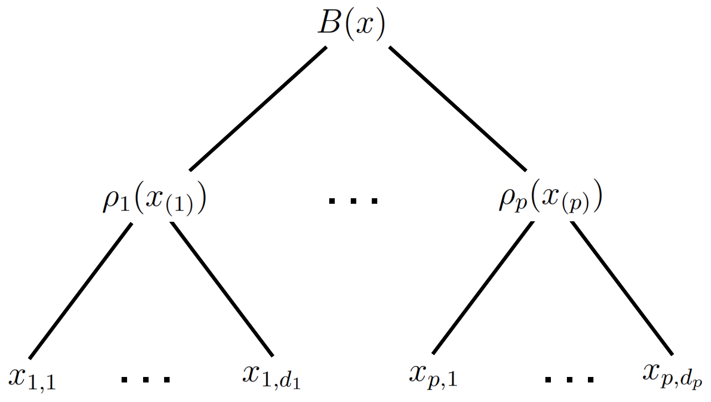
3.6 Discussions

In this work, we introduce a family of distances for high dimensional Euclidean spaces. Built on the new distances, we propose a class of distance and kernel-based metrics for high-dimensional two-sample and independence testing. The proposed metrics overcome certain limitations of the traditional metrics constructed based on the Euclidean distance. The new distance we introduce corresponds to a semi-norm given by

$$B(x) = \sqrt{\rho_1(x_{(1)}) + \ldots, \rho_p(x_{(p)})},$$

where $\rho_i(x_{(i)}) = \rho_i(x_{(i)}, 0_{d_i})$ and $x = (x_{(1)}, \ldots, x_{(p)}) \in \mathbb{R}^{\tilde{p}}$ with $x_{(i)} = (x_{i,1}, \ldots, x_{i,d_i})$. Such a semi-norm has an interpretation based on a tree as illustrated by Figure 3.3.

Figure 3.3: An interpretation of the semi-norm $B(\cdot)$ based on a tree



Tree structure provides useful information for doing grouping at different levels/depths. Theoretically, grouping allows us to detect a wider range of alternatives. For example, in two-sample testing, the difference between two one-dimensional marginals is always captured by the difference between two higher dimensional marginals that contain the two one-dimensional marginals. The same thing is true for dependence testing. Generally, one would like to find blocks which are nearly independent, but the variables inside a block have significant dependence among themselves. It is interesting to develop an algorithm for finding the optimal groups using the data and perhaps some auxiliary information. Another interesting direction is to study the semi-norm and distance constructed based on a more sophisticated tree structure. For example, in microbiome-wide association studies, phylogenetic tree or evolutionary tree which is a branching diagram or “tree” showing the evolutionary relationships among various biological species. Distance and kernel-based metrics constructed based on the distance utilizing the phylogenetic tree information is expected to be more powerful in signal detection. We leave these topics for future investigation.

4. NONPARAMETRIC MULTIPLE CHANGE-POINT DETECTION FOR HIGH DIMENSIONAL DATA

4.1 Background and notations

Change-point detection has been a classical and well-established problem in statistics, aiming to detect lack of homogeneity in a sequence of time-ordered observations. This finds abundance of applications in a wide variety of fields, for example, bioinformatics (see Picard et al. (2005), Curtis et al. (2012)), neuroscience (see Park et al. (2015)), digital speech processing (see Rabiner and Schäfer (2007)), social network analysis (see McCulloh (2009)), and so on. A nonparametric change-point detection procedure is concerned with detecting and localizing quite general types of changes in the data generating distribution, rather than only changes in mean. This challenging problem of detecting abrupt distributional changes in the nonparametric setting has been addressed in the literature over the last couple of decades. But many of the methodologies developed suffer from several limitations, for example, applicability only for real-valued data or in the low-dimensional setting, assumption that the number of true change-points is known, etc. Harchaoui and Cappé (2007) proposed a kernel-based procedure assuming a known number of change-points, which reduces its practical interest. Zou et al. (2014) proposed a nonparametric maximum likelihood approach for detecting multiple (unknown number of) change-points using BIC, but is only applicable for real-valued data. Lung-Yut-Fong et al. (2012) developed a nonparametric approach based on marginal rank statistics, which requires the number of observations to be larger than the dimension of the data. Arlot et al. (2012) proposed a kernel-based multiple change-point detection algorithm for multivariate (but fixed dimensional) or complex (non-Euclidean) data. Some graph-based tests have been proposed recently by Chen and Zhang (2015) and Chu and Chen (2019) for high-dimensional data, which allow us to detect only one or two change-points. Matteson and James (2014) proposed a procedure for estimating multiple change-point locations, namely E-Divisive, built upon an energy distance based test that applies to multivariate observations of arbitrary (but fixed) dimensions. Biau et al. (2016) rigorously derived the asymptotic distribution of the statistic proposed by Matteson and James (2014), thereby adding theoretical justifications to their methodology. However, some recent research revelations on the performance of energy distance for growing dimensions put a question on its performance when we have a sequence of high-dimensional observations. To the best of our knowledge, the literature in general on nonparametric multiple change-point detection in the high-dimensional setup is quite scarce till date.

Energy distance, proposed by Székely et al. (2004, 2005) and Baringhaus and Franz (2004), is a classical distance-based measure of equality of two multivariate distributions, taking the value zero if and only if the two random vectors are identically distributed. Such a complete characterization of homogeneity of distributions lends itself for reasonable use in one-sample goodness-of-fit testing and two-sample testing for equality of distributions, and has been widely studied in the literature over the last couple of decades. In a very recent paper, Chakraborty and Zhang (2019) showed a striking result that energy distance based on the usual Euclidean distance cannot completely characterize the homogeneity of the two high-dimensional distributions in the sense that it can only detect the *equality of means and the traces of covariance matrices* of the two high-dimensional random vectors. In other words, the Euclidean energy distance fails to detect inhomogeneity between two high-dimensional distributions beyond the first two moments. To overcome such a limitation, the authors proposed a new class of homogeneity metrics which inherits the desirable properties of energy distance in the low-dimensional setting. And more importantly, in the high-dimensional setup the new class of homogeneity metrics is capable of detecting the *pairwise homogeneity of the low-dimensional marginal distributions*, going beyond the scope of the Euclidean energy distance. In other words, the proposed class of homogeneity metrics can capture a wider range of inhomogeneity of distributions compared to the classical Euclidean energy distance in the high-dimensional framework. The core of their methodology is a new way of defining the distance between sample points (interpoint distance) in the high-dimensional Euclidean spaces.

This work focuses on estimating an unknown number of multiple change-point locations in an independent sequence of $\mathbb{R}^p$-valued observations of size n, where p can by far exceed n. Our approach essentially rests upon distance-based nonparametric two-sample tests for homogeneity of two high-dimensional distributions. We first construct a single change-point location estimator M_n based on the homogeneity metrics proposed by Chakraborty and Zhang (2019) via defining a cumulative sum process in an embedded Hilbert space. It essentially generalizes the single change-point location estimator developed by Matteson and James (2014) and Biau et al. (2016) in the high-dimensional setup, providing a unifying framework. Testing for the statistical significance of the estimated candidate change-point location necessitates determining the quantiles of the distribution of M_n. The key theoretical innovation of this paper is to rigorously derive the asymptotic null distribution of M_n as both the dimension p and the sample size n grow to infinity, with n growing at a smaller rate compared to p. Such a setup is typically known in the literature as the high dimension medium sample size (HDMSS) framework. The intrinsic difficulty is to establish the uniform weak convergence of an underlying stochastic process under certain mild assumptions, which has been non-trivial and challenging. Because of the pivotal nature of the limiting null distribution, its quantiles can be approximated using a large number of Monte Carlo simulations. We propose an algorithm for single change-point detection based on a permutation procedure to better approximate the quantiles of the distribution of M_n. Subsequently, we combine the idea of Wild Binary Segmentation (WBS) proposed by Fryzlewicz (2014) to recursively estimate and test for the significance of (an unknown number of) multiple change-point locations. The superior performance of our procedure compared to several of the existing methodologies is illustrated via both simulated and real datasets.

Notation. Denote by $\|\cdot\|_p$ the Euclidean norm of $\mathbb{R}^p$ (we shall use it interchangeably with $\|\cdot\|$ when there is no confusion). Let 0_p be the origin of $\mathbb{R}^p$. We use "$X \overset{d}{=} Y$" to indicate that X and Y are identically distributed. Let X', X'', X''' be independent copies of X. 'O' and 'o' stand for the usual notations in mathematics : 'is of the same order as' and 'is ultimately smaller than'. We use the symbol "$a \lesssim b$" to indicate that $a \leq C\,b$ for some constant $C > 0$. We utilize the order in probability notations such as stochastic boundedness O_p (big O in probability), convergence in probability o_p (small o in probability) and equivalent order $\asymp$, which is defined as follows: for a sequence of random variables $\{Z_n\}_{n=1}^{\infty}$ and a sequence of real numbers $\{a_n\}_{n=1}^{\infty}$, $Z_n \asymp_p a_n$ if and only if $Z_n/a_n = O_p(1)$ and $a_n/Z_n = O_p(1)$ as $n \to \infty$. If $Z_n \xrightarrow{P} Z$ as $n \to \infty$, then we say $\operatorname{plim}_{n\to\infty} Z_n = Z$. For a metric space $(\mathcal{X}, d_{\mathcal{X}})$, let $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}_1(\mathcal{X})$ denote the set of all finite signed Borel measures on $\mathcal{X}$ and all probability measures on $\mathcal{X}$, respectively. Define $\mathcal{M}^1_{d_{\mathcal{X}}}(\mathcal{X}) := \{v \in \mathcal{M}(\mathcal{X}) \,:\, \exists\, x_0 \in \mathcal{X} \text{ s.t. } \int_{\mathcal{X}} d_{\mathcal{X}}(x, x_0)\, d|v|(x) < \infty\}$. For a matrix $A = (a_{kl})_{k,l=1}^{n} \in \mathbb{R}^{n\times n}$, define its $\mathcal{U}$-centered version $\tilde{A} = (\tilde{a}_{kl}) \in \mathbb{R}^{n\times n}$ as follows

$$\tilde{a}_{kl} = \begin{cases} a_{kl} - \dfrac{1}{n-2}\displaystyle\sum_{j=1}^{n} a_{kj} - \dfrac{1}{n-2}\displaystyle\sum_{i=1}^{n} a_{il} + \dfrac{1}{(n-1)(n-2)}\displaystyle\sum_{i,j=1}^{n} a_{ij}, & k \neq l, \\ 0, & k = l, \end{cases} \tag{4.1}$$

for $k, l = 1, \ldots, n$. Let $\mathbb{1}(A)$ denote the indicator function associated with a set A. Finally, denote by $\lfloor a \rfloor$ and $\{a\}$ the integer and fractional part of $a \in \mathbb{R}$, respectively.

4.2 An overview

4.2.1 Energy Distance

Energy distance (see Székely et al. (2004, 2005), Baringhaus and Franz (2004)) or the Euclidean energy distance between two random vectors $X, Y \in \mathbb{R}^p$ and $X \perp\!\!\!\perp Y$ with $\mathbb{E}\|X\|_p < \infty$ and $\mathbb{E}\|Y\|_p < \infty$, is defined as

$$ED(X, Y) \;=\; \frac{1}{c_p}\int_{\mathbb{R}^p} \frac{|f_X(t) - f_Y(t)|^2}{\|t\|_p^{1+p}}\, dt\,, \tag{4.2}$$

where f_X and f_Y are the characteristic functions of X and Y respectively, and $c_p = \pi^{(1+p)/2}/\,\Gamma((1+p)/2)$ is a constant with $\Gamma(\cdot)$ being the complete gamma function. Theorem 1 in Székely et al. (2005) shows that $ED(X, Y) \geq 0$ and the equality holds if and only if $X \overset{d}{=} Y$. In other words, energy distance can completely characterize the homogeneity between two multivariate distributions.

Alternatively an equivalent expression for $ED(X,Y)$ is given by

$$ED(X,Y) \;=\; 2\,\mathbb{E}\|X-Y\|_p - \mathbb{E}\|X-X'\|_p - \mathbb{E}\|Y-Y'\|_p\;, \tag{4.3}$$

where (X',Y') is an independent copy of (X,Y).

In general, for an arbitrary metric space $(\mathcal{X},K)$, the generalized energy distance between $X \sim P_X$ and $Y \sim P_Y$ where $P_X, P_Y \in \mathcal{M}_1(\mathcal{X}) \cap \mathcal{M}_K^1(\mathcal{X})$ is defined as

$$ED_K(X,Y) \;=\; 2\,\mathbb{E}\,K(X,Y) - \mathbb{E}\,K(X,X') - \mathbb{E}\,K(Y,Y')\;. \tag{4.4}$$

Definition **4.2.1** (Spaces of negative type). *A metric space $(\mathcal{X},K)$ is said to have negative type if for all $n \geq 1$, $x_1,\ldots,x_n \in \mathcal{X}$ and $\alpha_1,\ldots,\alpha_n \in \mathbb{R}$ with $\sum_{i=1}^n \alpha_i = 0$, we have*

$$\sum_{i,j=1}^{n} \alpha_i\,\alpha_j\,K(x_i,x_j) \leq 0\;. \tag{4.5}$$

The metric space $(\mathcal{X},K)$ is said to be of strong negative type if the equality in (4.5) holds only when $\alpha_i = 0$ for all $i \in \{1,\ldots,n\}$.

By Theorem 3.16 in Lyons (2013), every separable Hilbert space is of strong negative type. In particular, Euclidean spaces are separable Hilbert spaces and therefore have strong negative type.

If $(\mathcal{X},K)$ has strong negative type, then $ED_K(X,Y)=0$ if and only if $X \overset{d}{=} Y$. In other words, the completely characterization of the homogeneity of two distributions holds good in any metric spaces of strong negative type (we refer the reader to Lyons (2013) and Sejdinovic et al. (2013) for detailed discussions). Thus the quantification of homogeneity of distributions by the Euclidean energy distance given in (4.3) is just a special case when K is the Euclidean distance on $\mathcal{X}=\mathbb{R}^p$.

This quantification of homogeneity of distributions lends itself for reasonable use in one-sample goodness-of-fit testing and two-sample testing for equality of distributions. Suppose $\{X_i\}_{i=1}^n$ and $\{X_j\}_{j=n+1}^N$ are two independent i.i.d samples on X and Y taking values in $(\mathcal{X},K)$. An U-statistic type estimator of the generalized energy distance between X and Y is defined as

$$\begin{aligned} E_{K;\,1,N,n} \;=\; & \frac{2}{n(N-n)} \sum_{i_1=1}^{n} \sum_{i_2=n+1}^{N} K(X_{i_1}, X_{i_2}) \;-\; \frac{1}{n(n-1)} \sum_{1 \le i_1 \ne i_2 \le n} K(X_{i_1}, X_{i_2}) \\ & -\; \frac{1}{(N-n)(N-n-1)} \sum_{n+1 \le i_1 \ne i_2 \le N} K(X_{i_1}, X_{i_2})\,. \end{aligned} \tag{4.6}$$

We refer the reader to Section A.1 in the supplementary materials of Chakraborty and Zhang (2019) for a comprehensive overview of the properties and asymptotic behavior of the U-statistic type estimator of $E_K(X,Y)$ in the low-dimensional setting.

4.2.2 Modifications to the classical energy distance in high dimensions

The question of interest is how do the classical distance-based homogeneity metrics like energy distance behave in the high-dimensional framework. Consider two $\mathbb{R}^p$-valued random vectors $X = (X_1, \ldots, X_p)$ and $Y = (Y_1, \ldots, Y_p)$. Chakraborty and Zhang (2019) in their recent paper showed a striking result that when dimension grows high, the Euclidean energy distance between X and Y can only capture the equality of the means and the first spectral means, viz. $\mu_X = \mu_Y$ and $\operatorname{tr}\Sigma_X = \operatorname{tr}\Sigma_Y$, where μ_X and μ_Y, and, Σ_X and Σ_Y are the mean vectors and the covariance matrices of X and Y, respectively.

To illustrate, consider the case $X \sim N(\mu, I_p)$ with $\mu = (1, \ldots, 1) \in \mathbb{R}^p$ and $Y_i \overset{ind}{\sim}$ Exponential (1) for $1 \le i \le p$. That is, $\mu_X = \mu_Y$ and $\operatorname{tr}\Sigma_X = \operatorname{tr}\Sigma_Y$ although X and Y have different distributions. Section 6.1 in Chakraborty and Zhang (2019) demonstrates that when p is much larger than the sample sizes observed, the Euclidean energy distance does a poor job in detecting the in-homogeneity of the two distributions.

Such a limitation of the classical Euclidean energy distance arises essentially due to the use of Euclidean distance. The authors proposed a new class of homogeneity metrics to overcome such a limitation of the Euclidean energy distance, which is based on a new way of defining the distance between sample points (interpoint distance) in the high-dimensional Euclidean spaces. For $x, x' \in \mathbb{R}^p$ with $x = (x_1, \ldots, x_p)$ and $x' = (x'_1, \ldots, x'_p)$, consider the distance metric

$$K(x,x') \;=\; \Big(\sum_{j=1}^{p} |x_j - x'_j|\Big)^{1/2} \;=\; \|x - x'\|_1^{1/2}\,, \tag{4.7}$$

where $\|x\|_1 = \sum_{j=1}^{p} |x_j|$ is the l_1 or the absolute norm on $\mathbb{R}^p$.

Based on the new distance metric defined in (4.7), the following homogeneity metric $\mathcal{E}$ is proposed to quantify the homogeneity of the distributions of X and Y:

$$\mathcal{E}(X,Y) \;=\; 2\,\mathbb{E}\,K(X,Y) \,-\, \mathbb{E}\,K(X,X') \,-\, \mathbb{E}\,K(Y,Y')\,, \tag{4.8}$$

which is essentially a generalized energy distance as defined in (4.4) with $\mathcal{X} = \mathbb{R}^p$ and K as defined in (4.7). Under the assumption that $\sup_{1\le i\le p} \mathbb{E}\,|X_i|^{1/2} < \infty$, $\mathcal{E}$ is finite.

For fixed p, $(\mathbb{R}^p, K)$ is shown to have strong negative type and hence $\mathcal{E}(X,Y) = 0$ if and only if $X \overset{d}{=} Y$. In other words, $\mathcal{E}(X,Y)$ completely characterizes the homogeneity of the distributions of X and Y in the low-dimensional setting. Theorem 4.1 and Lemma 4.1 in this paper show that when dimension grows high, $\mathcal{E}(X,Y)$ can capture the pairwise homogeneity of the univariate marginal distributions of X_i and Y_i. Clearly $X_i \overset{d}{=} Y_i$ for $1 \le i \le p$ implies $\mu_X = \mu_Y$ and $\operatorname{tr}\Sigma_X = \operatorname{tr}\Sigma_Y$, and therefore the proposed class of homogeneity metrics can capture a wider range of in-homogeneity of distributions compared to the Euclidean energy distance in the high-dimensional framework. Completely characterizing the discrepancy between two high-dimensional distributions might have some intrinsic difficulties and remains as an open problem for future investigation.

Consider i.i.d. samples $\{X_k\}_{k=1}^{n}$ and $\{X_l\}_{l=n+1}^{N}$ from the respective distributions of X and Y. The authors propose an unbiased U-statistic type estimator $\mathcal{E}_{1,N,n}$ of $\mathcal{E}(X,Y)$ as in equation (4.6) with K being the metric defined in (4.7).

Denote $X_{1:k} = \{X_1, \ldots, X_k\}$ and $X_{(k+1):N} = \{X_{k+1}, \ldots, X_N\}$ for $1 \le k \le N$. Also denote $v_s = s(s-3)/2$ for $s = n,\, N-n$. The pooled variance estimator $S_{n,N-n}$ of $\mathcal{E}_{1,N,n}$ is constructed as

$$S_{n,N-n} \;:=\; \frac{4(n-1)(N-n-1)\,cdCov^2_{n,N-n}(X_{1:n}, X_{(n+1):N}) \;+\; 4\,v_n\,\widetilde{\mathcal{D}^2_n}(X_{1:n}) \;+\; 4\,v_{N-n}\,\widetilde{\mathcal{D}^2_{N-n}}(X_{(n+1):N})}{(n-1)(N-n-1) + v_n + v_{N-n}}\,,$$

where

$$cdCov^2_{n,N-n}(X_{1:n}, X_{(n+1):N}) \;:=\; \frac{1}{(n-1)(N-n-1)} \sum_{i_1=1}^{n} \sum_{i_2=n+1}^{N} \widehat{K}(X_{i_1}, X_{i_2})^2 \,,$$

$$\widehat{K}(X_{i_1}, X_{i_2}) := K(X_{i_1}, X_{i_2}) - \frac{1}{n}\sum_{i_3=1}^{n} K(X_{i_3}, X_{i_2}) - \frac{1}{N-n}\sum_{i_4=n+1}^{N} K(X_{i_1}, X_{i_4}) + \frac{1}{n(N-n)} \sum_{i_3=1}^{n}\sum_{i_4=n+1}^{N} K(X_{i_3}, X_{i_4}) \qquad \text{for } 1 \le i_1 \le n,\, n+1 \le i_2 \le N\,,$$

$$\widetilde{\mathcal{D}^2_n}(X_{1:n}) \;:=\; \frac{1}{n(n-3)} \sum_{1 \le k \ne l \le n} \tilde{a}^2_{kl} \;,\; \text{ and } \; \widetilde{\mathcal{D}^2_{N-n}}(X_{(n+1):N}) \;:=\; \frac{1}{(N-n)(N-n-3)} \sum_{n+1 \le k \ne l \le N} \tilde{b}^2_{kl}\,,$$

with $\tilde{A} = (\tilde{a}_{kl})_{k,l=1}^n \in \mathbb{R}^{n \times n}$ and $\tilde{B} = (\tilde{b}_{kl})_{k,l=n+1}^N \in \mathbb{R}^{(N-n)\times(N-n)}$ being the $\mathcal{U}$-centered versions (see (4.1)) of the distance matrices $A := (a_{kl})_{k,l=1}^n := (K(X_k, X_l))_{k,l=1}^n$ and $B := (b_{kl})_{k,l=n+1}^N := (K(X_k, X_l))_{k,l=n+1}^N$, respectively.

Based on $\mathcal{E}_{1,N,n}$ and its pooled variance estimator, the authors propose a two-sample test statistic

$$T_{1,N,n}(X) \;=\; \frac{\mathcal{E}_{1,N,n}}{a_{n,N-n}\; S^{1/2}_{n,N-n}}\,, \tag{4.9}$$

where

$$a_{n,N-n} \;=\; \sqrt{\frac{1}{n(N-n)} + \frac{1}{2n(n-1)} + \frac{1}{2(N-n)(N-n-1)}}\;.$$

We will denote $T_{1,N,n}(X)$ simply by $T_{1,N,n}$ henceforth. Likewise, henceforth we will simply denote

$cdCov^2_{n,N-n}(X_{1:n}, X_{(n+1):N})$, $\widetilde{\mathcal{D}^2_n}(X_{1:n})$ and $\widetilde{\mathcal{D}^2_{N-n}}(X_{(n+1):N})$ respectively by $cdCov^2_{n,N-n}$, $\widetilde{\mathcal{D}^2_n}$ and $\widetilde{\mathcal{D}^2_{N-n}}$ for notational simplicities. Note that the construction of the pooled variance estimator $S_{n,N-n}$ and hence the two-sample statistic $T_{1,N,n}$ requires $n \ge 4$ and $N-n \ge 4$, i.e., $4 \le n \le N-4$.

Under certain mild assumptions, it is shown in Theorem B.1 in their paper that under $H_0 : X \overset{d}{=} Y$, $T_{1,N,n} \xrightarrow{d} N(0,1)$ as $p \to \infty$ and $n, (N-n) \to \infty$ at a slower rate than p. Based on the asymptotic behavior of $T_{1,N,n}$ for growing dimensions, the authors propose a test for H_0 against a general alternative. Owing to the pivotal nature of the limiting distribution of $T_{1,N,n}$, no resampling-based inference is needed.

4.3 Methodology

Consider an independent sequence of $\mathbb{R}^p$-valued observations $\{X_t\}_{t=1}^n$, where the dimension p is typically much higher than the sample size n. We are concerned with testing the null hypothesis $H_0 \,:\, X_t \sim F_0\,,\; t = 1, \dots, n$ against the single change-point alternative

$$H_1 \;:\; \exists\; 1 \leq \tau_0 < n\,, \qquad X_t \sim \begin{cases} F_0, & 1 \leq t \leq \tau_0, \\ F_1, & \tau_0 + 1 \leq t \leq n, \end{cases} \tag{4.10}$$

or the multiple change-point alternative

$$H_2 \;:\; \exists\; N_0 \in \mathbb{Z},\; N_0 \geq 2,\; 1 \leq \tau_1 < \cdots < \tau_{N_0} < n, \qquad X_t \sim \begin{cases} F_0, & 1 \leq t \leq \tau_1, \\ F_1, & \tau_1 + 1 \leq t \leq \tau_2, \\ \vdots & \\ F_{N_0}, & \tau_{N_0} + 1 \leq t \leq n, \end{cases} \tag{4.11}$$

where the probability distributions $F_0, F_1, \dots, F_{N_0}$ differ on a set of non-zero measure.

4.3.1 Construction of a single change-point location estimator via cumulative sum process in embedded spaces

The starting point of many change-point detection procedures rest upon the so-called cumulative sum process. In this subsection, we illustrate the idea behind the construction of our proposed test statistic in Section 4.3.2 for estimation of a single change-point location. The idea essentially rests upon the construction of a cumulative sum process in embedded spaces.

[Equivalent characterization of spaces of negative type] A metric space $(\mathcal{X}, K)$ is of negative type if and only if there is a Hilbert space $\mathcal{H}$ and an embedding $\phi : \mathcal{X} \to \mathcal{H}$ such that $K(x, x') =$

$\|\phi(x) - \phi(x')\|_{\mathcal{H}}^2$ for all $x, x' \in \mathcal{X}$, where $\| \cdot \|_{\mathcal{H}}$ is the norm associated with $\mathcal{H}$ (see Section 3 in Lyons (2013)).

It is well known that $\mathbb{R}^p$ equipped with the usual Euclidean distance is a separable Hilbert space, and therefore has strong negative type (by Theorem 3.16 in Lyons (2013)). This combined with Result 4.3.1 ensures the existence of an embedding $\phi : \mathbb{R}^p \to \mathcal{H}$ for some Hilbert space $\mathcal{H}$ such that

$$\|x - x'\|_p \;=\; \|\phi(x) - \phi(x')\|_{\mathcal{H}}^2 \;=\; \langle\, \phi(x) - \phi(x')\,,\, \phi(x) - \phi(x')\,\rangle_{\mathcal{H}}\,, \tag{4.12}$$

where $x, x' \in \mathbb{R}^p$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product associated with $\mathcal{H}$. Therefore we get from (4.12)

$$\langle\, \phi(x)\,,\, \phi(x')\,\rangle_{\mathcal{H}} \;=\; 2^{-1}\left(\|x\| \,+\, \|x'\| \,-\, \|x - x'\|\right) \;=:\; l(x, x')\,. \tag{4.13}$$

We define the cumulative sum process in the embedded space as

$$\begin{aligned} S_k \;&=\; \frac{1}{\sqrt{n}}\sum_{t=1}^{k}\left(\phi(X_t) - \bar{\phi}\right) \;=\; \frac{1}{\sqrt{n}}\Big(\sum_{t=1}^{k}\phi(X_t) \,-\, \frac{k}{n}\sum_{t=1}^{k}\phi(X_t) \,-\, \frac{k}{n}\sum_{t=k+1}^{n}\phi(X_t)\Big) \\ &=\; \frac{(n-k)k}{n^{3/2}}\,\Big(\frac{1}{k}\sum_{t=1}^{k}\phi(X_t) \,-\, \frac{1}{n-k}\sum_{t=k+1}^{n}\phi(X_t)\Big) \end{aligned} \tag{4.14}$$

for $1 \le k \le n$, where $\bar{\phi} = \frac{1}{n}\sum_{t=1}^{n}\phi(X_t)$. The squared norm of S_k induced by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is given by

$$\begin{aligned} \frac{n^3}{(n-k)^2\,k^2}\,\|S_k\|^2 \;&=\; \frac{1}{k^2}\sum_{t,t'=1}^{k} l(X_t, X_{t'}) \,+\, \frac{1}{(n-k)^2}\sum_{t,t'=k+1}^{n} l(X_t, X_{t'}) \,-\, \frac{2}{k(n-k)}\sum_{t=1}^{k}\sum_{t'=k+1}^{n} l(X_t, X_{t'}) \\ =\; &\frac{1}{k(n-k)}\sum_{t=1}^{k}\sum_{t'=k+1}^{n}\|X_t - X_{t'}\| \,-\, \frac{1}{2k^2}\sum_{t,t'=1}^{k}\|X_t - X_{t'}\| \,-\, \frac{1}{2(n-k)^2}\sum_{t,t'=k+1}^{n}\|X_t - X_{t'}\|\,, \end{aligned} \tag{4.15}$$

which essentially follows from (4.13). If there is a single change-point in the sequence of data observations, a natural statistic to consider is the maximizer of the cumulative sum statistic $\|S_k\|^2$

over $1 \leq k \leq n$, viz.

$$\begin{aligned} V_n \;=\; \max_{1\leq k\leq n} \|S_k\|^2 \;=\; \max_{1\leq k\leq n} \frac{(n-k)^2k^2}{n^3} \Big(& \frac{1}{k(n-k)} \sum_{t=1}^{k} \sum_{t'=k+1}^{n} \|X_t - X_{t'}\| \;-\; \frac{1}{2k^2} \sum_{t,t'=1}^{k} \|X_t - X_{t'}\| \\ & - \frac{1}{2(n-k)^2} \sum_{t,t'=k+1}^{n} \|X_t - X_{t'}\| \Big) \,, \end{aligned} \tag{4.16}$$

which gives a candidate change-point location that needs to be tested against a certain threshold. A U-statistic version of the statistic in (4.16) is given by

$$U_n \;=\; \max_{1\leq k\leq n} \frac{(n-k)^2k^2}{2n^3}\, E_{1,n,k}\,, \tag{4.17}$$

where $E_{1,n,k}$ is the U-statistic type estimator of the Euclidean energy distance between the two samples $X_{1:k}$ and $X_{(k+1):n}$ (with d as the Euclidean distance in (4.6)). We want to point out to the reader that U_n constructed as above looks quite similar to the statistic considered by Matteson and James (2014) for estimation of a single change-point location, slightly differing in the scaling factor.

Consequently the single change-point location can be estimated as $\widehat{\tau}_{0U} := \mathrm{argmax}_{1\leq k\leq n}\, \frac{(n-k)^2k^2}{2n^3}\, E_{1,n,k}$. The statistical significance of the candidate change-point location $\widehat{\tau}_{0U}$ remains to be tested, which requires rigorously deriving the null distribution of U_n.

4.3.2 Estimation of a single change-point location

The setup considered in Section 4.3.1 for the construction of the single change-point location estimator U_n is essentially low dimensional, i.e., when the dimension of the observations is much smaller than the sample size. The motivation of this work is to develop a nonparametric methodology for detection of change-point locations when we have a sequence of high-dimensional observations, i.e., when p can by far exceed n. Our approach essentially rests upon testing for homogeneity of two high-dimensional distributions based upon the modifications to the usual Euclidean energy distance discussed in Section 4.2.2.

Building upon the insights from Section 4.3.1, we propose the following statistic :

$$M_n \;:=\; \max_{4\,\leq k\,\leq n-4}\, T(1,n\,;\,k) \;=\; \max_{4\,\leq k\,\leq n-4}\, \frac{(n-k)k}{n^2}\, T_{1,n,k}\ , \tag{4.18}$$

where $T_{1,n,k}$ is the two-sample statistic stated in (4.9), computed based on the two samples $X_{1:k}$ and $X_{(k+1):n}$, and $T(1,n\,;\,k) := \frac{(n-k)k}{n^2}\, T_{1,n,k}$. M_n essentially estimates a candidate single change-point location as

$$\hat{\tau_0} \;:=\; \operatorname*{argmax}_{4\,\leq k\,\leq n-4}\; T(1,n\,;\,k)\ . \tag{4.19}$$

The statistical significance of the estimated change-point location $\hat{\tau_0}$ needs to be tested, which necessitates determining the null distribution of M_n. The key theoretical innovation of this paper is to rigorously derive the asymptotic null distribution of M_n as $n, p \to \infty$, with n growing at a smaller rate compared to p. The intrinsic difficulty is to derive a uniform weak convergence result for the stochastic process $T_{1,n,k}$, as clearly a pointwise weak convergence result won't suffice.

Towards that end, we begin with introducing some technical definitions. Define
$\tau^2 := \mathbb{E}\, K(X,X')^2$.

ASSUMPTION **4.3.1**. *There exist constants a and A such that uniformly over p*

$$0 < a \leq \inf_{1\leq i\leq p} \mathbb{E}\,|X_i - X_i'| \leq \sup_{1\leq i\leq p} \mathbb{E}\,|X_i - X_i'| \leq A < \infty\,.$$

Under Assumption 4.3.1, it is easy to see that $\tau \asymp p^{1/2}$. The following proposition (Proposition 4.1 in Chakraborty and Zhang (2019)) presents an expansion formula for the distance metric K when the dimension is high, which plays a key role in the theoretical analysis.

PROPOSITION **4.3.1**. *Under Assumption 4.3.1, we have*

$$\frac{K(X,X')}{\tau} = 1 + \frac{1}{2}L(X,X') + R(X,X')\,,$$

where $L(X,X') := \frac{K^2(X,X')-\tau^2}{\tau^2}$ is the leading term and $R(X,X')$ is the remainder term. In addition, if $L(X,X')$ is a $o_p(1)$ random variable as $p \to \infty$, then $R(X,X') = O_p\left(L^2(X,X')\right)$.

Define $H(X_k, X_l) := \frac{1}{\tau}\sum_{i=1}^{p} d_{kl}(i)$ for $1 \le k, l \le n$, where

$$d_{kl}(i) := |X_{ki} - X_{li}| - \mathbb{E}\left[|X_{ki} - X_{li}|\,\middle|X_{ki}\right] - \mathbb{E}\left[|X_{ki} - X_{li}|\,\middle|X_{li}\right] + \mathbb{E}\left[\,|X_{ki} - X_{li}|\,\right]$$

is the double-centered distance between X_{ki} and X_{li}.

Assumption **4.3.2**. *As $n, p \to \infty$,*

$$\frac{1}{n^2}\frac{\mathbb{E}\left[H^4(X,X')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} = o(1), \quad \frac{1}{n}\frac{\mathbb{E}\left[H^2(X,X'')\,H^2(X',X'')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} = o(1),$$
$$\frac{\mathbb{E}\left[H(X,X'')\,H(X',X'')\,H(X,X''')\,H(X',X''')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} = o(1).$$

Remark **4.3.1**. *We refer the reader to Section 2.2 in Zhang et al. (2018) for an illustration of Assumption 4.3.2.*

Assumption **4.3.3**. *Suppose $\mathbb{E}\left[L^2(X,X')\right] = O(\alpha_p^2)$ where α_p is a positive real sequence such that $\tau\alpha_p^2 = o(1)$ as $p \to \infty$. Further assume that as $n, p \to \infty$,*

$$\frac{n^4\,\tau^4\,\mathbb{E}\left[R^4(X,X')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} = o(1)\,.$$

Remark **4.3.2**. *We refer the reader to Remark 4.1 in Chakraborty and Zhang (2019) which illustrates some sufficient conditions under which $\alpha_p = O(\frac{1}{\sqrt{p}})$ and consequently $\tau\alpha_p^2 = o(1)$ holds, as $\tau \asymp p^{1/2}$. In similar lines of Remark D.1 in the supplementary materials of their paper, it can be argued that $\mathbb{E}\left[R^4(X,X')\right] = O\left(\frac{1}{p^4}\right)$. Further with a mild assumption that $\sigma^2 := \lim_{p\to\infty} \mathbb{E}\left[H^2(X_k,X_l)\right]$, we have $\mathbb{E}\left[H^2(X,X')\right] \asymp 1$. Combining all the above, it is easy to verify that $\frac{n^4\,\tau^4\,\mathbb{E}\left[R^4(X,X')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} = o(1)$ holds provided $n = o(p^{1/2})$.*

The following theorem establishes a uniform weak convergence result of the stochastic process $\{T_{1,n,\lfloor nr\rfloor}\}_{r\in[0,1]}$ which plays a key role in deriving the limiting null distribution of M_n as $n, p \to$

∞.

THEOREM **9**. *Under Assumptions 4.3.2 and 4.3.3, as $n, p \to \infty$,*

$$\left\{T_{1,n,\lfloor nr \rfloor}\right\}_{r\in[0,1]} \xrightarrow{d} G_0 \qquad \textit{in} \;\; L^{\infty}\left([0,1]\right) \,,$$

where $G_0(r) := Q(0,1) - \frac{1}{r}\,Q(0,r) - \frac{1}{1-r}\,Q(r,1)$ for $r \in (0,1)$ and zero otherwise, and Q is a centered gaussian process with covariance function given by

$$\textit{cov}\left(Q(a_1,b_1)\,,\,Q(a_2,b_2)\right) \;=\; \left(b_1 \wedge b_2 \,-\, a_1 \vee a_2\right)^2 \,\mathbb{1}\left(b_1 \wedge b_2 > a_1 \vee a_2\right).$$

In particular, $var\,\left(Q(a,b)\right) = (b-a)^2\,\mathbb{1}(b > a)$.

The proof of this theorem is non-trivial, requiring the finite dimensional weak convergence and stochastic equicontinuity of the process $\{T_{1,n,\lfloor nr \rfloor}\}_{r\in[0,1]}$ to be established (see Theorem 10.2 in Pollard et al. (1990)). Because of its extremely long and technical nature, we relegate it to the supplementary materials.

Theorem B.1 in the supplementary materials of Chakraborty and Zhang (2019) essentially proves that for fixed $r \in (0,1)$, $T_{1,n,\lfloor nr \rfloor} \xrightarrow{d} N(0,1)$ as $n, p \to \infty$, under the same Assumptions 4.3.2 and 4.3.3. Note that in Theorem 9, for fixed $r \in (0,1)$, $G_0(r)$ has a gaussian distribution with zero mean. From the covariance structure of the gaussian process Q given in Theorem 9, it is not hard to verify that var $(G_0(r)) = 1$. This illustrates that the uniform weak convergence result established in Theorem 9 in fact generalizes the pointwise weak convergence result proven in Theorem B.1 in Chakraborty and Zhang (2019).

As a consequence of Theorem 9, we derive the limiting null distribution of M_n, which serves as the main theoretical innovation of the paper.

THEOREM **10**. *Under Assumptions 4.3.2 and 4.3.3, as $n, p \to \infty$, $M_n \xrightarrow{d} \sup\limits_{r\in(0,1)} r\,(1-r)\,G_0(r)$.*

Theorem 10 essentially follows from Theorem 9 and continuous mapping theorem. One thing to be noted is that the limiting null distribution is pivotal in nature and the quantiles of the limiting distribution can be approximated via a large number of Monte Carlo simulations.

With the limiting null distribution of M_n being rigorously established, we now present in Algorithm 1 the pseudocode of the procedure to test for H_0 against the single change-point alternative H_1. We use a permutation procedure to approximate the quantiles of the distribution of M_n aiming to achieve more accurate results.

Algorithm 1 Single change-point detection

```
Input : $\mathbb{R}^p$-valued observations $\{X_1, \dots, X_n\}$; level of significance $\alpha \in (0,1)$; number of permutation replicates $B$.
Compute the value of the test statistic $M_n$ and the candidate change-point location $\hat{\tau}_0$.
for $j = 1, 2, \dots, B$ do
    Generate a random permutation of the observations $\{X_1, \dots, X_n\}$.
    Compute the value of the test statistic for the permuted data, call it $M_n^j$.
end for
Compute $M_\alpha$, the $100(1-\alpha)^{th}$ percentile of $\{M_n^1, \dots, M_n^B\}$.
if $M_n > M_\alpha$ then
    Reject $H_0$ at level $\alpha$.
    Return $\hat{\tau}_0$ as the estimated change-point.
end if
```

4.3.3 Recursive estimation of multiple change-point locations

In practice, both the number and locations of change-points are unknown and need to be estimated. We need a ‘greedy’ procedure to sequentially detect multiple change-point locations, with each stage relying on the previously detected change-points, which are never re-visited. We combine our proposed test statistic M_n with the Wild Binary Segmentation (WBS) procedure proposed by Fryzlewicz (2014) to recursively estimate and test for the significance of multiple change-point locations.

The main idea is quite simple. In the beginning, instead of computing the statistic M_n over the entire sample $\{X_1, \dots, X_n\}$, we randomly draw (hence the term ‘wild’) M sub-samples $\{X_{s_m}, \dots, X_{e_m}\}$, $1 \leq m \leq M$, where s_m, e_m are integers satisfying $1 \leq s_m \leq n-7$ and $s_m + 7 \leq e_m \leq n$. We compute the statistic $T(s_m, e_m\,;\,b) = \frac{(e_m-b)(b-s_m+1)}{(e_m-s_m+1)^2}\ T_{s_m,e_m,b}$ for each sub-sample with b ranging over $\{s_m + 3, \dots, e_m - 4\}$. We require $s_m + 7 \le e_m$ to ensure there are $e_m - s_m + 1 \ge 8$ observations in the sub-sample $\{X_{s_m}, \dots, X_{e_m}\}$. We choose the largest maximizer over all the sub-samples to be the first change-point candidate to be tested against a certain threshold. We determine that threshold using a permutation procedure with B replicates. If the candidate change-point location turns out to be statistically significant, the same procedure is then repeated to the left and right of it. The recursive search quits a bisected sub-interval if either it doesn't contain at least 8 observations, or, if no further significant change-point locations are detected within that sub-interval. We illustrate in Algorithm 2 the pseudocode of the WBS procedure for detecting significant multiple change-point locations within a generic interval (s, e).

Algorithm 2 WBS procedure for multiple change-point detection

```
function WBS($s, e$)
    if ($e - s < 7$) then
        STOP;
    else
        $\mathcal{M}_{s,e} := \{(s_m, e_m), m = 1, \dots, M : s \le s_m \le e_m - 7 \le e - 7\}$.
        $(m_0, b_0) := \operatorname*{argmax}_{\substack{m \in \mathcal{M}_{s,e} \\ b = s_m+3, \dots, e_m-4}} T(s_m, e_m\,;\, b)$.
        for $j = 1, 2, \dots, B$ do
            Generate a random permutation of the observations $\{X_s, \dots, X_e\}$.
            Compute $T^j := \max_{\substack{m \in \mathcal{M}_{s,e} \\ b = s_m+3, \dots, e_m-4}} T(s_m, e_m\,;\, b)$.
        end for
        Compute $\zeta_\alpha$, the $100(1 - \alpha)^{th}$ percentile of $\{T^1, \dots, T^B\}$.
        if ($T(s_{m_0}, e_{m_0}\,;\, b_0) > \zeta_\alpha$) then
            Add $b_0$ to the set of estimated change-points.
            WBS($s, b_0$)
            WBS($b_0 + 1, e$)
        end if
    end if
end function
```

4.4 Numerical studies

4.4.1 Simulation studies

In this subsection, we examine the finite sample performance of our proposed methodology for multiple change-point detection via simulation studies. We first consider the following examples of the single change-point alternative.

EXAMPLE **4.4.1**. *(No change point)*

1. $X_k \sim N(0, I_p)$ *for* $1 \leq k \leq n$.

2. $X_k \sim N(0, \Sigma)$ *for* $1 \leq k \leq n$*, where* $\Sigma = (\sigma_{ij})_{i,j=1}^p$ *with* $\sigma_{ij} = 0.7^{|i-j|}$.

EXAMPLE **4.4.2**. *(Single change in mean)*

1. $X_k \sim N(0, I_p)$ *for* $1 \leq k \leq \lfloor n/2 \rfloor$ *and* $X_k \sim N(\mu, I_p)$ *for* $\lfloor n/2 \rfloor + 1 \leq k \leq n$*, where* $\mu = (0.6, \ldots, 0.6) \in \mathbb{R}^p$.

2. $X_k \sim N(0, \Sigma)$ *for* $1 \leq k \leq \lfloor n/2 \rfloor$ *and* $X_k \sim N(\mu, \Sigma)$ *for* $\lfloor n/2 \rfloor + 1 \leq k \leq n$*, where* $\Sigma = (\sigma_{ij})_{i,j=1}^p$ *with* $\sigma_{ij} = 0.7^{|i-j|}$*, and* $\mu = (0.6, \ldots, 0.6) \in \mathbb{R}^p$.

EXAMPLE **4.4.3**. *(Single change in distribution)*

1. $X_k \sim N(\mu, I_p)$ *with* $\mu = (1, \ldots, 1) \in \mathbb{R}^p$ *for* $1 \leq k \leq \lfloor n/2 \rfloor$ *and* $X_{ki} \overset{ind}{\sim}$ *Exponential* (1) *for* $i = 1, \ldots, p$ *and* $\lfloor n/2 \rfloor + 1 \leq k \leq n$.

2. $X_k = \underbrace{(X_{k1}, \ldots, X_{kp})}_{\overset{i.i.d.}{\sim} Poisson(1)} - 1$ *for* $1 \leq k \leq \lfloor n/2 \rfloor$ *and* $X_k = (X_{k1}, \ldots, X_{k\lfloor \beta p \rfloor}, X_{k(\lfloor \beta p \rfloor + 1)}, \ldots, X_{kp})$ *where* $X_{k1}, \ldots, X_{k\lfloor \beta p \rfloor} \overset{i.i.d.}{\sim}$ *Poisson* $(1) - 1$*, and* $X_{k(\lfloor \beta p \rfloor + 1)}, \ldots, X_{kp} \overset{i.i.d.}{\sim}$ *Rademacher* (0.5) *for* $\lfloor n/2 \rfloor + 1 \leq k \leq n$.

3. $X_k = R^{1/2} Z_{1k}$ *for* $1 \leq k \leq \lfloor n/2 \rfloor$ *and* $X_k = R^{1/2} Z_{2k}$ *for* $\lfloor n/2 \rfloor + 1 \leq k \leq n$*, where* $R = (r_{ij})_{i,j=1}^p$ *with* $r_{ii} = 1$ *for* $i = 1, \ldots, p$*,* $r_{ij} = 0.25$ *if* $1 \leq |i - j| \leq 2$ *and* $r_{ij} = 0$ *otherwise,* $Z_{1k} \sim N(0, I_p)$ *and* $Z_{2k} = \underbrace{(Z_{2k1}, \ldots, Z_{2kp})}_{\overset{i.i.d.}{\sim} Exponential(1)} - 1$.

We try $n = 100$, $p = 100, 200$ and $\beta = 1/2$. We implement Algorithm 1 with $B = 199$ permutation replicates and a significance level of $\alpha = 0.05$. We cluster the observations based on the estimated significant change-point location and compute the Adjusted Rand Index (ARI) (Morey and Agresti (1984)). The ARI is a positive value between 0 and 1. The ARI value is 0 when there is no change-point estimated when there does exist one (or more), or there is no change-point but the method estimates one (or more) change-point locations. The ARI value is 1 when the estimation is perfect. Higher the value of ARI, more accurate is the estimation of the change-point location. We consider 100 simulations of each of the above examples, for each of which we compute the ARI value and report it in the table below. We compare our test with

- the E-Divisive procedure proposed by Matteson and James (2014) with $R = 199$ random permutations (using the 'ecp' R package); (denote by MJ)
- the test based on the graph-based original scan statistic proposed by Chen and Zhang (2015) with 199 permutations (using the 'gSeg' R package); (denote by CZ)
- the max-type edge-count test proposed by Chu and Chen (2019) with 199 permutations (using the 'gSeg' R package); (denote by CC)
- the test proposed by Wang et al. (2019); (denote by WVS) and
- the INSPECT procedure proposed by Wang and Samworth (2018) (using the 'InspectChangepoint' R package) (denote by WS).

Although the methodologies proposed by Wang et al. (2019) and Wang and Samworth (2018) are aimed at detecting a mean shift for high dimensional data, we compare our method with theirs to illustrate that our method can capture inhomogeneities among a sequence of high-dimensional observations beyond the first moment. The results from Table 4.1 indicate that almost all the methods perform nearly equally good when there is no true change-point or when there is a mean shift. When there is no true change-point, the procedure proposed by Wang and Samworth (2018) still detects one, leading to a zero ARI value.

Table 4.1: Comparison of average ARI values for different methods over 100 simulations.

		n	p	Our test	MJ	CC	CZ	WVS	WS
Ex 4.4.1	(1)	100	100	0.98	0.97	0.97	0.97	0.90	0.00
	(1)	100	200	0.97	0.98	0.96	0.96	0.97	0.00
	(2)	100	100	0.93	0.97	0.91	0.92	0.94	0.00
	(2)	100	200	0.97	0.97	0.95	0.98	0.93	0.00
Ex 4.4.2	(1)	100	100	1	1	0.997	0.999	0.963	1
	(1)	100	200	1	1	0.999	0.999	0.969	1
	(2)	100	100	0.984	0.986	0.867	0.946	0.949	0.981
	(2)	100	200	0.996	0.996	0.978	0.983	0.962	0.993
Ex 4.4.3	(1)	100	100	0.993	0.014	0.004	0.027	0.052	0.390
	(1)	100	200	1	0.030	0.007	0.037	0.035	0.414
	(2)	100	100	0.999	0.034	0.001	0.059	0.063	0.468
	(2)	100	200	1	0.032	0.001	0.055	0.067	0.502
	(3)	100	100	0.978	0.024	0.021	0.065	0.074	0.402
	(3)	100	200	0.992	0.029	0.006	0.040	0.044	0.363

Most interestingly, when there is a change in distribution among the sequence of the high-dimensional observations, our method performs way better than any of the other competitors in terms of accurately estimating the single change-point location. That it clearly beats the E-Divisive procedure, is quite expected as the Euclidean energy distance fails to capture any inhomogeneity between two high-dimensional distributions beyond the first two moments.

The following examples are illustrate the performance of our methodology in case of two change-points alternative.

EXAMPLE **4.4.4**. *(Two changes in mean)*

1. *$X_k \sim N(0, I_p)$ for $1 \leq k \leq \lfloor n/3 \rfloor$ and $2\lfloor n/3 \rfloor + 1 \leq k \leq n$, and $X_k \sim N(\mu, I_p)$ for $\lfloor n/3 \rfloor + 1 \leq k \leq 2\lfloor n/3 \rfloor$, where $\mu = (0.6, \ldots, 0.6) \in \mathbb{R}^p$.*

2. *$X_k \sim N(0, \Sigma)$ for $1 \leq k \leq \lfloor n/3 \rfloor$ and $2\lfloor n/3 \rfloor + 1 \leq k \leq n$, and $X_k \sim N(\mu, \Sigma)$ for $\lfloor n/3 \rfloor + 1 \leq k \leq 2\lfloor n/3 \rfloor$, where $\Sigma = (\sigma_{ij})_{i,j=1}^{p}$ with $\sigma_{ij} = 0.7^{|i-j|}$ and $\mu = (0.6, \ldots, 0.6) \in \mathbb{R}^p$.*

Table 4.2: Comparison of average ARI values for different methods over 100 simulations.

		n	p	Our test	MJ	CC	CZ	WS
Ex 4.4.4	(1)	100	100	0.991	1	0.975	0.960	0.916
	(2)	100	100	0.962	0.978	0.747	0.885	0.643

4.4.2 Real data illustration

We consider the daily closed stock prices of $p = 72$ companies under the Consumer Defensive sector, listed under the NYSE and NASDAQ stock exchanges, on the first dates of each month during the time period between January 1, 2005 and December 31, 2010. The data has been downloaded from Yahoo Finance via the R package 'quantmod'. At each time t, denote the closed stock prices of these companies by $X_t = (X_{1t}, \dots, X_{pt})$ for $1 \leq t \leq 72$. We consider the stock returns $S_t^X = (S_{1t}^X, \dots, S_{pt}^X)$ for $1 \leq t \leq 71$, where $S_{it}^X = \log \frac{X_{i,t+1}}{X_{it}}$ for $1 \leq i \leq p$.

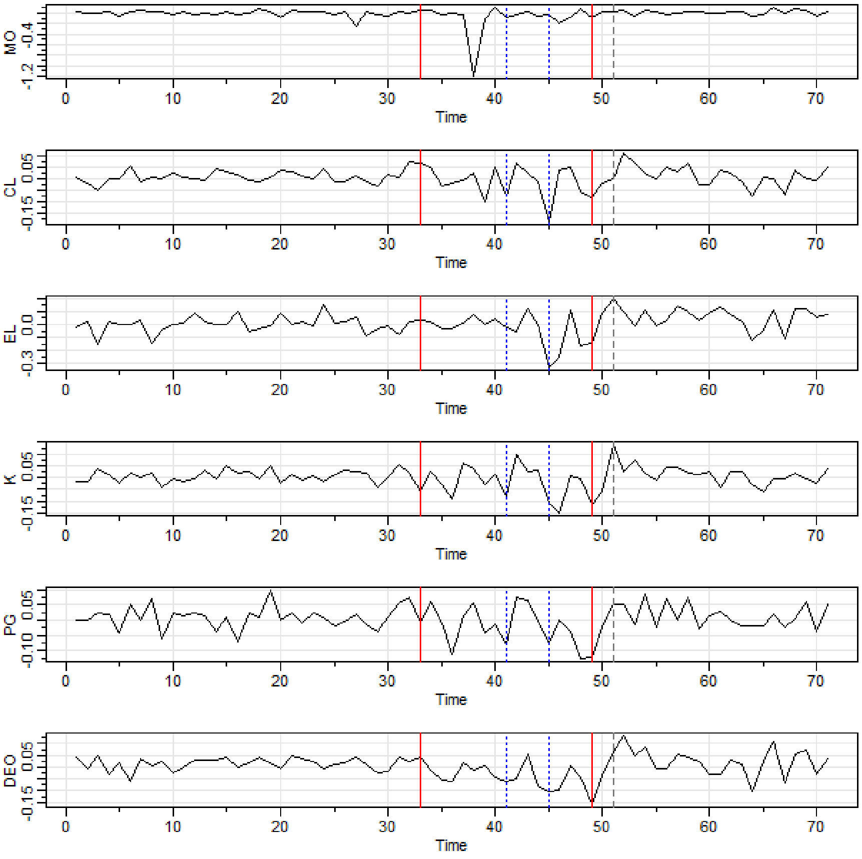
According to the US National Bureau of Economic Research, the Great Recession began in the United States in December 2007. The government responded with an unprecedented \$700 billion bank bailout in October 2008 and \$787 billion fiscal stimulus package in February 2009 to save existing jobs, provide temporary relief programs for those most affected by the recession, invest in infrastructure, education, health and renewable energy, etc. The recession officially lasted till June 2009, thus extending over 19 months.

When a recession or an economic slowdown occurs, markets tend to become volatile, prompting investors to sell stocks. Although all industrial sectors are susceptible to economic changes, some are less sensitive or more resistant to recessions (for example, Consumer Defensive, Utility or Healthcare sectors) compared to some others (for example, Real Estate, Finance, Oil and Gas, Automobiles, etc.). The goal is to consider stock returns of companies under a sector that is known to perform relatively well even when a recession hits the market, and see if our proposed methodology and the other state-of-the-art methods can detect change-points in the stock returns data.

When applied on the stock returns dataset (with $n = 71$ and $p = 72$) for the Consumer Defensive sector :

- our procedure detects two change-points, viz. September 1, 2007 and January 1, 2009, which seems quite reasonable given the historical sequence of eventualities;
- the E-Divisive procedure by Matteson and James (2014) fails to detect any change-point over that time period;
- the test based on the graph-based original scan statistic proposed by Chen and Zhang (2015) detects only one change-point, viz. March 1, 2009;
- the max-type edge-count test proposed by Chu and Chen (2019) detects two change-points, viz. May 1, 2008 and September 1, 2008; and
- the methodology proposed by Wang and Samworth (2018) detects as many as 18 change-points over the aforesaid period of time.

Figure 4.1: Time series plots of the stock returns for six companies under the Consumer Defensive sector. The solid red lines represent the change-point locations detected by our methodology. The dotted blue and gray lines represent the change-point locations detected by the procedures proposed by Chu and Chen (2019) and Chen and Zhang (2015), respectively.



5. SUMMARY AND CONCLUSIONS

To summarize, measuring and testing for independence and homogeneity of distributions are some fundamental problems in statistics, finding applications in a wide variety of areas. The first work (Chapter 2) aims at quantifying and testing for joint independence among $d \geq 2$ random vectors, which is a quite different and more ambitious task than testing for pairwise independence of a collection of random vectors. The second work (Chapter 3) upholds the limitations of the classical distance and kernel-based homogeneity and dependence metrics for growing dimensions, and proposes a new class of homogeneity/dependence metrics to perform two-sample/independence testing in the high-dimensional setup. The third work (Chapter 4) proposes a methodology to detect an unknown number of change-points in an independent sequence of high-dimensional time-ordered observations. The key idea essentially rests upon nonparametric testing for equality of two high-dimensional distributions.

5.1 Future research

There are several intriguing problems that are worthy of investigation in the future.

- The direct implementation of JdCov has a time complexity of the order of $O(n^2)$, where n is the sample size. This quadratic computational cost might be prohibitive in many applications with large-scale datasets. One possible direction is to come up with a fast computational algorithm to compute JdCov.

- Another potential direction for future research might be to develop a computationally and statistically efficient algorithm to learn the correct causal structure among a collection of random variables based on nonparametric tests for joint independence.

- Extension of our methodology for multiple change-point detection for a weakly dependent high-dimensional time series seems to be of obvious interest, as temporal dependence is natural to expect in many practical applications. The problem, though quite intriguing, seems absolutely non-immediate and non-trivial from both methodological and theoretical perspectives because of the additional complexity brought in by the temporal dependence.

- It would be interesting to develop theoretical consistency results for the wild binary segmentation procedure we implemented for multiple change-point detection, similar to Theorem 3.2 in Fryzlewicz (2014). This again seems non-trivial and challenging, and we leave it as a topic for future research.

REFERENCES

[1] ANDREWS, D.W.K.(1992). Generic Uniform Convergence. *Econometric Theory*, *8*(2) 241-257.

[2] ARLOT, S., CELISSE, A. and HARCHAOUI, Z. (2019). A Kernel Multiple Change-point Algorithm via Model Selection. arXiv:1202.3878v3.

[3] BACH, F. R. and JORDAN, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, *3* 1-48.

[4] BARINGHAUS, L. and FRANZ, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, *88*(1), 190-206.

[5] BERGSMA, W. and DASSIOS, A. (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, *20*(2) 1006-1028.

[6] BIAU, G., BLEAKLEY, K. and MASON, D.M. (2016). Long signal change-point detection. *Electronic Journal of Statistics*, *10*(2), 2097-2123.

[7] BICKEL, P. J. (1969). A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case. *The Annals of Mathematical Statistics*, *40*(1) 1-23.

[8] BÖTTCHER, B. (2017). Dependence structures - estimation and visualization using distance multivariance. arxiv:1712.06532.

[9] BRADLEY, R. C. (2005). Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probability Surveys*, *2* 107-144.

[10] BÜHLMANN, P., PETERS, J. and ERNEST, J. (2014). CAM : Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, *42*(6) 2526-2556.

[11] Chakraborty, S. and Zhang, X. (2019). Distance Metrics for Measuring Joint Dependence with Application to Causal Inference. *Journal of the American Statistical Association*, *114*(528), 1638-1650.

[12] Chakraborty, S., and Zhang, X. (2019). A New Framework for Distance and Kernel-based Metrics in High Dimensions. arXiv:1909.13469.

[13] Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *Annals of Statistics*, *43*(1), 139-176.

[14] Chen, H. and Friedman, J. H. (2017). A New Graph-Based Two-Sample Test for Multivariate and Object Data. *Journal of the American Statistical Association*, *112*(517), 397-409.

[15] Chu, L. and Chen, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. *Annals of Statistics*, *47*(1), 382-414.

[16] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

[17] Curtis, R., Xiang, J., Parikh, A., Kinnaird, P. and Xing, E.P. (2012). Enabling dynamic network analysis through visualization in TVNViewer. *BMC Bioinformatics*, *13*(204).

[18] Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises Tests. *The Annals of Mathematical Statistics*, *29*(3) 842-851.

[19] Dau, H. A., Keogh, E., Kamgar, K., Yeh, C. C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Chen, Y., Hu, B., Begum, N., Bagnall, A., Mueen, A. and Batista, G. (2018). The UCR Time Series Classification Archive. URL `https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`.

[20] David, H. T. (1958). A Three-Sample Kolmogorov-Smirnov Test. *The Annals of Mathematical Statistics*, *28*(4) 823-838.

[21] DAVIS, R. A., MATSUI, M., MIKOSCH, T. and WAN, P. (2016). Applications of distance correlation to time series, arXiv:1606.05481.

[22] DOUKHAN, P. and NEUMANN, M.H. (2008). The notion of ψ-weak dependence and its applications to bootstrapping time series. *Probability Surveys*, *5* 146-168.

[23] DUECK, J., EDELMANN, D., GNEITING, T. and RICHARDS, D. (2014). The affinely invariant distance correlation. *Bernoulli*, *20*(4) 2305-2330.

[24] EDELMANN, D., FOKIANOS, K. and PITSILLOU, M. (2018). An Updated Literature Review of Distance Correlation and its Applications to Time Series. arxiv:1710.01146.

[25] FANO, R. M. (1961). *Transmission of Information*. New York: M.I.T Press.

[26] FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *Annals of Statistics*, *7*(4) 697-717.

[27] FRYZLEWICZ, P.(2014). Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, *42*(6) 2243-2281.

[28] GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *Algorithmic Learning Theory, Springer-Verlag*, 63-77.

[29] GRETTON, A., FUKUMIZU, C. H. TEO., SONG, L., SCHÖLKOPF, B. and SMOLA, A. (2007). A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, *20* 585-592.

[30] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, *13* 723-773.

[31] HALL, P. and HEYDE, C. C. (1980). Martingale Limit Theory and Its Applications . *Academic press*.

[32] HUANG, C. and HUO, X. (2017). A statistically and numerically efficient independence test based on random projections and distance covariance, arXiv:1701.06054.

[33] HUO, X. and SZÉKELY, G. J. (2016). Fast computing for distance covariance. *Technometrics*, *58*(4) 435-446.

[34] JIN, Z. and MATTESON, D. S. (2017). Generalizing Distance Covariance to Measure and Test Multivariate Mutual Dependence. https://arxiv.org/abs/1709.02532.

[35] JOSSE, J. and HOLMES, S. (2014). Tests of independence and Beyond. arxiv:1307.7383.

[36] KIM, I., BALAKRISHNAN, S. and WASSERMAN, L. (2018). Robust multivariate nonparametric tests via projection-pursuit. arXiv:1803.00715.

[37] LANCASTER, H. O. (1969). *The Chi-Squared Distribution*. London: Wiley.

[38] LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*. Springer-Verlag.

[39] LI, J. (2018). Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika*, 105(3), 529-546.

[40] LI, Q. and RACINE, J. C. (2007). Nonparametric Econometrics : Theory and Practice . *Princeton University press*.

[41] LUNG-YUT-FONG, A., LÉVY-LEDUC, C. and CAPPÉ, O. (2015). Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Francaise de Statistique*, *156*(4), 133-162.

[42] LYONS, R. (2013). Distance covariance in metric spaces. *Annals of Probability*, *41*(5) 3284-3305.

[43] MAA, J. -F., PEARL, D. K. and BARTOSZYŃSKI, R. (1996). Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Annals of Statistics*, *24*(3) 1069-1074.

[44] MATTESON, D.S. and JAMES, N.A. (2014). A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data. *Journal of the American Statistical Association*, *109*(505), 334-345.

[45] MATTESON, D. S. and TSAY, R. S. (2017). Independent component analysis via distance covariance. *Journal of the American Statistical Association*, *112*(518), 623-637.

[46] MCCULLOH, I.(2009). Detecting Changes in a Dynamic Social Network. PhD thesis, Institute for Software Research, School of Computer Science, Carnegie Mellon University. CMU-ISR-09-104.

[47] MCGILL, W. J (1954). Multivariate information transmission. *Psychometrika*, *19*(4) 97-116.

[48] MOREY, L.C. and AGRESTI, A. (1984). The Measurement of Classification Agreement : An Adjustment to the Rand Statistic for Chance Agreement. *Educational and Psychological Measurement*, *44*(1), 33-37.

[49] NEUHAUS, G.(1977). Functional Limit Theorems for U-Statistics in the Degenerate Case. *Journal of Multivariate Analysis*, *7*, 424-439.

[50] PETERS, J., MOOIJ, J.M., JANZING, D. and SCHÖLKOPF, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, *15* 2009-2053.

[51] PFISTER, N., BÜHLMANN, P., SCHÖLKOPF, B. and PETERS, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society, Series B*, *80*(1) 5-31.

[52] PICARD, F., ROBIN, S., LAVIELLE, M., VAISSE, C. and DAUDIN, J.-J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, *6*(27).

[53] POLLARD, D.(1990). Empirical Processes: Theory and Applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, *2* i-iii+v+vii-viii+1-86.

[54] RABINER, L.R. and SCHÄFER, R.W. (2007). Introduction to Digital Speech Processing. *Foundations and Trends in Signal Processing*, *1*(1-2), 1-194.

[55] RAMDAS, A., REDDI, S. J., POCZOS, B., SINGH, A. and WASSERMAN, L. (2015a). Adaptivity and Computation-Statistics Tradeoffs for Kernel and Distance based High Dimensional Two Sample Testing. arXiv:1508.00655.

[56] RAMDAS, A., REDDI, S. J., POCZOS, B., SINGH, A. and WASSERMAN, L. (2015b). On the Decreasing Power of Kernel and Distance Based Nonparametric Hypothesis Tests in High Dimensions. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[57] READ, T. and CRESSIE. N. (1988). *Goodness-Of-Fit Statistics for Discrete Multivariate Analysis.* New York: Springer-Verlag.

[58] RESNICK, S. I. (1999). *A Probability Path.* Springer.

[59] ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42(1), 43.

[60] SCHILLING, M. F. (1986). Multivariate Two-Sample Tests Based on Nearest Neighbors. *Journal of the American Statistical Association* , *81*(395) 799-806.

[61] SEJDINOVIC, D., GRETTON, A. and Bergsma, W. (2013). A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems* (NIPS 26), 1124-1132.

[62] SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A. and FUKUMIZU, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, *41*(5) 2263-2291.

[63] SEN, P. K. (1977). Almost Sure Convergence of Generalized U-Statistics. *Annals of Probability*, *5*(2) 287-290.

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

[65] SERFLING, R. J. (1980). Approximation Theorems of Mathematical Statistics . *Wiley* , New York.

[66] SHANNON, C. E. and Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

[67] SHAO, X. and ZHANG, J. (2014). Martingale Difference Correlation and Its Use in High-Dimensional Variable Screening. *Journal of the American Statistical Association*, *109*(507) 1302-1318.

[68] SRIPERUMBUDUR, B., GRETTON, A., FUKUMIZU, K., SCHÖLKOPF, B. and LANCKRIET, G.R.G (2010). Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, *11* 1517-1561.

[69] STREITBERG, B. (1990). Lancaster interactions revisited. *Annals of Statistics*, *18*(4) 1878-1885.

[70] SZÉKELY, G. J. (2002). E-Statistics: the Energy of Statistical Samples. Technical report.

[71] SZÉKELY, G. J. and RIZZO, M. L.. (2004). Testing for equal distributions in high dimension. *InterStat*, *5*.

[72] SZÉKELY, G. J. and RIZZO, M. L.. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification*, *22* 151-183.

[73] SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing independence by correlation of distances. *Annals of Statistics*, *35*(6) 2769-2794.

[74] SZÉKELY, G. J. and RIZZO, M. L.. (2009). Brownian distance covariance. *Annals of Applied Statistics*, *3*(4) 1236-1265.

[75] SZÉKELY, G. J. and RIZZO, M. L. (2012). On the uniqueness of distance covariance. *Statistics and Probability Letters*, *82* 2278-2282.

[76] SZÉKELY, G. J. and RIZZO, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, *117* 193-213.

[77] SZÉKELY, G. J. and RIZZO, M. L.. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, *143* (8) 1249-1272 .

[78] SZÉKELY, G. J. and RIZZO, M. L. (2014). Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, *42*(6) 2382-2412.

[79] WANG, T. and SAMWORTH, R.J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society, Series B*, *80*(1), 57-83.

[80] WANG, R., VOLGUSHEV, S. and SHAO, X. (2019). Inference for Change Points in High Dimensional Data. arXiv:1905.08446.

[81] WANG, X., WENLIANG, P., HU, W., TIAN, Y. and ZHANG, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, *110*(512) 1726-1734.

[82] WOOD, S. N. and AUGUSTIN, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, *157* (2-3) 157-177.

[83] YAO, S., ZHANG, X. and SHAO, X. (2018). Testing mutual independence in high dimension via distance covariance. *Journal of the Royal Statistical Society, Series B*, to appear.

[84] ZHANG, X., YAO, S. and SHAO, X. (2018). Conditional Mean and Quantile Dependence Testing in High Dimension. *Annals of Statistics*, *46*(1) 219-246.

[85] ZHOU, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, *33*(3), 438-457.

[86] ZHU, L., XU, K., LI, R., and ZHONG, W. (2017). Projection correlation between two random vectors. *Biometrika*, *104*(4) 829-843.

[87] ZHU, C., YAO, S., ZHANG, X. and SHAO, X. (2020). Distance-based and RKHS-based Dependence Metrics in High-dimension. *Annals of Statistics*, to appear.

[88] ZOU, C., YIN, G., FENG, L. and WANG, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *Annals of Statistics*, *42*(3) 970-1002.

APPENDIX A

This is the Appendix for Chapter 2.

Proof of Lemma 2.2.1. By Lemma 1 of Székely et al. (2007), we have

$$\begin{aligned}RHS =& \int_{\mathbb{R}^p} \left\{ \mathbb{E}e^{\imath\langle t, X_i - X_i'\rangle} + e^{\imath\langle t, x-x'\rangle} - \mathbb{E}e^{\langle t, x - X_i'\rangle} - \mathbb{E}e^{\imath\langle t, X_i - x'\rangle} \right\} w_{p_i}(t)dt \\ =&\mathbb{E}\int_{\mathbb{R}^p} \Bigg\{ \cos(\langle t, X_i - X_i'\rangle) - 1 + \cos(\langle t, x - x'\rangle) - 1 + 1 - \cos(\langle t, x - X_i'\rangle) \\ &+ 1 - \cos(\langle t, X_i - x'\rangle) \Bigg\} w_{p_i}(t)dt + \imath \int_{\mathbb{R}} \mathbb{E}\Bigg\{ \sin(\langle t, X_i - X_i'\rangle) + \sin(\langle t, x - x'\rangle) \\ &- \sin(\langle t, x - X_i'\rangle) - \sin(\langle t, X_i - x'\rangle) \Bigg\} w_{p_i}(t)dt \\ =&\mathbb{E}|x - X_i'| + \mathbb{E}|X_i - x'| - |x - x'| - \mathbb{E}|X_i - X_i'| \;=\; U_i(x, x').\end{aligned}$$

Here we have used the fact that $\int_{\mathbb{R}}\{\sin(\langle t, X - X'\rangle) + \sin(\langle t, x - x'\rangle) - \sin(\langle t, x - X'\rangle) - \sin(\langle t, X - x'\rangle)\} w_{p_i}(t)dt = 0$. $\diamondsuit$

Proof of Proposition 2.2.1. To show (1), notice that for a_i, c_i and orthogonal transformations $A_i \in \mathbb{R}^{p_i \times p_i}$,

$$\mathbb{E}\prod_{i\in S} U_i(a_i + c_i A_i X_i, a_i + c_i A_i X_i') = \prod_{i\in S}|c_i|\,\mathbb{E}\prod_{i\in S} U_i(X_i, X_i'),$$

where $S \subset \{1, 2, \ldots, d\}$. The conclusion follows directly. $\diamondsuit$

Proof of Proposition 2.2.2. The proof is essentially similar to the proof of Lemma 1.2 in the supplementary material of Yao et al. (2018). It is easy to verify that

$$E\prod_{i=1}^{d} U_i(X_i, X_i') = C\int_{\mathbb{R}^d} |A|^2 \frac{dt_1}{t_1^2} \cdots \frac{dt_d}{t_d^2}\ ,$$

where C is some constant and

$$\begin{aligned} A = e^{-\frac{t_1^2+\cdots+t_d^2}{2}} \Big[& e^{-\rho t_1 t_2} + e^{-\rho t_1 t_3} + \dots \ \binom{d}{2} \text{ similar terms} \\ & - e^{-\rho t_1 t_2 - -\rho t_1 t_3 - \rho t_2 t_3} - e^{-\rho t_1 t_2 - \rho t_1 t_4 - \rho t_2 t_4} - \dots \ \binom{d}{3} \text{ similar terms} \\ & + e^{-\rho t_1 t_2 - -\rho t_1 t_3 - \rho t_1 t_4 - \rho t_2 t_3 - \rho t_2 t_4 - \rho t_3 t_4} + \dots \ \binom{d}{4} \text{ similar terms} \\ & + \ \dots \ - (d-1) \ \Big] . \end{aligned} \tag{A.1}$$

For example, if $d \geq 4$ and we use the Taylor's expansion $e^x = 1 + x + \frac{x^2}{2!} + \sum_{l=3}^{\infty} \frac{x^l}{l!}$, then keeping in mind the multinomial theorem

$$(a_1 + \cdots + a_q)^2 = \sum_{\substack{l_1,\ldots,l_q=0 \\ l_1+\cdots+l_q=2}}^{2} \frac{2!}{l_1! \ldots l_q!} \ , \ \ q \geq 2 \ ,$$

it is easy to check that the leading terms and their coefficients (upto some constants) are

Leading terms	Coefficients (upto some constants)
$t_i t_j$	$1 - \binom{d-2}{1} + \binom{d-2}{2} - \dots \quad (= 0 \text{ for } d > 2)$
$t_i^2 t_j^2$	$1 - \binom{d-2}{1} + \binom{d-2}{2} - \dots \quad (= 0 \text{ for } d > 2)$
$t_i^2 t_j t_k$	$-1 + \binom{d-3}{1} - \binom{d-3}{2} + \dots \quad (= 0 \text{ for } d > 3)$
$t_i t_j t_k t_l$	$1 - \binom{d-4}{1} + \binom{d-4}{2} - \dots \quad (= 1 \text{ if } d = 4, = 0 \text{ for } d > 4)$

.

To get a non-trivial upper bound for $E \prod_{i=1}^{d} U_i(X_i, X_i')$, we need to consider the Taylor's expansion $e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^k}{k!} + \sum_{l=k+1}^{\infty} \frac{x^l}{l!}$, when $d = 2k-1$ or $d = 2k$, $k \geq 2$, and the only leading term in the Taylor's expansion of (A.1) that would lead to a term with non-vanishing coefficient, is $\frac{x^k}{k!}$. To see this, note that when $d = 4$, i.e., $k = 2$, it is shown in Lemma 1.2 in the supplementary material of Yao et al. (2018) that the only non-vanishing term is $t_1 t_2 t_3 t_4$ (upto some constants). Likewise for $d = 5$ and 6, the only non-vanishing leading terms (upto some constants) are :

d	k	The only non-vanishing term (upto some constants)
5	3	$(t_it_j)^1(t_it_a)^1(t_lt_m)^1 = t_i^2t_jt_at_lt_m$
6	3	$(t_it_j)^1(t_at_l)^1(t_mt_n)^1 = t_it_jt_at_lt_mt_n$

$, i \neq j \neq a \neq l \neq m \neq n \, .$

In general when $d = 2k-1$, for $k \geq 2$, the only non-vanishing term (upto some constants) is $t_{i_1}^2 t_{i_2} \dots t_{i_d}$, where $(i_1, \dots, i_d)$ is any permutation of $(1, 2, \dots, d)$. Suppose P_d denotes the set of all possible permutations of $(1, 2, \dots, d)$. Then

$$\begin{aligned} E\prod_{i=1}^{d} U_i(X_i, X_i') &= c_0 \int_{\mathbb{R}^d} |e^{-\frac{t_1^2+\cdots+t_d^2}{2}} (\, c_1 \rho^k \sum_{(i_1,\dots,i_d)\in P_d} t_{i_1}^2 t_{i_2} \dots t_{i_d} + R\,)\,|^2 \, \frac{dt_1}{t_1^2} \cdots \frac{dt_d}{t_d^2} \\ &= A_0 + A_1 + A_2 + A_3 \ , \end{aligned}$$

where

$$A_0 \;=\; \tilde{c}_0 \, \rho^{2k} \sum_{(i_1,\dots,i_d)\in P_d} \int_{\mathbb{R}^d} e^{-(t_1^2+\cdots+t_d^2)} \, t_{i_1}^4 t_{i_2}^2 \dots t_{i_d}^2 \, \frac{dt_1}{t_1^2} \cdots \frac{dt_d}{t_d^2} \ ,$$

$$A_1 \;=\; \tilde{c}_1 \, |\rho|^k \sum_{(i_1,\dots,i_d)\in P_d} \int_{\mathbb{R}^d} e^{-(t_1^2+\cdots+t_d^2)} \, t_{i_1}^2 t_{i_2} \dots t_{i_d} \times R \, \frac{dt_1}{t_1^2} \cdots \frac{dt_d}{t_d^2} \ ,$$

$$A_2 \;=\; \tilde{c}_2 \, \rho^{2k} \sum_{(i_1,\dots,i_d)\in P_d} \int_{\mathbb{R}^d} e^{-(t_1^2+\cdots+t_d^2)} \, t_{i_1}^3 t_{i_2}^3 t_{i_3}^2 \dots t_{i_d}^2 \, \frac{dt_1}{t_1^2} \cdots \frac{dt_d}{t_d^2} \ ,$$

and

$$A_3 \;=\; \tilde{c}_3 \int_{\mathbb{R}^d} e^{-(t_1^2+\cdots+t_d^2)} \, \times R^2 \, \frac{dt_1}{t_1^2} \cdots \frac{dt_d}{t_d^2} \ ,$$

$c_0, c_1, \tilde{c}_0, \tilde{c}_1$, $\tilde{c}_2$ and $\tilde{c}_3$ being some constants and R being the remainder term from the Taylor's expansion. Following the similar arguments of Yao et al. (2018), it can be shown that

$$A_0 \;=\; O(|\rho|^{2k}) \ , \ A_1 \;=\; O(|\rho|^{2k+1}) \ , \ A_2 \;=\; O(|\rho|^{2k}) \text{ and } A_3 \;=\; O(|\rho|^{2k+2}) \ .$$

Thus for $d = 2k-1, k \geq 2$,

$$E \prod_{i=1}^{d} U_i(X_i, X_i') \;=\; O(|\rho|^{2k}) \;.$$

And when $d = 2k$, for $k \geq 2$, the only non-vanishing term (upto some constants) is $t_1 t_2 \ldots t_d$. Consequently

$$\begin{aligned} E \prod_{i=1}^{d} U_i(X_i, X_i') \;&=\; c_0' \int_{\mathbb{R}^d} |e^{-\frac{t_1^2+\cdots+t_d^2}{2}} (\, c_1' \rho^k t_1 t_2 \ldots t_d + R' \,)\,|^2 \, \frac{dt_1}{t_1^2} \ldots \frac{dt_d}{t_d^2} \\ &= A_0' + A_1' + A_2' \;, \end{aligned}$$

where

$$\begin{aligned} A_0' \;&=\; \tilde{c}_0' \, \rho^{2k} \int_{\mathbb{R}^d} e^{-(t_1^2+\cdots+t_d^2)} \; t_1^2 t_2^2 \ldots t_d^2 \, \frac{dt_1}{t_1^2} \ldots \frac{dt_d}{t_d^2} \;, \\ A_1' \;&=\; \tilde{c}_1' \, |\rho|^k \int_{\mathbb{R}^d} e^{-(t_1^2+\cdots+t_d^2)} \; t_1 t_2 \ldots t_d \times R' \, \frac{dt_1}{t_1^2} \ldots \frac{dt_d}{t_d^2} \;, \end{aligned}$$

and

$$A_2' \;=\; \tilde{c}_2' \int_{\mathbb{R}^d} e^{-(t_1^2+\cdots+t_d^2)} \;\times R'^2 \, \frac{dt_1}{t_1^2} \ldots \frac{dt_d}{t_d^2} \;,$$

$c_0', c_1', \tilde{c}_0', \tilde{c}_1'$ and $\tilde{c}_2'$ being some constants and R' being the remainder term from the Taylor's expansion. Again following the similar arguments of Yao et al. (2018), it can be shown that

$$A_0' \;=\; O(|\rho|^{2k}) \;,\; A_1' \;=\; O(|\rho|^{2k+1}) \text{ and } A_2' \;=\; O(|\rho|^{2k+2}) \;.$$

Thus for $d = 2k, k \geq 2$,

$$E \prod_{i=1}^{d} U_i(X_i, X_i') \;=\; O(|\rho|^{2k}) \;,$$

which completes the proof.

$\diamondsuit$

Proposition **A.0.1**. *(1)* $dCov^2(X_1, \ldots, X_d) \leq \mathbb{E}[\prod_{j=1}^{d} \min\{a_j(X_j), a_j(X_j')\}]$ *with* $a_j(x) =$

$\max\{\mathbb{E}|X_j - X_j'|, |\mathbb{E}|X_j - X_j'| - 2\mathbb{E}|x - X_j||\}$. *For any partition* $S_1 \cup S_2 = \{1, 2, \ldots, d\}$ *and* $S_1 \cap S_2 = \emptyset$*, we have* $dCov^2(X_1, \ldots, X_d) \leq \mathbb{E}[\prod_{i \in S_1} a_j(X_j)]\, \mathbb{E}[\prod_{i \in S_2} a_j(X_j)]$.

(2) $dCov^2(X_1, \ldots, X_d) \leq \prod_{i=1}^d \left\{\mathbb{E}[|U_i(X_i, X_i')|^d]\right\}^{1/d}$. *In particular, when* d *is even,* $dCov^2(X_1, \ldots, X_d) \leq \prod_{i=1}^d dCov^2(\underbrace{X_i, \ldots, X_i}_{d})^{1/d}$.

(3) Denote by μ_j *the uniform probability measure on the unit sphere* $\mathcal{S}^{p_j - 1}$*. Then*

$$dCov^2(X_1, \ldots, X_d) = C \int_{\prod_{j=1}^d \mathcal{S}^{p_j-1}} dCov^2(\langle u_1, X_1 \rangle, \ldots, \langle u_d, X_d \rangle) d\mu_1(u_1) \cdots d\mu_d(u_d),$$

and

$$\begin{aligned} & JdCov^2(X_1, \ldots, X_d; c) \\ =& C' \int_{\prod_{j=1}^d \mathcal{S}^{p_j-1}} JdCov^2(\langle u_1, X_1 \rangle, \ldots, \langle u_d, X_d \rangle; c) d\mu_1(u_1) \cdots d\mu_d(u_d), \end{aligned}$$

for some positive constants C *and* C'.

Proof of Proposition A.0.1. To prove (1), we have by the triangle inequality

$$|\mathbb{E}[|X_j - x'|] - |x - x'|| \leq \mathbb{E}[|x - X_j'|].$$

Thus we have $|U_j(x, x')| \leq \min\{a_j(x), a_j(x')\}$, which implies that

$$\mathbb{E}\left[\prod_{j=1}^d U_j(X_j, X_j')\right] \leq \mathbb{E}\left[\prod_{j=1}^d \min\{a_j(X_j), a_j(X_j')\}\right].$$

For any partition $S_1 \cup S_2 = \{1, 2, \ldots, d\}$ and $S_1 \cap S_2 = \emptyset$, using the independence between X_j and X_j', we get

$$dCov^2(X_1, X_2, \ldots, X_d) \leq \mathbb{E}\left[\prod_{i \in S_1} a_j(X_j)\right] \mathbb{E}\left[\prod_{i \in S_2} a_j(X_j)\right].$$

(2) follows from the Hölder's inequality directly. Finally, by the change of variables: $t_1 = r_i u_i$ where $r_i \in (0, +\infty)$ and $u_i \in \mathcal{S}^{p_i - 1}$, we have

$$
\begin{aligned}
& dCov^2(X_1, X_2, \ldots, X_d) \\
=& \int_{\mathbb{R}^{p_0}} \left| \mathbb{E}\left[\prod_{i=1}^{d} (f_i(t_i) - e^{\imath \langle t_i, X_i \rangle}) \right] \right|^2 dw \\
=& C_1 \int_{\mathcal{S}_+^{p_1}} \cdots \int_{\mathcal{S}_+^{p_d}} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \left| \mathbb{E}\left[\prod_{i=1}^{d} (\mathbb{E} e^{\imath r_i \langle u_i, X_i \rangle} - e^{\imath r_i \langle u_i, X_i \rangle}) \right] \right|^2 \prod_{i=1}^{d} d\mu_i(u_i) dr_i \\
=& C_2 \int_{\mathcal{S}_+^{p_1}} \cdots \int_{\mathcal{S}_+^{p_d}} JdCov^2(\langle u_1, X_1 \rangle, \ldots, \langle u_d, X_d \rangle; c) d\mu_1(u_1) \cdots d\mu_d(u_d) \\
=& C_3 \int_{\mathcal{S}^{p_1}} \cdots \int_{\mathcal{S}^{p_d}} JdCov^2(\langle u_1, X_1 \rangle, \ldots, \langle u_d, X_d \rangle; c) d\mu_1(u_1) \cdots d\mu_d(u_d),
\end{aligned}
$$

where C_1, C_2, C_3 are some positive constants. ♢

Property (1) gives an upper bound for $dCov^2(X_1, X_2, \ldots, X_d)$, which is motivated by Lemma 2.1 of Lyons (2013), whereas an alternative upper bound is given in Property (2) which follows directly from the Hölder's inequality. Property (3) allows us to represent $dCov$ of random vectors of any dimensions as an integral of $dCov$ of univariate random variables, which are the projections of the aforementioned random vectors.

Proof of Proposition 2.2.3. The "if" part is trivial. To prove the "only if" part, we proceed using induction. Clearly this is true if $d = 2$. Suppose the result holds for $d = m$. Note that $dCov^2(X_1, X_2, \ldots, X_{m+1}) = 0$ implies that $\mathbb{E}\left[\prod_{i=1}^{m+1}(f_i(t_i) - e^{\imath\langle t_i X_i \rangle})\right] = 0$ almost everywhere. Thus we can write the higher order effect $f_{12\cdots(m+1)}(t_1, \ldots, t_{m+1}) - \prod_{i=1}^{m+1} f_i(t_i)$ as a linear combination of the lower order effects. By the assumption that $(X_{i_1}, \ldots, X_{i_m})$ are mutually independent for any m-tuples in I_m^d with $m < d$, we know $f_{12\cdots(m+1)}(t_1, \ldots, t_{m+1}) - \prod_{i=1}^{m+1} f_i(t_i) = 0$. ♢

Proof of Proposition 2.2.4. Notice that

$$\begin{aligned}&\prod_{i=1}^{d}(U_i(X_i,X_i')+c)\\=&c^d+c^{d-1}\sum_{i=1}^{d}U_i(X_i,X_i')+c^{d-2}\sum_{(i_1,i_2)\in I_2^d}U_{i_1}(X_{i_1},X_{i_1}')U_{i_2}(X_{i_2},X_{i_2}')\\&+\cdots+\prod_{i=1}^{d}U_i(X_i,X_i').\end{aligned}$$

The conclusion follows from the fact that $\mathbb{E}[U_i(X_i,X_i')]=0$, equation (2.4) and the definition of JdCov. ♢

Proof of Proposition 2.2.7. We only prove the "if" part. If $dcf(t_1,\dots,t_d)$ can be factored, $U(X_i,X_i')$ are independent. Therefore, it is easy to see that $JdCov^2(X_1,\dots,X_d;c)=0$, which implies that $\{X_1,\dots,X_d\}$ are mutually independent by Proposition 2.2.3. ♢

Proof of Lemma 2.3.1. The RHS of (2.13) in Chapter 2 is equal to

$$\frac{1}{n^2}\sum_{k,l=1}^{n}\prod_{i=1}^{m}\int_{\mathbb{R}}(\hat{f}_i(t_i)-e^{\imath\langle t_i,X_{ki}\rangle})(\hat{f}_i(-t_i)-e^{-\imath\langle t_i,X_{li}\rangle})w_{p_i}(t_i)dt_i.$$

Thus it is enough to prove that

$$\int(\hat{f}_i(t_i)-e^{\imath\langle t_i,X_{ki}\rangle})(\hat{f}_i(-t_i)-e^{-\imath\langle t_i,X_{li}\rangle})\,w_{p_i}(t_i)dt_i\;=\;\widehat{U}_i(k,l).$$

To this end, we note that

$$
\begin{aligned}
&\left(\hat{f}_i(t_i) - e^{\imath\langle t_i, X_{ki}\rangle}\right)\left(\hat{f}_i(-t_i) - e^{-\imath\langle t_i, X_{li}\rangle}\right)\\
&= \hat{f}_i(t_i)\hat{f}_i(-t_i) \;-\; e^{\imath\langle t_i, X_{ki}\rangle}\hat{f}_i(-t_i) \;-\; \hat{f}_i(t_i)e^{-\imath\langle t_i, X_{li}\rangle} \;+\; e^{\imath\langle t_i, (X_{ki}-X_{li})\rangle}\\
&= \frac{1}{n^2}\sum_{k,l=1}^{n} e^{\imath\langle t_i, (X_{ki}-X_{li})\rangle} \;-\; \frac{1}{n}\sum_{l=1}^{n} e^{\imath\langle t_i, (X_{ki}-X_{li})\rangle} \;-\; \frac{1}{n}\sum_{k=1}^{n} e^{\imath\langle t_i, (X_{ki}-X_{li})\rangle} \;+\; e^{\imath\langle t_i, (X_{ki}-X_{li})\rangle}\\
&= \frac{1}{n}\sum_{l=1}^{n}(1 - e^{\imath\langle t_i, (X_{ki}-X_{li})\rangle}) \;+\; \frac{1}{n}\sum_{k=1}^{n}(1 - e^{\imath\langle t_i, (X_{ki}-X_{li})\rangle}) - (1 - e^{\imath\langle t_i, (X_{ki}-X_{li})\rangle})\\
&\quad - \frac{1}{n^2}\sum_{k,l=1}^{n}(1 - e^{\imath\langle t_i, (X_{ki}-X_{li})\rangle})\ .
\end{aligned}
$$

Using (2.11) of Székely et al. (2007), we obtain

$$
\begin{aligned}
&\int (\hat{f}_i(t_i) - e^{\imath\langle t_i, X_{ki}\rangle})(\overline{\hat{f}_i(t_i)} - e^{-\imath\langle t_i, X_{li}\rangle})\, w_{p_i}(t_i)dt_i\\
&= \frac{1}{n}\sum_{l=1}^{n}|X_{ki}-X_{li}|_{p_i} \;+\; \frac{1}{n}\sum_{k=1}^{n}|X_{ki}-X_{li}|_{p_i} \;-\; |X_{ki}-X_{li}|_{p_i} \;-\; \frac{1}{n^2}\sum_{k,l=1}^{n}|X_{ki}-X_{li}|_{p_i}\\
&= \widehat{U}_i(k,l)\ .
\end{aligned}
$$

Finally, (2.14) in Chapter 2 follows from (2.13) in Chapter 2 and the definition of $\widehat{JdCov}^2$. ♢

Proof of Proposition 2.3.1. Define

$$
\xi(t_1, t_2, \dots, t_d) = \mathbb{E}\left[\prod_{i=1}^{d}(f_i(t_i) - e^{\imath\langle t_i, X_i\rangle})\right], \quad \xi_n(t_1, t_2, \dots, t_d) = \frac{1}{n}\sum_{j=1}^{n}\prod_{i=1}^{d}(\hat{f}_i(t_i) - e^{\imath\langle t_i, X_{ji}\rangle}),
$$

and note that

$$
\begin{aligned}
dCov^2(X_1, X_2, \cdots, X_d) &= \int |\xi(t_1, t_2, \dots, t_d)|^2\, dw,\\
\widehat{dCov^2}(X_1, X_2, \cdots, X_d) &= \int |\xi_n(t_1, t_2, \dots, t_d)|^2\, dw.
\end{aligned}
$$

Direct calculation shows that

$$\begin{aligned}\xi(t_1,t_2,\ldots,t_d) = & \prod_{j=1}^{d} f_j - d\prod_{j=1}^{d} f_j + \left(f_{12}\prod_{j\neq 1,2} f_j + f_{13}\prod_{j\neq 1,3} f_j + \cdots\right) \\ & - \left(f_{123}\prod_{j\neq 1,2,3} f_j + f_{124}\prod_{j\neq 1,2,4} f_j + \cdots\right) + \cdots + (-1)^d f_{12,\ldots,d},\end{aligned}$$

and $\xi_n(t_1,t_2,\ldots,t_d)$ has the same expression by replacing the characteristic functions by their empirical counterparts in $\xi(t_1,t_2,\ldots,t_d)$. Then by the strong law of large numbers, we have for any fixed $(t_1,t_2,\ldots,t_d)$,

$$\xi_n(t_1,t_2,\ldots,t_d) \xrightarrow{a.s} \xi(t_1,t_2,\ldots,t_d).$$

For complex numbers $x_1,x_2,\ldots,x_n$ with $n \geq 2$, the CR inequality says that for any $r > 1$

$$\left|\sum_{i=1}^{n} x_i\right|^r \leq n^{r-1}\sum_{i=1}^{n}|x_i|^r. \tag{A.2}$$

Using (A.2), we get

$$\begin{aligned}|\xi_n(t_1,t_2,\ldots,t_d)|^2 = & \left|\frac{1}{n}\sum_{j=1}^{n}\prod_{i=1}^{d}(\hat{f}_i(t_i) - e^{\imath\langle t_i,X_{ji}\rangle})\right|^2 \leq \frac{1}{n^2}\,n^{2-1}\sum_{j=1}^{n}\prod_{i=1}^{d}\left|\hat{f}_i(t_i) - e^{\imath\langle t_i,X_{ji}\rangle}\right|^2 \\ = & \frac{1}{n}\sum_{j=1}^{n}\prod_{i=1}^{d} 4 = 4^d.\end{aligned}$$

For any $\delta > 0$, define $D(\delta) = \{(t_1,t_2,\ldots,t_d) : \delta \leq |t_i|_{p_i} \leq 1/\delta\,,\ i = 1,2,\ldots,d\}$. Notice that

$$\begin{aligned}\widehat{dCov^2}(X_1,X_2,\ldots,X_d) = & \int_{D(\delta)} |\,\xi_n(t_1,t_2,\ldots,t_d)\,|^2\,dw + \int_{D^c(\delta)} |\,\xi_n(t_1,t_2,\ldots,t_d)\,|^2 dw \\ = & \ D^{(1)}_{n,\delta} + D^{(2)}_{n,\delta} \quad \text{(say)},\end{aligned}$$

where $D^{(1)}_{n,\delta} \leq \int_{D(\delta)} 4^d < \infty$. Using the Dominated Convergence Theorem (DCT), we have as

$n \to \infty$,

$$D_{n,\delta}^{(1)} \xrightarrow{a.s} \int_{D(\delta)} |\,\xi(t_1, t_2, .., t_d)\,|^2 \, dw \;=\; D_\delta^{(1)} \quad \text{(say).}$$

So, almost surely

$$\lim_{\delta\to 0}\lim_{n\to\infty} D_{n,\delta}^{(1)} \;= \lim_{\delta\to 0} D_\delta^{(1)} = \int |\,\xi(t_1, t_2, .., t_d)\,|^2 \, dw = dCov^2(X_1, X_2, \,.., X_d).$$

The proof will be complete if we can show almost surely

$$\lim_{\delta\to 0}\lim_{n\to\infty} D_{n,\delta}^{(2)} \;=\; 0.$$

To this end, write $D^c(\delta) = \bigcup_{i=1}^d (A_i^1 \cup A_i^2)$, where $A_i^1 = \{|t_i|_{p_i} < \delta\}$ and $A_i^2 = \{|t_i|_{p_i} > \frac{1}{\delta}\}$ for $i = 1, 2, \ldots, d$. Then we have

$$D_{n,\delta}^{(2)} = \int_{D^c(\delta)} |\,\xi_n(t_1, t_2, \ldots, t_d)\,|^2 \, dw \;\le \sum_{\substack{i=1,2,\ldots,d \\ k=1,2}} \int_{A_i^k} |\xi_n(t_1, t_2, \ldots, t_d)|^2 dw.$$

Define $u_j^i = e^{\imath\langle t_i, X_{ji}\rangle} - f_i(t_i)$ for $1 \le j \le n$ and $1 \le i \le d$. Following the proof of Theorem 2 of Székely *et al.* (2007), we have for $i = 1, 2, \ldots, d$,

$$\int_{\mathbb{R}^{p_i}} \frac{|u_j^i|^2}{c_{p_i}|t_i|_{p_i}^{1+p_i}} \, dt_i \;\le\; 2\,(\,|X_{ji}| + \mathbb{E}|X_i|\,), \tag{A.3}$$

$$\int_{|t_i|_{p_i}<\delta} \frac{|u_j^i|^2}{c_{p_i}|t_i|_{p_i}^{1+p_i}} \, dt_i \;\le\; 2\,\mathbb{E}[|X_{ji} - X_i||X_{ji}]\, G(\,|X_{ji} - X_i|\delta\,), \tag{A.4}$$

$$\int_{|t_i|_{p_i}>1/\delta} \frac{|u_j^i|^2}{c_{p_i}|t_i|_{p_i}^{1+p_i}} \, dt_i \;\le\; 4\delta, \tag{A.5}$$

where

$$G(y) = \int_{|z|<y} \frac{1 - \cos z_1}{|z|^{1+p}} dz,$$

which satisfies that $G(y) \leq c_p$ and $\lim\limits_{y \to 0} G(y) = 0$. Notice that

$$\xi_n(t_1, t_2, \dots, t_d) = \frac{1}{n} \sum_{j=1}^{n} \prod_{i=1}^{d} \left(\frac{1}{n} \sum_{k=1}^{n} u_k^i - u_j^i \right).$$

Some algebra yields that

$$\begin{aligned}
&\xi_n(t_1, t_2, \dots, t_d) \\
=& \prod_{i=1}^{d} \left(\frac{1}{n} \sum_{k=1}^{n} u_k^i \right) - d \prod_{i=1}^{d} \left(\frac{1}{n} \sum_{k=1}^{n} u_k^i \right) \\
&+ \left\{ \left(\frac{1}{n} \sum_{k=1}^{n} u_k^1 u_k^2 \right) \prod_{i \neq 1,2} \left(\frac{1}{n} \sum_{k=1}^{n} u_k^i \right) + \left(\frac{1}{n} \sum_{k=1}^{n} u_k^1 u_k^3 \right) \prod_{i \neq 1,3} \left(\frac{1}{n} \sum_{k=1}^{n} u_k^i \right) + \cdots \right\} \\
&+ \left\{ \left(\frac{1}{n} \sum_{k=1}^{n} u_k^1 u_k^2 u_k^3 \right) \prod_{i \neq 1,2,3} \left(\frac{1}{n} \sum_{k=1}^{n} u_k^i \right) + \left(\frac{1}{n} \sum_{k=1}^{n} u_k^1 u_k^2 u_k^4 \right) \prod_{i \neq 1,2,4} \left(\frac{1}{n} \sum_{k=1}^{n} u_k^i \right) + \cdots \right\} \\
&+ (-1)^d \left(\frac{1}{n} \sum_{k=1}^{n} u_k^1 u_k^2 \cdots u_k^d \right).
\end{aligned}$$

By the CR-inequality, we get

$$\begin{aligned}
&|\xi_n(t_1, t_2, \dots, t_d)|^2 \\
=& C \Bigg[\prod_{i=1}^{d} \left(\frac{1}{n} \sum_{k=1}^{n} |u_k^i|^2 \right) + d^2 \prod_{i=1}^{d} \left(\frac{1}{n} \sum_{k=1}^{n} |u_k^i|^2 \right) \\
&+ \left\{ \left(\left| \frac{1}{n} \sum_{k=1}^{n} u_k^1 u_k^2 \right|^2 \right) \prod_{i \neq 1,2} \left(\frac{1}{n} \sum_{k=1}^{n} |u_k^i|^2 \right) + \left(\left| \frac{1}{n} \sum_{k=1}^{n} u_k^1 u_k^3 \right|^2 \right) \prod_{i \neq 1,3} \left(\frac{1}{n} \sum_{k=1}^{n} u_k^i \right) + \cdots \right\} \\
&+ \left\{ \left(\left| \frac{1}{n} \sum_{k=1}^{n} u_k^1 u_k^2 u_k^3 \right|^2 \right) \prod_{i \neq 1,2,3} \left(\frac{1}{n} \sum_{k=1}^{n} |u_k^i|^2 \right) + \left(\left| \frac{1}{n} \sum_{k=1}^{n} u_k^1 u_k^2 u_k^4 \right|^2 \right) \prod_{i \neq 1,2,4} \left(\frac{1}{n} \sum_{k=1}^{n} |u_k^i|^2 \right) + \cdots \right\} \\
&+ (-1)^d \left(\left| \frac{1}{n} \sum_{k=1}^{n} u_k^1 u_k^2 \cdots u_k^d \right|^2 \right) \Bigg],
\end{aligned}$$

for some positive constant $C > 0$. By the Cauchy-Schwarz inequality, we have for any $2 \leq q \leq d$,

$$\left| \frac{1}{n} \sum_{k=1}^{n} u_k^1 \, u_k^2 \cdots u_k^q \right|^2 \; \leq \; \frac{1}{n} \sum_{k=1}^{n} \prod_{i \in S_{q_1}} |u_k^i|^2 \cdot \frac{1}{n} \sum_{k=1}^{n} \prod_{i \in S_{q_2}} |u_k^i|^2, \tag{A.6}$$

where $S_{q_1} \cup S_{q_2} = \{1, 2, \ldots, d\}$. By Assumption 4.3.2 and (A.3)-(A.5), we have

$$\lim_{\delta \to 0} \lim_{n \to \infty} \int_{|t_i|_{p_i} < \delta} |\xi_n(t_1, t_2, .., t_d)|^2 \, dw = 0 \quad a.s,$$
$$\lim_{\delta \to 0} \lim_{n \to \infty} \int_{|t_i|_{p_i} > 1/\delta} |\xi_n(t_1, t_2, .., t_d)|^2 \, dw = 0 \quad a.s,$$

for every $i \in \{1, 2, \ldots, d\}$. This implies that $\lim\limits_{\delta \to 0} \lim\limits_{n \to \infty} D_{n,\delta}^{(2)} = 0$ almost surely and thus completes the proof. $\diamondsuit$

Proof of Proposition 2.3.2. Define the empirical process

$$\Gamma_n(t) \; = \sqrt{n} \, \xi_n(t_1, t_2, .., t_d) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \prod_{i=1}^{d} (\hat{f}_i(t_i) - e^{\imath \langle t_i, X_{ji} \rangle}).$$

Then $n\widehat{dcov^2}(X_1, X_2, \ldots, X_d) \; = \; \|\Gamma_n\|^2 \; := \; \int \Gamma_n(t_1, t_2, \ldots, t_d)^2 dw$. Under the assumption of independence, we have $\mathbb{E}(\Gamma_n(t)) = 0$ and

$$\begin{aligned}
\Gamma_n(t) \, \overline{\Gamma_n(t_0)} \; &= \frac{1}{n} \sum_{k,l=1}^{n} \prod_{i=1}^{d} (\hat{f}_i(t_i) - e^{\imath \langle t_i, X_{ki} \rangle})(\hat{f}_i(-t_{i0}) - e^{-\imath \langle t_{i0}, X_{li} \rangle}) \\
&= \frac{1}{n} \left\{ \sum_{k=1}^{n} \prod_{i=1}^{d} (\hat{f}_i(t_i) - e^{\imath \langle t_i, X_{ki} \rangle})(\hat{f}_i(-t_{i0}) - e^{-\imath \langle t_{i0}, X_{ki} \rangle}) \right. \\
&\quad \left. + \sum_{k \neq l}^{n} \prod_{i=1}^{d} (\hat{f}_i(t_i) - e^{\imath \langle t_i, X_{ki} \rangle})(\hat{f}_i(-t_{i0}) - e^{-\imath \langle t_{i0}, X_{li} \rangle}) \right\},
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathbb{E}\big[\Gamma_n(t)\,\overline{\Gamma_n(t_0)}\big] &= \frac{1}{n}\Big\{n\,\prod_{i=1}^{d}\mathbb{E}(\hat{f}_i(t_i)-e^{\imath\langle t_i,X_{ki}\rangle})(\hat{f}_i(-t_{i0})-e^{-\imath\langle t_{i0},X_{ki}\rangle}) \\
&\quad +\,n(n-1)\,\prod_{i=1}^{d}\mathbb{E}(\hat{f}_i(t_i)-e^{\imath\langle t_i,X_{ki}\rangle})(\hat{f}_i(-t_{i0})-e^{-\imath\langle t_{i0},X_{li}\rangle})\Big\} \\
&= \frac{1}{n}\big\{n\,A\,+\,n(n-1)\,B\big\} \quad \text{(say)}.
\end{aligned}$$

Direct calculation shows that

$$\begin{aligned}
A &= \prod_{i=1}^{d}\mathbb{E}\Big\{\frac{1}{n^2}\sum_{a,b=1}^{n}e^{\imath\langle t_i,X_{ai}\rangle-\imath\langle t_{i0},X_{bi}\rangle}\,-\,\frac{1}{n}\sum_{b=1}^{n}e^{\imath\langle t_i,X_{ki}\rangle-\imath\langle t_{i0},X_{bi}\rangle} \\
&\qquad -\,\frac{1}{n}\sum_{a=1}^{n}e^{-\imath\langle t_{i0},X_{ki}\rangle+\imath\langle t_i,X_{ai}\rangle}\,+\,e^{\imath\langle t_i-t_{i0},X_{ki}\rangle}\Big\} \\
&= \prod_{i=1}^{d}\Big[\frac{1}{n^2}\big\{n\,f_i(t_i-t_{i0})\,+\,n(n-1)f_i(t_i)f_i(-t_{i0})\big\} \\
&\qquad -\,\frac{2}{n}\big\{f_i(t_i-t_{i0})\,+\,(n-1)f_i(t_i)f_i(-t_{i0})\big\}\,+\,f_i(t_i-t_{i0})\Big] \\
&= \left(\frac{n-1}{n}\right)^{d}\prod_{i=1}^{d}\big\{f_i(t_i-t_{i0})\,-\,f_i(t_i)f_i(-t_{i0})\big\},
\end{aligned}$$

and

$$\begin{aligned}
B &= \prod_{i=1}^{d}\mathbb{E}\Big[\frac{1}{n^2}\sum_{a,b=1}^{n}e^{\imath\langle t_i,X_{ai}\rangle-\imath\langle t_{i0},X_{bi}\rangle}\,-\sum_{b=1}^{n}e^{\imath\langle t_i,X_{ki}\rangle-\imath\langle t_{i0},X_{bi}\rangle} \\
&\qquad -\,\frac{1}{n}\sum_{a=1}^{n}e^{-\imath\langle t_{i0},X_{li}\rangle+\imath\langle t_i,X_{ai}\rangle}\,+\,e^{\imath\langle t_i,X_{ki}\rangle-\imath\langle t_{i0},X_{li}\rangle}\Big] \\
&= \prod_{i=1}^{d}\Big[\frac{1}{n^2}\big\{n\,f_i(t_i-t_{i0})\,+\,n(n-1)f_i(t_i)f_i(-t_{i0})\big\} \\
&\qquad -\,\frac{2}{n}\big\{f_i(t_i-t_{i0})\,+\,(n-1)f_i(t_i)f_i(-t_{i0})\big\}\,+\,f_i(t_i)f_i(-t_{i0})\Big] \\
&= \left(-\frac{1}{n}\right)^{d}\prod_{i=1}^{d}\big\{f_i(t_i-t_{i0})\,-\,f_i(t_i)f_i(-t_{i0})\big\}.
\end{aligned}$$

Hence we obtain

$$\mathbb{E}\big[\Gamma_n(t)\overline{\Gamma_n(t_0)}\big] = c_n \prod_{i=1}^{d} \big\{f_i(t_i - t_{i0}) - f_i(t_i)f_i(-t_{i0})\big\}, \tag{A.7}$$

where $c_n = \left(\frac{n-1}{n}\right)^d + (n-1)\left(-\frac{1}{n}\right)^d$. To prove $\|\Gamma_n\|^2 \xrightarrow{d} \|\Gamma\|^2$, we construct a sequence of random variables $\{Q_n(\delta)\}$ such that

1. $Q_n(\delta) \xrightarrow{d} Q(\delta)$ as $n \to \infty$, for any fixed $\delta > 0$;

2. $\limsup\limits_{n\to\infty} \mathbb{E}|\,Q_n(\delta) - \|\Gamma_n\|^2\,| \to 0$ as $\delta \to 0$;

3. $Q(\delta) \xrightarrow{d} \|\Gamma\|^2$ as $\delta \to 0$.

Then $\|\Gamma_n\|^2 \xrightarrow{d} \|\Gamma\|^2$ follows from Theorem 8.6.2 of Resnick (1999).

We first show (1). Define

$$Q_n(\delta) \;=\; \int_{D(\delta)} |\Gamma_n(t)|^2\, dw, \quad Q(\delta) \;=\; \int_{D(\delta)} |\Gamma(t)|^2\, dw.$$

Given $\epsilon > 0$, choose a partition $\{D_k\}_{k=1}^N$ of $D(\delta)$ into N measurable sets with diameter at most ϵ. Then

$$Q_n(\delta) \;=\; \sum_{k=1}^{N} \int_{D_k} |\Gamma_n(t)|^2\, dw, \quad Q(\delta) \;=\; \sum_{k=1}^{N} \int_{D_k} |\Gamma(t)|^2\, dw.$$

Define

$$Q_n^{\epsilon}(\delta) \;=\; \sum_{k=1}^{N} \int_{D_k} |\Gamma_n(t^k)|^2\, dw, \quad Q^{\epsilon}(\delta) \;=\; \sum_{k=1}^{N} \int_{D_k} |\Gamma(t^k)|^2\, dw,$$

where $\{t^k\}_{k=1}^N$ are a set of distinct points such that $t^k \in D_k$. In view of Theorem $8.6.2$ of Resnick (1999), it suffices to show that

i) $\limsup\limits_{\epsilon\to 0}\; \limsup\limits_{n\to\infty} \mathbb{E}|\,Q_n^{\epsilon}(\delta) - Q_n(\delta)\,| \;=\; 0$;

ii) $\limsup\limits_{\epsilon\to 0} \mathbb{E}|\,Q^{\epsilon}(\delta) - Q(\delta)\,| \;=\; 0$;

iii) $Q_n^{\epsilon}(\delta) \xrightarrow{d} Q^{\epsilon}(\delta)$ as $n \to \infty$, for any fixed $\delta > 0$.

To this end, define $\beta_n(\epsilon) \;=\; \sup_{t,t_0} \mathbb{E}\big|\,|\Gamma_n(t)|^2 - |\Gamma_n(t_0)|^2\,\big|$ and $\beta(\epsilon) \;=\; \sup_{t,t_0} \mathbb{E}\big|\,|\Gamma(t)|^2 - |\Gamma(t_0)|^2\,\big|$, where the supremum is taken over all all $t = (t_1, .., t_d)$ and $t_0 = (t_{10}, .., t_{d0})$ such that $\delta < |t_i|, |t_{i0}| < 1/\delta$ for $i = 1, 2, \ldots, d$, and $\sum_{i=1}^d |t_i - t_{i0}|_{p_i}^2 < \epsilon^2$. Since the function inside the supremum is continuous in t and t_0, and using the fact that a continuous function on a compact support is uniformly continuous, it follows that $\lim\limits_{\epsilon \to 0} \beta(\epsilon) = 0$ and $\lim\limits_{\epsilon \to 0} \beta_n(\epsilon) = 0$ for fixed $\delta > 0$ and fixed n. Thus (i) and (ii) hold. To show (iii), it is enough to show

$$\begin{pmatrix} \Gamma_n(t^1) \\ \Gamma_n(t^2) \\ \vdots \\ \Gamma_n(t^N) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \Gamma(t^1) \\ \Gamma(t^2) \\ \vdots \\ \Gamma(t^N) \end{pmatrix},$$

where $(t^1, \ldots, t^N) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \times \cdots \times \mathbb{R}^{p_d}$ is fixed. The rest follows from the Continuous Mapping Theorem and the Cramer-Wold Device. Notice that $\Gamma_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \prod_{i=1}^d \Big[\big(\hat{f}_i(t_i) - f_i(t_i)\big) - \big(e^{\imath \langle t_i, X_{ji} \rangle} - f_i(t_i)\big) \Big]$. By some algebra and the weak law of large number, we have

$$\begin{pmatrix} \Gamma_n(t^1) \\ \Gamma_n(t^2) \\ \vdots \\ \Gamma_n(t^N) \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{j=1}^n \mathcal{Z}_j \; + \; o_p(1),$$

where $\mathcal{Z}_j = (\mathcal{Z}_{j1}, \ldots, \mathcal{Z}_{jN})'$ with $\mathcal{Z}_{jk} = \prod_{i=1}^d \big(f_i(t_i^k) - e^{\imath \langle t_i^k, X_{ji} \rangle}\big)$ for $1 \le k \le N$. By the independence assumption, $\mathbb{E}[X_j] = 0$ and for $1 \le l, m \le N$,

$$\mathbb{E}[\mathcal{Z}_{jl} \overline{\mathcal{Z}}_{jm}] = \prod_{i=1}^d \mathbb{E}\big\{e^{\imath \langle t_i^l, X_{ji} \rangle} - f_i(t_i^l)\big\}\big\{e^{-\imath \langle t_i^m, X_{ji} \rangle} - f_i(t_i^m)\big\} = R(t^l, t^m).$$

By the Central Limit Theorem (CLT) and Stutsky's theorem, as $n \to \infty$,

$$\begin{pmatrix} \Gamma_n(t^1) \\ \Gamma_n(t^2) \\ \vdots \\ \Gamma_n(t^N) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \Gamma(t^1) \\ \Gamma(t^2) \\ \vdots \\ \Gamma(t^N) \end{pmatrix},$$

which completes the proof of (1).

To prove (2), define $u_i = e^{\imath\langle t_i, X_i\rangle} - f_i(t_i)$. Then $|u_i|^2 = 1 + |f_i(t_i)|^2 - e^{\imath\langle t_i, X_i\rangle}\,\overline{f_i(t_i)} - e^{-\imath\langle t_i, X_i\rangle}\,f_i(t_i)$, and hence

$$\mathbb{E}|u_i|^2 = 1 - |f_i(t_i)|^2. \tag{A.8}$$

Following the similar steps as in the proof of Theorem 5 in Székely *et al.*(2007) and using the Fubini's Theorem,

$$\begin{aligned} \mathbb{E}\,|\,Q_n(\delta) - \|\Gamma_n(t)\|^2\,| &= \mathbb{E}\,\Big|\int_{D(\delta)} |\Gamma_n(t)|^2\,dw - \int |\Gamma_n(t)|^2\,dw\,\Big| \\ &\leq \int_{|t_1|_{p_1}<\delta} \mathbb{E}|\Gamma_n(t)|^2\,dw + \int_{|t_1|_{p_1}>1/\delta} \mathbb{E}|\Gamma_n(t)|^2\,dw \\ &\quad + \cdots + \int_{|t_d|_{p_d}<\delta} \mathbb{E}\,|\Gamma_n(t)|^2\,dw + \int_{|t_d|_{p_d}>1/\delta} \mathbb{E}\,|\Gamma_n(t)|^2\,dw. \end{aligned} \tag{A.9}$$

Using (A.7) and (A.8), we have $\mathbb{E}\,|\Gamma_n(t)|^2 = c_n \prod_{i=1}^d \mathbb{E}|u_i|^2$. Along with the independence assumption, we have

$$\begin{aligned} \int_{|t_1|_{p_1}<\delta} \mathbb{E}\,|\Gamma_n(t)|^2\,dw &= c_n \int_{|t_1|_{p_1}<\delta} \frac{\mathbb{E}|u_1|^2}{c_{p_1}\,|t_1|_{p_1}^{1+p_1}}\,dt_1 \prod_{i=2}^d \int \frac{\mathbb{E}|u_i|^2}{c_{p_i}\,|t_i|_{p_i}^{1+p_i}}\,dt_i \\ &\leq 2c_n \mathbb{E}|X_1 - X_1^{'}|_{p_1}\, G(|X_1 - X_1^{'}|_{p_1}\delta) \prod_{i=2}^d 4\mathbb{E}|X_i|_{p_i}. \end{aligned}$$

Therefore

$$\lim_{\delta\to 0}\lim_{n\to\infty} \int_{|t_1|_{p_1}<\delta} \mathbb{E}\,|\Gamma_n(t)|^2\,dw = 0\,.$$

Similarly

$$\begin{aligned}\int_{|t_1|_{p_1}>1/\delta} \mathbb{E}\,|\Gamma_n(t)|^2\, dw &= c_n \int_{|t_1|_{p_1}>1/\delta} \frac{\mathbb{E}|u_1|^2}{c_{p_1}\,|t_1|_{p_1}^{1+p_1}}\, dt_1 \cdot \prod_{i=2}^{d} \int \frac{\mathbb{E}|u_i|^2}{c_{p_i}\,|t_i|_{p_i}^{1+p_i}}\, dt_i \\ &\le 4\delta c_n \prod_{i=2}^{d} 4\mathbb{E}|X_i|_{p_i} .\end{aligned}$$

Therefore

$$\lim_{\delta\to 0}\lim_{n\to\infty} \int_{|t_1|_{p_1}>1/\delta} \mathbb{E}\,|\Gamma_n(t)|^2\, dw \;=\; 0\ .$$

Applying similar argument to the remaining summands in $(A.9)$, we get

$$\lim_{\delta\to 0}\lim_{n\to\infty} \mathbb{E}|\,Q_n(\delta) - \|\Gamma_n(t)\|^2\,| \;=\; 0\ .$$

To prove (3), we note that

$$\Gamma(t)\,\mathbf{1}\big(t\in D(\delta)\big) \;\xrightarrow{a.s}\; \Gamma(t)\,\mathbf{1}\big(t\in \mathbb{R}^{p_1}\times\mathbb{R}^{p_2}\times\cdots\times\mathbb{R}^{p_d}\big),$$

as $\delta \to 0$. Again by the Fubini's Theorem and equation (2.5) of Székely *et al.* (2007),

$$\begin{aligned}\mathbb{E}\|\Gamma\|^2 \;&=\; \int \prod_{i=1}^{d} \big(1-|f_i(t_i)|^2\big)\, dw \;=\; \prod_{i=1}^{d} \int \frac{\big(1-|f_i(t_i)|^2\big)}{c_{p_i}\,|t_i|_{p_i}^{1+p_i}}\, dt_i \\ &= \prod_{i=1}^{d} \mathbb{E}\int \frac{1-\cos\langle t_i, X_i - X_i^{'}\rangle}{c_{p_i}\,|t_i|_{p_i}^{1+p_i}}\, dt_i \\ &= \prod_{i=1}^{d} \mathbb{E}\,|X_i - X_i^{'}|_{p_i} \;<\; \infty\ .\end{aligned}$$

Hence $\|\Gamma\|^2 < \infty$ almost surely. By DCT, $Q(\delta) \xrightarrow{a.s} \|\Gamma\|^2$ as $\delta \to 0$, which completes the proof.
$\diamondsuit$

LEMMA **A.0.1**. *$\widetilde{U}_i(k,l)$ can be composed as*

$$\begin{aligned}\widetilde{U}_i(k,l) =& \frac{n-3}{(n-1)(n-2)}\sum_{u\notin\{k,l\}} U_i(X_{ui},X_{li}) + \frac{n-3}{(n-1)(n-2)}\sum_{v\notin\{k,l\}} U_i(X_{ki},X_{vi}) \\ &- \frac{n-3}{n-1}U_i(X_{ki},X_{li}) + \frac{2}{(n-1)(n-2)}\sum_{u,v\notin\{k,l\},u<v} U_i(X_{ui},X_{vi}),\end{aligned}$$

where the four terms are uncorrelated with each other.

Proof of Lemma A.0.1. The result follows from direct calculation. ♢

PROPOSITION **A.0.2**. $\mathbb{E}[\widetilde{dCov}^2(X_i,X_j)] = dCov^2(X_i,X_j)$.

Proof of Proposition A.0.2. Using Lemma A.0.1 and the fact that $dCov^2(X_i,X_j) = \mathbb{E}[U_i(X_{ki},X_{li})U_j(X_{kj},X_{lj})]$ for $k\neq l$, we have for $k\neq l$,

$$\begin{aligned}&\mathbb{E}[U_i(X_{ki},X_{li})U_j(X_{kj},X_{lj})] \\ =&\left\{\frac{(n-3)^2}{(n-1)^2} + \frac{2(n-3)^2}{(n-1)^2(n-2)} + \frac{2(n-3)}{(n-1)^2(n-2)}\right\}\mathbb{E}[U_i(X_{ki},X_{li})U_j(X_{kj},X_{lj})] \\ =&\frac{n-3}{n-1}dCov^2(X_i,X_j).\end{aligned}$$

It thus implies that

$$\mathbb{E}[\widetilde{dCov^2}(X_i,X_j)] = \frac{n-1}{n-3}\mathbb{E}[U_i(X_{ki},X_{li};\alpha)U_j(X_{kj},X_{lj})] = dCov^2(X_i,X_j),$$

which completes the proof. ♢

Proof of Proposition 2.4.1. Denote by $\mathbf{X} = \{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$. By independence of the bootstrap samples, we have $\mathbb{E}\left[\Gamma_n^*(t)|\,\mathbf{X}\right] = 0$. Proceeding in the similar way as in the proof of PROPOSITION 2.3.2, it can be shown that

$$\mathbb{E}\big[\Gamma_n^*(t)\overline{\Gamma_n^*(t_0)}\mid\mathbf{X}\big] = c_n\prod_{i=1}^d\big\{\hat{f}_i(t_i-t_{i0}) - \hat{f}_i(t_i)\hat{f}_i(-t_{i0})\big\}, \tag{A.10}$$

where $c_n = \left(\frac{n-1}{n}\right)^d + (n-1)\left(-\frac{1}{n}\right)^d$.

To prove $\|\Gamma_n^*\|^2 \xrightarrow{d} \|\Gamma\|^2$ almost surely, we construct a sequence of random variables $\{Q_n^*(\delta)\}$ such that

1. $Q_n^*(\delta) \xrightarrow{d} Q(\delta)$ almost surely as $n \to \infty$, for any fixed $\delta > 0$;
2. $\limsup\limits_{n\to\infty} \mathbb{E}\left[Q_n^*(\delta) - \|\Gamma_n^*\|^2 \mid \mathbf{X}\right] \to 0$ almost surely as $\delta \to 0$;
3. $Q(\delta) \xrightarrow{d} \|\Gamma\|^2$ as $\delta \to 0$.

Then $\|\Gamma_n^*\|^2 \xrightarrow{d} \|\Gamma\|^2$ almost surely follows from Theorem 8.6.2 of Resnick (1999).

We first show (1). Define

$$Q_n^*(\delta) = \int_{D(\delta)} |\Gamma_n^*(t)|^2 \, dw, \quad Q(\delta) = \int_{D(\delta)} |\Gamma(t)|^2 \, dw.$$

Given $\epsilon > 0$, choose a partition $\{D_k\}_{k=1}^N$ of $D(\delta)$ into N measurable sets with diameter at most ϵ. Then

$$Q_n^*(\delta) = \sum_{k=1}^N \int_{D_k} |\Gamma_n^*(t)|^2 \, dw, \quad Q(\delta) = \sum_{k=1}^N \int_{D_k} |\Gamma(t)|^2 \, dw.$$

Define

$$Q_n^{\epsilon *}(\delta) = \sum_{k=1}^N \int_{D_k} |\Gamma_n^*(t^k)|^2 \, dw, \quad Q^{\epsilon}(\delta) = \sum_{k=1}^N \int_{D_k} |\Gamma(t^k)|^2 \, dw,$$

where $\{t^k\}_{k=1}^N$ are a set of distinct points such that $t^k \in D_k$. In view of Theorem 8.6.2 of Resnick (1999), it suffices to show that

i) $\limsup\limits_{\epsilon\to 0} \limsup\limits_{n\to\infty} \mathbb{E}\left[\,|\,Q_n^{\epsilon *}(\delta) - Q_n^*(\delta)\,|\,\middle|\,\mathbf{X}\right] = 0$ almost surely ;

ii) $\limsup\limits_{\epsilon\to 0} \mathbb{E}\left[\,|\,Q^{\epsilon}(\delta) - Q(\delta)\,|\,\right] = 0$;

iii) $Q_n^{\epsilon *}(\delta) \xrightarrow{d} Q^{\epsilon}(\delta)$ almost surely as $n \to \infty$, for any fixed $\delta > 0$.

To this end, define

$$\beta_n^*(\epsilon) \;=\; \sup_{t,t_0} \mathbb{E}\,\left[\,\left|\,|\Gamma_n^*(t)|^2 - |\Gamma_n^*(t_0)|^2\,\right|\mathbf{X}\,\right],$$

and,

$$\beta(\epsilon) \;=\; \sup_{t,t_0} \mathbb{E}\left[\,\left|\,|\Gamma(t)|^2 - |\Gamma(t_0)|^2\,\right|\,\right],$$

where the supremum is taken over all all $t = (t_1, .., t_d)$ and $t_0 = (t_{10}, .., t_{d0})$ such that $\delta < |t_i|, |t_{i0}| < 1/\delta$ for $i = 1, 2, \ldots, d$, and $\sum_{i=1}^d |t_i - t_{i0}|_{p_i}^2 < \epsilon^2$. Then for fixed $\delta > 0$, $\lim\limits_{\epsilon\to 0} \beta(\epsilon) = 0$ and $\lim\limits_{\epsilon\to 0} \beta_n^*(\epsilon) = 0$ almost surely for fixed n. Thus (i) and (ii) hold. To show (iii), it is enough to show

$$\begin{pmatrix} \Gamma_n^*(t^1) \\ \Gamma_n^*(t^2) \\ \vdots \\ \Gamma_n^*(t^N) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \Gamma(t^1) \\ \Gamma(t^2) \\ \vdots \\ \Gamma(t^N) \end{pmatrix} \; almost\ surely,$$

where $(t^1, \ldots, t^N) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \times \cdots \times \mathbb{R}^{p_d}$ is fixed. The rest follows from the Continuous Mapping Theorem and the Cramer-Wold Device. Notice that $\Gamma_n^*(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \prod_{i=1}^d \Big[\big(\hat{f}_i^*(t_i) - \hat{f}_i(t_i)\big) - \big(e^{\imath\langle t_i, X_{ji}^*\rangle} - \hat{f}_i(t_i)\big) \Big]$. Using Markov's inequality and Triangle inequality, observe that

$$\begin{aligned}
&\sum_{n=1}^\infty P\left(\left| \frac{1}{n} \sum_{k=1}^n (e^{\imath\langle t_i, X_{ki}^*\rangle} - \hat{f}_i(t_i) \right| > \epsilon \right) \\
&= \sum_{n=1}^\infty P\left(\left| \sum_{k=1}^n Y_{ki} \right| > n\epsilon \right) \;=\; \sum_{n=1}^\infty P\left(\left| \sum_{k=1}^n Y_{ki} \right|^2 > n^2\epsilon^2 \right) \\
&= \sum_{n=1}^\infty P\left(\sum_{k,l=1}^n Y_{ki}\overline{Y_{li}} > n^2\epsilon^2 \right) \;\le\; \sum_{n=1}^\infty \frac{1}{(n\epsilon)^4}\, E\left[(\sum_{k,l=1}^n Y_{ki}\overline{Y_{li}})^2 \,\middle|\, \mathbf{X} \right] \\
&= \sum_{n=1}^\infty \frac{1}{(n\epsilon)^4}\, E\left[\left(\sum_{\substack{k_1,l_1,\\ k_2,l_2=1}}^n Y_{k_1 i}\overline{Y_{l_1 i}}\, Y_{k_2 i}\overline{Y_{l_2 i}} \right) \,\middle|\, \mathbf{X} \right] \\
&\le \sum_{n=1}^\infty \frac{1}{(n\epsilon)^4} \,.\, Cn^2 \;<\; \infty,
\end{aligned}$$

where $C > 0$, $Y_k = e^{\imath\langle t_i, X^*_{ki}\rangle} - \hat{f}_i(t_i)$, and $|Y_k| \le 2$ for any $1 \le k \le n$.

By Borel-Cantelli Lemma, as $n \to \infty$, $\hat{f}^*_i(t_i) - \hat{f}_i(t_i) \xrightarrow{a.s} 0$ almost surely. By some algebra and the weak law of large number, we have

$$\begin{pmatrix} \Gamma^*_n(t^1) \\ \Gamma^*_n(t^2) \\ \vdots \\ \Gamma^*_n(t^N) \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathcal{Z}_j \; + \; U \; ,$$

where $\mathcal{Z}_j = (\mathcal{Z}_{j1}, \ldots, \mathcal{Z}_{jN})'$ with $\mathcal{Z}_{jk} = \prod_{i=1}^{d} \left(\hat{f}_i(t^k_i) - e^{\imath\langle t^k_i, X^*_{ji}\rangle}\right)$ for $1 \le k \le N$, and, $U \xrightarrow{a.s} 0$, almost surely. By the independence of Bootstrap samples, $\mathbb{E}[\mathcal{Z}_j | \, \mathbf{X}] = 0$ and for $1 \le l, m \le N$,

$$\begin{aligned} \mathbb{E}[\mathcal{Z}_{jl}\overline{\mathcal{Z}}_{jm}] &= \prod_{i=1}^{d} \mathbb{E}\left[(e^{\imath\langle t^l_i, X^*_{ji}\rangle} - \hat{f}_i(t^l_i))\,(e^{-\imath\langle t^m_i, X^*_{ji}\rangle} - \hat{f}_i(-t^m_i)) | \, \mathbf{X} \right] \\ &= \prod_{i=1}^{d} \left\{ \hat{f}_i(t^l_i - t^m_i) - \hat{f}_i(t^l_i)\hat{f}_i(-t^m_i) \right\}. \end{aligned}$$

Let R_n and R be $N \times N$ matrices with the $(l, m)^{th}$ element being

$$R_n(l, m) = \prod_{i=1}^{d} \left\{ \hat{f}_i(t^l_i - t^m_i) - \hat{f}_i(t^l_i)\hat{f}_i(-t^m_i) \right\} ,$$

and,

$$R(l, m) = \prod_{i=1}^{d} \left\{ f_i(t^l_i - t^m_i) - f_i(t^l_i) f_i(-t^m_i) \right\} .$$

By Multivariate CLT,

$$R_n^{-\frac{1}{2}} \begin{pmatrix} \Gamma_n^*(t^1) \\ \Gamma_n^*(t^2) \\ \vdots \\ \Gamma_n^*(t^N) \end{pmatrix} \xrightarrow{d} N(0, I_N) \;\; almost\ surely \; ,$$

which, along with the fact $R_n \xrightarrow{a.s} R$ and Slutsky's Theorem, implies

$$\begin{pmatrix} \Gamma_n^*(t^1) \\ \Gamma_n^*(t^2) \\ \vdots \\ \Gamma_n^*(t^N) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \Gamma(t^1) \\ \Gamma(t^2) \\ \vdots \\ \Gamma(t^N) \end{pmatrix} \;\; almost\ surely \; ,$$

and thus completes the proof of (1).

To prove (2), define $u_i^* = e^{\imath\langle t_i, X_i^*\rangle} - \hat{f_i}^*(t_i)$. Then $|u_i|^2 = 1 + |\hat{f}_i(t_i)|^2 - e^{\imath\langle t_i, X_i^*\rangle}\, \overline{\hat{f}_i(t_i)} - e^{-\imath\langle t_i, X_i^*\rangle}\, \hat{f}_i(t_i)$, and hence

$$\mathbb{E}\left[\, |u_i^*|^2 \,|\mathbf{X} \right] \;=\; 1 - |\hat{f}_i(t_i)|^2. \tag{A.11}$$

Following the similar steps as in the proof of Theorem 5 in Székely *et al.*(2007) and using the Fubini's Theorem,

$$\begin{aligned}
&\mathbb{E}\left[\, |Q_n^*(\delta) - \|\Gamma_n^*(t)\|^2| \,\middle|\, \mathbf{X} \right] \\
&= \mathbb{E}\left[\Big| \int_{D(\delta)} |\Gamma_n^*(t)|^2 \, dw \;-\; \int |\Gamma_n^*(t)|^2 \, dw \Big| \,\middle|\, \mathbf{X} \right] \\
&\leq \int_{|t_1|_{p_1} < \delta} \mathbb{E}\left[\, |\Gamma_n^*(t)|^2 \,\middle|\, \mathbf{X} \right] dw + \int_{|t_1|_{p_1} > 1/\delta} \mathbb{E}\left[\, |\Gamma_n^*(t)|^2 \,\middle|\, \mathbf{X} \right] dw \\
&\quad + \cdots + \int_{|t_d|_{p_d} < \delta} \mathbb{E}\left[\, |\Gamma_n^*(t)|^2 \,\middle|\, \mathbf{X} \right] dw + \int_{|t_d|_{p_d} > 1/\delta} \mathbb{E}\left[\, |\Gamma_n^*(t)|^2 \,\middle|\, \mathbf{X} \right] dw.
\end{aligned} \tag{A.12}$$

Using (A.10) and (A.11), we have $\mathbb{E}\left[\, |\Gamma_n^*(t)|^2 \,\middle|\, \mathbf{X} \right] = c_n \prod_{i=1}^d \mathbb{E}\left[\, |u_i^*|^2 \,\middle|\, \mathbf{X} \right]$. Along with the independence assumption, we have

$$
\begin{aligned}
&\int_{|t_1|_{p_1}<\delta} \mathbb{E}\,\left[\,|\Gamma_n^*(t)|^2\,\middle|\,\mathbf{X}\right]\,dw \\
&= c_n \int_{|t_1|_{p_1}<\delta} \frac{\mathbb{E}\,\left[\,|u_1^*|^2\,\middle|\,\mathbf{X}\right]}{c_{p_1}\,|t_1|_{p_1}^{1+p_1}}\,dt_1 \prod_{i=2}^{d} \int \frac{\mathbb{E}\,\left[\,|u_i^*|^2\,\middle|\,\mathbf{X}\right]}{c_{p_i}\,|t_i|_{p_i}^{1+p_i}}\,dt_i \\
&\le 2\,c_n\,\mathbb{E}\,\left[\,|X_1^* - X_1^{*'}|_{p_1}\,G(|X_1^* - X_1^{*'}|_{p_1}\delta)\,|\,\mathbf{X}\right] \prod_{i=2}^{d} 4\,\mathbb{E}\,\left[\,|X_i^*|_{p_i}\,|\,\mathbf{X}\right] \\
&= 2\,c_n\,\frac{1}{n^2}\sum_{j,k=1}^{n} |X_{j1} - X_{k1}|_{p_1}\,G(|X_{j1} - X_{k1}|_{p_1}\delta) \prod_{i=2}^{d} 4\,\frac{1}{n}\sum_{j=1}^{n} |X_{ji}|_{p_i} \\
&\overset{a.s}{\to} 2\,\mathbb{E}\,\left[\,|X_1 - X_1^{'}|_{p_1}\,G(|X_1 - X_1^{'}|_{p_1}\delta)\right] \prod_{i=2}^{d} 4\,\mathbb{E}\,\left[\,|X_i|_{p_i}\,\right] \qquad as\;\; n \to \infty.
\end{aligned}
$$

Therefore

$$
\lim_{\delta\to 0}\,\lim_{n\to\infty} \int_{|t_1|_{p_1}<\delta} \mathbb{E}\,\left[\,|\Gamma_n^*(t)|^2\,\middle|\,\mathbf{X}\right]\,dw \;=\; 0 \quad almost\ surely.
$$

Similarly

$$
\begin{aligned}
\int_{|t_1|_{p_1}>1/\delta} \mathbb{E}\,\left[\,|\Gamma_n^*(t)|^2\,\middle|\,\mathbf{X}\right]\,dw &= c_n \int_{|t_1|_{p_1}>1/\delta} \frac{\mathbb{E}\,\left[\,|u_1^*|^2\,\middle|\,\mathbf{X}\right]}{c_{p_1}\,|t_1|_{p_1}^{1+p_1}}\,dt_1 \cdot \prod_{i=2}^{d} \int \frac{\mathbb{E}\,\left[\,|u_i^*|^2\,\middle|\,\mathbf{X}\right]}{c_{p_i}\,|t_i|_{p_i}^{1+p_i}}\,dt_i \\
&\le 4\delta c_n \prod_{i=2}^{d} 4\,\mathbb{E}\,\left[\,|X_i^*|_{p_i}\,|\,\mathbf{X}\right].
\end{aligned}
$$

Therefore

$$
\lim_{\delta\to 0}\,\lim_{n\to\infty} \int_{|t_1|_{p_1}>1/\delta} \mathbb{E}\,\left[\,|\Gamma_n^*(t)|^2\,\middle|\,\mathbf{X}\right]\,dw \;=\; 0 \quad almost\ surely.
$$

Applying similar argument to the remaining summands in $(A.12)$, we get

$$
\lim_{\delta\to 0}\,\lim_{n\to\infty} \mathbb{E}\,\left[\,|\,Q_n^*(\delta) - \|\Gamma_n^*(t)\|^2\,|\,\middle|\,\mathbf{X}\right] \;=\; 0 \quad almost\ surely.
$$

The proof of part (3) is exactly the same as its counterpart in the proof of Proposition 2.3.2, which completes the proof.

$\diamondsuit$

Let G_n be the set of all functions from $\{1, 2, \ldots, n\}$ to $\{1, 2, \ldots, n\}$. Define a map g: $\mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ as the following

$$g(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \begin{pmatrix} X_{g_1(1),1} & X_{g_2(1),2} & \cdots & X_{g_d(1),d} \\ X_{g_1(2),1} & X_{g_2(2),2} & \cdots & X_{g_d(2),d} \\ \vdots & \vdots & \ddots & \vdots \\ X_{g_1(n),1} & X_{g_2(n),2} & \cdots & X_{g_d(n),d} \end{pmatrix}$$

where $g_i \in G_n$ for $1 \le i \le d$. With some abuse of notation, we denote by $\widehat{JdCov}^2(g(\mathbf{X}_1, \ldots, \mathbf{X}_n))$ the sample (squared) JdCov computed based on the sample $g(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Conditional on the sample, the resampling distribution function $\widehat{F}_n : [0, +\infty) \to [0, 1]$ of the bootstrap statistic is defined for all $t \in \mathbb{R}$ as

$$\widehat{F}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)(t) := \frac{1}{n^{nd}} \sum_{g \in G_n^d} 1_{\{\, n\widehat{JdCov}^2(g(\mathbf{X}_1, \ldots, \mathbf{X}_n)) \le t \,\}}.$$

For $\alpha \in (0, 1)$, we define the α-level bootstrap-assisted test for testing H_0 against H_A as

$$\phi_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) := 1_{\{\, n\widehat{JdCov}^2(\psi(\mathbf{X}_1, .., \mathbf{X}_n)) > \left(\widehat{F}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)\right)^{-1}(1-\alpha) \,\}} \,. \tag{A.13}$$

Proof of Proposition 2.4.2. The proof is in similar lines of the proof of Theorem 3.7 in *Pfister et al.* (2018). There exists a set A_0 with $P(A_0) = 1$ such that for all $\omega \in A_0$ and $\forall\, t \in \mathbb{R}$,

$$\begin{aligned}
\lim_{n\to\infty} \widehat{F}_n(\mathbf{X}_1(\omega), .., \mathbf{X}_n(\omega))\,(t) \;&=\; \lim_{n\to\infty} \frac{1}{n^{nd}} \sum_{g \in G_n^d} \mathbb{1}_{\{\, n\widehat{JdCov}^2\left(g(\mathbf{X}_1(\omega), .., \mathbf{X}_n(\omega))\right) \le t\,\}} \\
&=\; \lim_{n\to\infty} E\left(\mathbb{1}_{\{\, n\widehat{JdCov}^2(g(\mathbf{X}_1(\omega), .., \mathbf{X}_n(\omega))\,) \le t\,\}}\right) \\
&=\; \lim_{n\to\infty} P\left(n\widehat{JdCov}^2\big(g\big(\mathbf{X}_1(\omega),\, .., \mathbf{X}_n(\omega)\big)\big) \;\le\; t\right) \\
&=\; G(t)\;,
\end{aligned}$$

where $G(\cdot)$ is the distribution function of $\sum_{j=1}^{+\infty} \lambda_j^{'} Z_j^2$.

Since G is continuous, for all $\omega \in A_0$ and $\forall\, t \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \left(\widehat{F}_n(\mathbf{X}_1(\omega),\, ..\, , \mathbf{X}_n(\omega))\right)^{-1}(t) \;=\; G^{-1}(t)\,.$$

In particular, for all $\omega \in A_0$, we have

$$\lim_{n\to\infty} \left(\widehat{F}_n(\mathbf{X}_1(\omega),\, ..\, , \mathbf{X}_n(\omega))\right)^{-1}(1-\alpha) \;=\; G^{-1}(1-\alpha)\,. \tag{A.14}$$

When H_0 is true, using Proposition 2.3.2, equation $(A.13)$ and Corollary $11.2.3$ in Lehmann and Romano (2005), we have

$$\begin{aligned}
&\limsup_{n\to\infty}\; P\,(\,\phi_n(\mathbf{X}_1,\ldots,\mathbf{X}_n) = 1\,)\\
=&\limsup_{n\to\infty}\; P\left(\,n\,\widehat{dcov^2}(\mathbf{X}_1,\, ..\, ,\mathbf{X}_n) \;>\; \left(\widehat{F}_n(\mathbf{X}_1,\, ..\, ,\mathbf{X}_n)\right)^{-1}(1-\alpha)\right)\\
=&\,1-\liminf_{n\to\infty}\; P\left(\,n\,\widehat{dcov^2}(\mathbf{X}_1,\, ..\, ,\mathbf{X}_n) \;\le\; \left(\widehat{F}_n(\mathbf{X}_1,\, ..\, ,\mathbf{X}_n)\right)^{-1}(1-\alpha)\right)\\
=&1-G\left(G^{-1}(1-\alpha)\right) \;=\; 1-(1-\alpha) \;=\; \alpha\,.
\end{aligned}$$

This completes the proof of the proposition. ♢

Proof of Proposition 2.4.3. The proof is in similar lines of the proof of Theorem 3.8 in *Pfister et al.* (2018). In the proof of Proposition 2.4.2, we showed that there exists a set A_0 with $P(A_0) = 1$ such that for all $\omega \in A_0$,

$$\lim_{n\to\infty} \left(\widehat{F}_n(\mathbf{X}_1(\omega),\, ..\, , \mathbf{X}_n(\omega))\right)^{-1}(1-\alpha) \;=\; G^{-1}(1-\alpha)\;.$$

Define the set

$$A_1 \;=\; \{\,\omega\; :\; \forall\, t\; \in \mathbb{R},\, \lim_{n\to\infty}\, \mathbb{1}_{\{\, n\,\widehat{dcov^2}(\mathbf{X}_1(\omega),\,..\,,\mathbf{X}_n(\omega))\,\le t\,\}}\; =\; 0\,\}\;. \tag{A.15}$$

Clearly, $P(A_1) = 1$ and hence $P(A_0 \cap A_1) = 1$. Fix $\omega \in A_0 \cap A_1$. Then by $(A.13)$ and $(A.14)$, there exists a constant $t^* \in \mathbb{R}$ such that $\forall n \in \mathbb{N}$,

$$\lim_{n\to\infty}\, \Big(\widehat{F}_n(\mathbf{X}_1(\omega),\,..\,,\mathbf{X}_n(\omega))\Big)^{-1}(1-\alpha)\;\le\; t^*\;.$$

Therefore,

$$\begin{aligned}&\lim_{n\to\infty}\; \mathbb{1}_{\{\, n\,\widehat{dcov^2}(\mathbf{X}_1(\omega),\,..\,,\mathbf{X}_n(\omega))\,\le\, \left(\widehat{F_n}(\mathbf{X}_1(\omega),\,..\,,\mathbf{X}_n(\omega))\,\right)^{-1}(1-\alpha)\,\}}\\ &\qquad\le\; \lim_{n\to\infty}\; \mathbb{1}_{\{\, n\,\widehat{dcov^2}(\mathbf{X}_1(\omega),\,..\,,\mathbf{X}_n(\omega))\,\le\, t^*\,\}}\;\;=\;\;0\;,\end{aligned}$$

i.e., $\mathbb{1}_{\{\, n\,\widehat{dcov^2}(\mathbf{X}_1,\,..\,,\mathbf{X}_n)\,\le\, \left(\widehat{F_n}(\mathbf{X}_1,\,..\,,\mathbf{X}_n)\,\right)^{-1}(1-\alpha)\,\}} \;\xrightarrow{a.s}\; 0$ as $n \to \infty$. It follows by dominated convergence theorem that

$$\begin{aligned}&\lim_{n\to\infty}\, P\left(\, n\,\widehat{dcov^2}(\mathbf{X}_1,\,..\,,\mathbf{X}_n)\;\le\;\Big(\widehat{F_n}(\mathbf{X}_1,\,..\,,\mathbf{X}_n)\Big)^{-1}(1-\alpha)\right)\\ &\quad=\;\lim_{n\to\infty}\, E\left(\,\mathbb{1}_{\{\, n\,\widehat{dcov^2}(\mathbf{X}_1,\,..\,,\mathbf{X}_n)\,\le\, \left(\widehat{F_n}(\mathbf{X}_1,\,..\,,\mathbf{X}_n)\,\right)^{-1}(1-\alpha)\,\}}\right)\;=\;0\,,\end{aligned}$$

which completes the proof of the proposition. $\diamondsuit$

Table A.1: Empirical size and power for the bootstrap-assisted joint independence tests (based on the U-statistics) for $c = 1$. The results are obtained based on 1000 replications and the number of bootstrap resamples is taken to be 500.

				$\widetilde{JdCov^2}$		$\widetilde{JdCov_S^2}$		$\widetilde{JdCov_R^2}$		dHSIC		T_{MT}	
		n	$\tilde{d}$	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
Ex 2.5.1	(1)	50	5	0.097	0.049	0.110	0.059	0.099	0.045	0.102	0.047	0.099	0.042
	(1)	50	10	0.100	0.050	0.108	0.053	0.101	0.053	0.091	0.042	0.068	0.034
	(2)	50	5	0.103	0.062	0.096	0.052	0.099	0.045	0.104	0.048	0.115	0.061
	(2)	50	10	0.119	0.062	0.121	0.056	0.101	0.053	0.105	0.041	0.106	0.056
	(3)	50	5	0.057	0.022	0.112	0.047	0.099	0.045	0.103	0.047	0.027	0.011
	(3)	50	10	0.050	0.017	0.100	0.051	0.101	0.053	0.091	0.040	0.013	0.006
	(1)	100	5	0.101	0.05	0.098	0.057	0.091	0.042	0.088	0.038	0.098	0.052
	(1)	100	10	0.105	0.045	0.085	0.043	0.102	0.053	0.091	0.038	0.098	0.059
	(2)	100	5	0.094	0.047	0.093	0.049	0.091	0.042	0.102	0.042	0.094	0.054
	(2)	100	10	0.115	0.063	0.102	0.06	0.102	0.053	0.104	0.049	0.106	0.06
	(3)	100	5	0.08	0.034	0.115	0.058	0.091	0.042	0.095	0.038	0.043	0.019
	(3)	100	10	0.066	0.025	0.104	0.052	0.102	0.053	0.111	0.047	0.021	0.005
Ex 2.5.2	(1)	50	5	0.606	0.474	0.510	0.381	0.626	0.513	0.229	0.142	0.607	0.490
	(1)	50	10	0.495	0.359	0.306	0.192	0.705	0.596	0.145	0.070	0.669	0.545
	(2)	50	5	0.813	0.720	0.732	0.632	0.835	0.751	0.342	0.219	0.805	0.706
	(2)	50	10	0.797	0.668	0.466	0.339	0.941	0.904	0.201	0.113	0.906	0.846
	(3)	50	5	0.877	0.817	0.815	0.764	0.886	0.840	0.374	0.242	0.849	0.787
	(3)	50	10	0.848	0.749	0.521	0.396	0.960	0.917	0.174	0.096	0.942	0.897
	(1)	100	5	0.903	0.854	0.834	0.767	0.93	0.881	0.405	0.278	0.913	0.863
	(1)	100	10	0.853	0.756	0.468	0.337	0.977	0.954	0.203	0.114	0.97	0.936
	(2)	100	5	0.989	0.981	0.968	0.946	0.99	0.983	0.618	0.491	0.987	0.975
	(2)	100	10	0.998	0.988	0.79	0.657	1	1	0.36	0.215	1	0.999
	(3)	100	5	0.998	0.994	0.988	0.98	0.997	0.991	0.649	0.518	0.995	0.991
	(3)	100	10	0.998	0.991	0.816	0.721	1	1	0.307	0.189	1	0.999
Ex 2.5.3	(1)	50	3	0.998	0.986	1.000	1.000	0.624	0.365	0.898	0.794	0.221	0.106
	(2)	50	3	1	1	1	1	1	1	1	1	1	1
	(1)	100	3	1	1	1	1	1	0.999	1	1	0.622	0.368
	(2)	100	3	1	1	1	1	1	1	1	1	1	1
Ex 2.5.4	(1)	100	5	0.339	0.195	0.523	0.379	0.122	0.07	0.219	0.114	0.073	0.038
	(1)	100	10	0.105	0.027	0.248	0.147	0.049	0.019	0.117	0.043	0.025	0.008
	(2)	100	5	0.369	0.235	0.466	0.362	0.162	0.09	0.406	0.25	0.241	0.161
	(2)	100	10	0.097	0.04	0.218	0.13	0.06	0.021	0.164	0.077	0.046	0.022
	(1)	200	5	0.813	0.676	0.929	0.865	0.238	0.128	0.378	0.224	0.085	0.044
	(1)	200	10	0.262	0.140	0.433	0.305	0.093	0.045	0.137	0.061	0.047	0.023
	(2)	200	5	0.773	0.662	0.778	0.689	0.398	0.263	0.797	0.665	0.581	0.505
	(2)	200	10	0.290	0.171	0.384	0.296	0.136	0.065	0.300	0.173	0.141	0.077

Note: In Examples 2.5.1-2.5.3, $\tilde{d}$ denotes the number of random variables d. In Example 2.5.4, $\tilde{d}$ stands for p.

Table A.2: Empirical size and power for the bootstrap-assisted joint independence tests (based on the U-statistics) for $c = 2$ and 0.5. The results are obtained based on 1000 replications and the number of bootstrap resamples is taken to be 500.

				$c = 2$						$c = 0.5$					
				$\widetilde{JdCov}^2$		$\widetilde{JdCov}^2_S$		$\widetilde{JdCov}^2_R$		$\widetilde{JdCov}^2$		$\widetilde{JdCov}^2_S$		$\widetilde{JdCov}^2_R$	
		n	$\tilde{d}$	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
Ex 2.5.1	(1)	50	5	0.097	0.05	0.099	0.052	0.094	0.045	0.102	0.05	0.115	0.061	0.099	0.054
	(1)	50	10	0.103	0.049	0.112	0.056	0.097	0.048	0.102	0.051	0.116	0.067	0.107	0.051
	(2)	50	5	0.106	0.057	0.102	0.058	0.094	0.045	0.113	0.053	0.110	0.058	0.099	0.054
	(2)	50	10	0.107	0.051	0.107	0.06	0.097	0.048	0.125	0.074	0.120	0.071	0.107	0.051
	(3)	50	5	0.063	0.017	0.101	0.048	0.094	0.045	0.058	0.019	0.105	0.052	0.099	0.054
	(3)	50	10	0.056	0.022	0.100	0.053	0.097	0.048	0.026	0.009	0.096	0.049	0.107	0.051
	(1)	100	5	0.087	0.043	0.098	0.049	0.085	0.046	0.097	0.059	0.107	0.066	0.098	0.042
	(1)	100	10	0.104	0.049	0.107	0.050	0.098	0.052	0.087	0.040	0.117	0.056	0.104	0.053
	(2)	100	5	0.088	0.046	0.091	0.039	0.085	0.046	0.104	0.059	0.108	0.057	0.098	0.042
	(2)	100	10	0.099	0.060	0.105	0.065	0.098	0.052	0.101	0.060	0.101	0.054	0.104	0.053
	(3)	100	5	0.080	0.034	0.113	0.057	0.085	0.046	0.086	0.034	0.120	0.063	0.098	0.042
	(3)	100	10	0.077	0.029	0.117	0.053	0.098	0.052	0.044	0.019	0.100	0.055	0.104	0.053
Ex 2.5.2	(1)	50	5	0.644	0.526	0.629	0.504	0.630	0.517	0.434	0.323	0.291	0.196	0.610	0.499
	(1)	50	10	0.690	0.580	0.603	0.473	0.718	0.610	0.220	0.125	0.163	0.105	0.615	0.498
	(2)	50	5	0.857	0.777	0.836	0.750	0.837	0.760	0.641	0.519	0.439	0.318	0.816	0.734
	(2)	50	10	0.944	0.887	0.872	0.798	0.953	0.914	0.313	0.212	0.221	0.165	0.887	0.811
	(3)	50	5	0.903	0.851	0.889	0.835	0.892	0.846	0.773	0.692	0.596	0.510	0.876	0.821
	(3)	50	10	0.957	0.918	0.912	0.842	0.967	0.929	0.370	0.254	0.266	0.198	0.915	0.868
	(1)	100	5	0.935	0.890	0.912	0.877	0.932	0.886	0.747	0.637	0.453	0.346	0.916	0.867
	(1)	100	10	0.979	0.943	0.927	0.860	0.983	0.963	0.308	0.194	0.188	0.129	0.949	0.890
	(2)	100	5	0.994	0.987	0.991	0.986	0.991	0.983	0.938	0.897	0.705	0.605	0.988	0.981
	(2)	100	10	1	1	1	0.999	1	1	0.476	0.352	0.274	0.210	1	1
	(3)	100	5	0.998	0.997	0.998	0.994	0.997	0.991	0.980	0.962	0.872	0.817	0.997	0.991
	(3)	100	10	1	1	1	0.999	1	1	0.559	0.444	0.336	0.274	1	0.998
Ex 2.5.3	(1)	50	3	0.797	0.567	0.978	0.893	0.267	0.155	1	1	1	1	1	0.984
	(2)	50	3	1	1	1	1	0.959	0.593	1	1	1	1	1	1
	(1)	100	3	1	0.999	1	1	0.704	0.458	1	1	1	1	1	1
	(2)	100	3	1	1	1	1	1	1	1	1	1	1	1	1
Ex 2.5.4	(1)	100	5	0.198	0.087	0.295	0.195	0.109	0.047	0.605	0.413	0.768	0.638	0.178	0.096
	(1)	100	10	0.074	0.018	0.171	0.092	0.045	0.017	0.149	0.046	0.357	0.221	0.050	0.020
	(2)	100	5	0.342	0.221	0.444	0.315	0.180	0.095	0.438	0.338	0.496	0.419	0.267	0.143
	(2)	100	10	0.083	0.034	0.179	0.105	0.055	0.016	0.134	0.056	0.266	0.176	0.066	0.027
	(1)	200	5	0.435	0.293	0.619	0.462	0.162	0.083	0.981	0.951	0.995	0.987	0.438	0.281
	(1)	200	10	0.146	0.063	0.243	0.146	0.077	0.032	0.465	0.308	0.664	0.528	0.132	0.057
	(2)	200	5	0.698	0.571	0.781	0.669	0.338	0.212	0.715	0.623	0.688	0.611	0.534	0.400
	(2)	200	10	0.214	0.129	0.316	0.213	0.120	0.052	0.349	0.241	0.442	0.352	0.169	0.082

Note: In Examples 2.5.1-2.5.3, $\tilde{d}$ denotes the number of random variables d. In Example 2.5.4, $\tilde{d}$ stands for p.

Table A.3: Empirical size and power for the bootstrap-assisted joint independence tests based on the V-statistic type estimators, with $c = 1$. The results are obtained based on 1000 replications and the number of bootstrap resamples is taken to be 500.

				$\widehat{JdCov}^2$		$\widehat{JdCov}^2_S$		$\widehat{JdCov}^2_R$	
		n	$\tilde{d}$	10%	5%	10%	5%	10%	5%
Ex 2.5.1	(1)	50	5	0.093	0.033	0.269	0.131	0.103	0.052
	(1)	50	10	0.130	0.067	0.257	0.139	0.110	0.063
	(2)	50	5	0.142	0.081	0.106	0.061	0.103	0.052
	(2)	50	10	0.452	0.130	0.077	0.020	0.110	0.063
	(3)	50	5	0.118	0.067	0.200	0.118	0.103	0.052
	(3)	50	10	0.124	0.069	0.195	0.111	0.110	0.063
	(1)	100	5	0.068	0.024	0.204	0.113	0.090	0.044
	(1)	100	10	0.086	0.042	0.184	0.092	0.107	0.058
	(2)	100	5	0.121	0.061	0.102	0.053	0.090	0.044
	(2)	100	10	0.222	0.050	0.056	0.013	0.107	0.058
	(3)	100	5	0.128	0.066	0.191	0.116	0.090	0.044
	(3)	100	10	0.114	0.061	0.168	0.102	0.107	0.058
Ex 2.5.2	(1)	50	5	0.485	0.299	0.649	0.450	0.637	0.528
	(1)	50	10	0.284	0.161	0.428	0.271	0.727	0.627
	(2)	50	5	0.746	0.571	0.806	0.659	0.846	0.768
	(2)	50	10	0.393	0.250	0.544	0.371	0.955	0.911
	(3)	50	5	0.822	0.725	0.877	0.788	0.895	0.848
	(3)	50	10	0.479	0.325	0.637	0.459	0.965	0.938
	(1)	100	5	0.850	0.717	0.830	0.693	0.932	0.886
	(1)	100	10	0.298	0.168	0.428	0.276	0.980	0.955
	(2)	100	5	0.985	0.947	0.974	0.922	0.992	0.985
	(2)	100	10	0.500	0.328	0.595	0.436	1.000	1.000
	(3)	100	5	0.995	0.983	0.989	0.977	0.998	0.993
	(3)	100	10	0.613	0.441	0.700	0.551	1.000	1.000
Ex 2.5.3	(1)	50	3	0.985	0.928	0.999	0.997	0.647	0.377
	(2)	50	3	1	1	1	1	1	1
	(1)	100	3	1	1	1	1	1	0.999
	(2)	100	3	1	1	1	1	1	1

Note: $\tilde{d}$ denotes the number of random variables d.

Table A.4: Empirical size and power for the bootstrap-assisted joint independence tests (based on the U-statistics) with c chosen according to the heuristic idea discussed in Remark 2.2.3. The results are obtained based on 1000 replications and the number of bootstrap resamples is taken to be 500.

				c		$\widetilde{JdCov^2}$		$\widetilde{JdCov^2_S}$		$\widetilde{JdCov^2_R}$	
		n	$\tilde{d}$	10%	5%	10%	5%	10%	5%	10%	5%
Ex 2.5.1	(1)	50	5	1.646	1.724	0.103	0.053	0.107	0.055	0.107	0.053
	(1)	50	10	1.657	1.732	0.101	0.049	0.106	0.056	0.095	0.052
	(2)	50	5	0.440	0.533	0.116	0.055	0.114	0.058	0.110	0.052
	(2)	50	10	1.519	1.636	0.099	0.052	0.087	0.045	0.094	0.050
	(3)	50	5	0.438	0.527	0.050	0.020	0.113	0.048	0.110	0.052
	(3)	50	10	0.438	0.527	0.027	0.011	0.094	0.051	0.107	0.048
	(1)	100	5	1.657	1.731	0.102	0.047	0.105	0.054	0.089	0.046
	(1)	100	10	1.656	1.731	0.108	0.049	0.101	0.046	0.101	0.060
	(2)	100	5	0.438	0.527	0.112	0.063	0.109	0.058	0.098	0.044
	(2)	100	10	0.484	0.620	0.104	0.064	0.104	0.048	0.116	0.066
	(3)	100	5	0.438	0.527	0.082	0.039	0.116	0.070	0.098	0.044
	(3)	100	10	0.438	0.527	0.051	0.020	0.100	0.051	0.107	0.058
Ex 2.5.2	(1)	50	5	1.646	1.724	0.637	0.517	0.603	0.484	0.630	0.502
	(1)	50	10	1.657	1.732	0.651	0.517	0.529	0.403	0.718	0.600
	(2)	50	5	1.646	1.724	0.842	0.761	0.815	0.728	0.844	0.760
	(2)	50	10	1.657	1.732	0.906	0.834	0.801	0.706	0.948	0.909
	(3)	50	5	1.646	1.724	0.901	0.844	0.882	0.819	0.889	0.845
	(3)	50	10	1.657	1.732	0.928	0.871	0.843	0.766	0.957	0.919
	(1)	100	5	1.657	1.731	0.923	0.884	0.891	0.845	0.929	0.883
	(1)	100	10	1.656	1.731	0.951	0.905	0.867	0.778	0.982	0.953
	(2)	100	5	1.657	1.731	0.990	0.986	0.985	0.977	0.992	0.985
	(2)	100	10	1.656	1.731	0.986	0.982	0.976	0.962	1.000	1.000
	(3)	100	5	1.657	1.731	0.998	0.996	0.996	0.990	0.996	0.992
	(3)	100	10	1.656	1.731	0.991	0.984	0.974	0.965	1.000	0.999
Ex 2.5.3	(1)	50	3	0.554	0.729	0.984	0.962	0.998	0.994	0.899	0.843
	(2)	50	3	0.438	0.527	1.000	1.000	1.000	1.000	1.000	1.000
	(1)	100	3	.439	0.530	1.000	1.000	1.000	1.000	1.000	1.000
	(2)	100	3	0.438	0.527	1.000	1.000	1.000	1.000	1.000	1.000
Ex 2.5.4	(1)	100	5	1.427	1.545	0.300	0.166	0.417	0.314	0.131	0.064
	(1)	100	10	1.589	1.680	0.090	0.023	0.204	0.111	0.050	0.018
	(2)	100	5	0.537	0.659	0.398	0.293	0.468	0.375	0.233	0.129
	(2)	100	10	1.040	1.198	0.106	0.040	0.212	0.132	0.058	0.018
	(1)	200	5	1.177	1.350	0.681	0.568	0.804	0.720	0.250	0.153
	(1)	200	10	1.503	1.609	0.221	0.117	0.340	0.234	0.086	0.040
	(2)	200	5	0.440	0.532	0.709	0.621	0.681	0.603	0.539	0.392
	(2)	200	10	0.606	0.735	0.290	0.179	0.381	0.277	0.149	0.065

Note: In Examples 2.5.1-2.5.3, $\tilde{d}$ denotes the number of random variables d. In Example 2.5.4, $\tilde{d}$ stands for p.

APPENDIX B

This is the Appendix for Chapter 3.

The appendix is organized as follows. In Section B.1 we explore our proposed homogeneity and dependence metrics in the low-dimensional setup. In Section B.2 we study the asymptotic behavior of our proposed homogeneity and dependence metrics in the high dimension medium sample size (HDMSS) framework where both the dimension(s) and the sample size(s) grow. Section B.3 illustrates an additional real data example for testing for independence in the high-dimensional framework. Finally, Section B.4 contains additional proofs of the main results in Chapter 3 and Sections B.1 and B.2 in the appendix.

B.1 Low-dimensional setup

In this section we illustrate that the new class of homogeneity metrics proposed in Chapter 3 inherits all the nice properties of generalized energy distance and MMD in the low-dimensional setting. Likewise, the proposed dependence metrics inherit all the desirable properties of generalized dCov and HSIC in the low-dimensional framework.

B.1.1 Homogeneity metrics

Note that in either Case S1 or S2, the Euclidean space equipped with distance K is of strong negative type. As a consequence, we have the following result.

THEOREM **11**. *$\mathcal{E}(X,Y)=0$ if and only if $X \overset{d}{=} Y$, in other words $\mathcal{E}(X,Y)$ completely characterizes the homogeneity of the distributions of X and Y.*

The following proposition shows that $\mathcal{E}_{n,m}(X,Y)$ is a two-sample U-statistic and an unbiased estimator of $\mathcal{E}(X,Y)$.

PROPOSITION **B.1.1**. *The U-statistic type estimator enjoys the following properties:*

1. *$\mathcal{E}_{n,m}$ is an unbiased estimator of the population $\mathcal{E}$.*

2. *$\mathcal{E}_{n,m}$ admits the following form :*

$$\mathcal{E}_{n,m}(X,Y) \;=\; \frac{1}{\binom{n}{2}\binom{m}{2}} \sum_{1\leq i<j\leq n}\,\sum_{1\leq k<l\leq m} h(X_i,X_j;Y_k,Y_l)\,,$$

where

$$\begin{aligned} h(X_i,X_j;Y_k,Y_l) \;=\; &\frac{1}{2}\Big(K(X_i,Y_k) \,+\, K(X_i,Y_l) \,+\, K(X_j,Y_k) \,+\, K(X_j,Y_l)\Big) \\ &-\, K(X_i,X_j) \,-\, K(Y_k,Y_l)\,. \end{aligned}$$

The following theorem shows the asymptotic behavior of the U-statistic type estimator of $\mathcal{E}$ for fixed p and growing n.

THEOREM **12**. *Under Assumption 3.3.5 and the assumption that* $\sup_{1\leq i\leq p}\mathbb{E}\rho_i(X_{(i)},0_{d_i})<\infty$ *and* $\sup_{1\leq i\leq p}\mathbb{E}\rho_i(Y_{(i)},0_{d_i})<\infty$*, as $m,n\to\infty$ with p remaining fixed, we have the following:*

1. *$\mathcal{E}_{n,m}(X,Y) \xrightarrow{a.s.} \mathcal{E}(X,Y)$.*

2. *When $X \overset{d}{=} Y$, $\mathcal{E}_{n,m}$ has degeneracy of order $(1,1)$, and*

$$\frac{(m-1)(n-1)}{n+m}\,\mathcal{E}_{n,m}(X,Y) \xrightarrow{d} \sum_{k=1}^{\infty}\lambda_k^2\left(Z_k^2\,-\,1\right)\,,$$

where $\{Z_k\}$ is a sequence of independent $N(0,1)$ random variables and λ_k's depend on the distribution of (X,Y).

Proposition B.1.1, Theorem 11 and Theorem 12 demonstrate that $\mathcal{E}$ inherits all the nice properties of generalized energy distance and MMD in the low-dimensional setting.

B.1.2 Dependence metrics

Note that Proposition 3.2.1 in Section 3.2 and Proposition 3.7 in Lyons (2013) ensure that $\mathcal{D}(X,Y)$ completely characterizes independence between X and Y, which leads to the following result.

THEOREM **13**. *Under Assumption 4.3.2, $\mathcal{D}(X,Y)=0$ if and only if $X \perp\!\!\!\perp Y$.*

The following proposition shows that $\widetilde{\mathcal{D}_n^2}(X,Y)$ is an unbiased estimator of $\mathcal{D}^2(X,Y)$ and is a U-statistic of order four.

PROPOSITION **B.1.2**. *The U-statistic type estimator $\widetilde{\mathcal{D}_n^2}$ (defined in (1.14) in Chapter 3) has the following properties:*

1. *$\widetilde{\mathcal{D}_n^2}$ is an unbiased estimator of the squared population $\mathcal{D}^2$.*

2. *$\widetilde{\mathcal{D}_n^2}$ is a fourth-order U-statistic which admits the following form:*

$$\widetilde{\mathcal{D}_n^2} = \frac{1}{\binom{n}{4}} \sum_{i<j<k<l} h_{i,j,k,l}\,,$$

where

$$\begin{aligned} h_{i,j,k,l} &= \frac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,k,l)} (d^X_{st} d^Y_{st} + d^X_{st} d^Y_{uv} - 2 d^X_{st} d^Y_{su}) \\ &= \frac{1}{6} \sum_{s<t,u<v}^{(i,j,k,l)} (d^X_{st} d^Y_{st} + d^X_{st} d^Y_{uv}) - \frac{1}{12} \sum_{(s,t,u)}^{(i,j,k,l)} d^X_{st} d^Y_{su}\,, \end{aligned}$$

the summation is over all possible permutations of the 4-tuple of indices (i,j,k,l). For example, when $(i,j,k,l)=(1,2,3,4)$, there exist 24 permutations, including $(1,2,3,4),\ldots,$ $(4,3,2,1)$. Furthermore, $\widetilde{\mathcal{D}_n^2}$ has degeneracy of order 1 when X and Y are independent.

The following theorem shows the asymptotic behavior of the U-statistic type estimator of $\mathcal{D}^2$ for fixed p, q and growing n.

THEOREM **14**. *Under Assumption 4.3.2, with fixed p,q and $n \to \infty$, we have the following as $n \to \infty$:*

1. *$\widetilde{\mathcal{D}_n^2}(X,Y) \xrightarrow{a.s.} \mathcal{D}^2(X,Y)$;*

2. *When $\mathcal{D}^2(X,Y)=0$ (i.e., $X \perp\!\!\!\perp Y$), $n\,\widetilde{\mathcal{D}^2_n}(X,Y) \xrightarrow{d} \sum_{i=1}^{\infty} \tilde{\lambda}_i^2(Z_i^2-1)$, where $Z_i's$ are i.i.d. standard normal random variables and $\tilde{\lambda}_i$'s depend on the distribution of (X,Y);*

3. *When $\mathcal{D}^2(X,Y)>0$, $n\,\widetilde{\mathcal{D}^2_n}(X,Y) \xrightarrow{a.s.} \infty$.*

Proposition B.1.2, Theorem 13 and Theorem 14 demonstrate that in the low-dimensional setting, $\mathcal{D}$ inherits all the nice properties of generalized dCov and HSIC.

B.2 High dimension medium sample size (HDMSS)

B.2.1 Homogeneity metrics

In this subsection, we consider the HDMSS setting where $p \to \infty$ and $n, m \to \infty$ at a slower rate than p. Under H_0, we impose the following conditions to obtain the asymptotic null distribution of the statistic $T_{n,m}$ under the HDMSS setup.

ASSUMPTION **B.2.1**. *As n, m and $p \to \infty$,*

$$\frac{1}{n^2}\,\frac{\mathbb{E}\,[H^4(X,X')]}{\left(\mathbb{E}\,[H^2(X,X')]\right)^2} = o(1), \quad \frac{1}{n}\,\frac{\mathbb{E}\,[H^2(X,X'')\,H^2(X',X'')]}{\left(\mathbb{E}\,[H^2(X,X')]\right)^2} = o(1),$$
$$\frac{\mathbb{E}\,[H(X,X'')\,H(X',X'')\,H(X,X''')\,H(X',X''')]}{\left(\mathbb{E}\,[H^2(X,X')]\right)^2} = o(1).$$

REMARK **B.2.1**. *We refer the reader to Section 2.2 in Zhang et al. (2018) and Remark A.2.2 in Zhu et al. (2020) for illustrations of Assumption B.2.1 where ρ_i has been considered to be the Euclidean distance or the squared Euclidean distance, respectively, for $1 \le i \le p$.*

ASSUMPTION **B.2.2**. *Suppose $\mathbb{E}\,[L^2(X,X')] = O(\alpha_p^2)$ where α_p is a positive real sequence such that $\tau_X \alpha_p^2 = o(1)$ as $p \to \infty$. Further assume that as $n, p \to \infty$,*

$$\frac{n^4\,\tau_X^4\,\mathbb{E}\,[R^4(X,X')]}{\left(\mathbb{E}\,[H^2(X,X')]\right)^2} = o(1)\,.$$

REMARK **B.2.2**. *We refer the reader to Remark 3.3.1 in Chapter 3 which illustrates some sufficient conditions under which $\alpha_p = O(\frac{1}{\sqrt{p}})$ and consequently $\tau_X \alpha_p^2 = o(1)$ holds, as $\tau_X \asymp p^{1/2}$. In* *similar lines of Remark B.4.1 in Section B.4 of the appendix, it can be argued that* $\mathbb{E}\left[R^4(X,X')\right] = O\left(\frac{1}{p^4}\right)$*. If we further assume that Assumption 3.3.4 holds, then we have* $\mathbb{E}\left[H^2(X,X')\right] \asymp 1$*. Combining all the above, it is easy to verify that* $\frac{n^4\,\tau_X^4\,\mathbb{E}\left[R^4(X,X')\right]}{(\mathbb{E}\left[H^2(X,X')\right])^2} = o(1)$ *holds provided* $n = o(p^{1/2})$.

The following theorem illustrates the limiting null distribution of $T_{n,m}$ under the HDMSS setup. We refer the reader to Section B.4 of the appendix for a detailed proof.

Theorem **15**. *Under* H_0 *and Assumptions 3.3.5, B.2.1 and B.2.2, as* n, m *and* $p \to \infty$*, we have*

$$T_{n,m} \xrightarrow{d} N(0,1).$$

B.2.2 Dependence metrics

In this subsection, we consider the HDMSS setting where $p, q \to \infty$ and $n \to \infty$ at a slower rate than p, q. The following theorem shows that similar to the HDLSS setting, under the HDMSS setup, $\widetilde{\mathcal{D}}_n^2$ is asymptotically equivalent to the aggregation of group-wise generalized dCov. In other words $\widetilde{\mathcal{D}}_n^2(X,Y)$ can quantify group-wise nonlinear dependence between X and Y in the HDMSS setup as well.

Assumption **B.2.3**. $\mathbb{E}[L_X(X,X')^2] = \alpha_p^2$*,* $\mathbb{E}[L_X(X,X')^4] = \gamma_p^2$*,* $\mathbb{E}[L_Y(Y,Y')^2] = \beta_q^2$ *and* $\mathbb{E}[L_Y(Y,Y')^4] = \lambda_q^2$*, where* $\alpha_p, \gamma_p, \beta_q, \lambda_q$ *are positive real sequences satisfying* $n\alpha_p = o(1)$*,* $n\beta_q = o(1)$*,* $\tau_X^2(\alpha_p\gamma_p + \gamma_p^2) = o(1)$*,* $\tau_Y^2(\beta_q\lambda_q + \lambda_q^2) = o(1)$*, and* $\tau_{XY}\,(\alpha_p\lambda_q + \gamma_p\beta_q + \gamma_p\lambda_q) = o(1)$*.*

Remark **B.2.3**. *Following Remark 3.3.1 in Chapter 3, we can write* $L(X,X') = O(\frac{1}{p})\sum_{i=1}^{p} (Z_i - \mathbb{E}\,Z_i)$*, where* $Z_i = \rho_i(X_{(i)}, X'_{(i)})$ *for* $1 \le i \le p$*. Assume that* $\sup_{1\le i\le p}\mathbb{E}\,\rho_i^4(X_{(i)}, 0_{d_i}) < \infty$*, which implies* $\sup_{1\le i\le p}\mathbb{E}\,Z_i^4 < \infty$*. Under certain weak dependence assumptions, it can be shown that* $\mathbb{E}\left(\sum_{i=1}^{p}(Z_i - \mathbb{E}\,Z_i)\right)^4 = O(p^2)$ *as* $p \to \infty$ *(see for example Theorem 1 in Doukhan et al. (1999)). Therefore we have* $\mathbb{E}[L(X,X')^4] = O(\frac{1}{p^2})$*. It follows from Hölder's inequality that* $\mathbb{E}[L(X,X')^2] = O(\frac{1}{p})$*. Similar arguments can be made about* $\mathbb{E}[L(Y,Y')^4]$ *and* $\mathbb{E}[L(Y,Y')^2]$ *as well.*

Theorem **16**. *Under Assumptions 3.3.2 and B.2.3, we can show that*

$$\widetilde{\mathcal{D}_n^2}(X,Y) = \frac{1}{4\tau_{XY}} \sum_{i=1}^{p}\sum_{j=1}^{q} \widetilde{D^2_{n\,;\,\rho_i,\rho_j}}(X_{(i)},Y_{(j)}) \,+\, \mathcal{R}_n\;, \tag{B.1}$$

where $\mathcal{R}_n$ is the remainder term satisfying that $\mathcal{R}_n = O_p(\tau_{XY}\,(\alpha_p\lambda_q + \gamma_p\beta_q + \gamma_p\lambda_q)) = o_p(1)$, i.e., $\mathcal{R}_n$ is of smaller order compared to the leading term and hence is asymptotically negligible.

The following theorem states the asymptotic null distribution of the studentized test statistic $\mathcal{T}_n$ (given in equation (3.26) in Chapter 3) under the HDMSS setup. Define

$$U(X_k,X_l) := \frac{1}{\tau_X}\sum_{i=1}^{p} d^X_{kl}(i), \quad \text{and} \quad V(Y_k,Y_l) := \frac{1}{\tau_Y}\sum_{i=1}^{q} d^Y_{kl}(i).$$

Assumption **B.2.4**. *Assume that*

$$\frac{\mathbb{E}\,[U(X,X')]^4}{\sqrt{n}\,(\mathbb{E}[U(X,X')]^2)^2} = o(1),$$
$$\frac{\mathbb{E}\,[U(X,X')\,U(X',X'')\,U(X'',X''')\,U(X''',X)]}{(\mathbb{E}[U(X,X')]^2)^2} = o(1),$$

and the same conditions hold for Y in terms of $V(Y,Y')$.

Remark **B.2.4**. *We refer the reader to Section 2.2 in Zhang et al. (2018) and Remark A.2.2 in Zhu et al. (2020) for illustrations of Assumption B.2.1 where ρ_i has been considered to be the Euclidean distance or the squared Euclidean distance, respectively.*

We can show that under H_0, the studentized test $\mathcal{T}_n$ converge to the standard normal distribution under the HDMSS setup.

Theorem **17**. *Under H_0 and Assumptions B.2.3-B.2.4, as $n,p,q \to \infty$, we have $\mathcal{T}_n \xrightarrow{d} N(0,1)$.*

B.3 Additional real data example

We consider the monthly closed stock prices of $\tilde{p} = 36$ companies under the transport sector and $\tilde{q} = 41$ companies under the utilities sector between January 1, 2017 and December 31,

2018. The companies under both the sectors are clustered or grouped according to their countries. The data has been downloaded from Yahoo Finance via the R package 'quantmod'. Under the transport sector, we have $q = 14$ countries or groups, viz. USA, Brazil, Canada, Greece, China, Panama, Belgium, Bermuda, UK, Mexico, Chile, Monaco, Ireland and Hong Kong, with $d = (5, 1, 2, 8, 4, 1, 1, 3, 1, 3, 1, 4, 1, 1)$. And under the utilities sector, we have $q = 21$ countries or groups, viz. USA, Mexico, UK, India, Canada, China, Hong Kong, Taiwan, Brazil, Cayman Islands, Israel, Argentina, Chile, Singapore, South Korea, Russia, France, Phillipines, Indonesia, Spain and Turkey, with $g = (5, 1, 3, 1, 5, 2, 3, 1, 4, 1, 1, 4, 1, 1, 2, 1, 1, 1, 1, 1, 1)$. At each time t, denote the closed stock prices of these companies from the two different sectors by $X_t = (X_{1t}, \ldots, X_{pt})$ and $Y_t = (Y_{1t}, \ldots, Y_{qt})$ for $1 \leq t \leq 24$. We consider the stock returns $S_t^X = (S_{1t}^X, \ldots, S_{pt}^X)$ and $S_t^Y = (S_{1t}^Y, \ldots, S_{qt}^Y)$ for $1 \leq t \leq 23$, where $S_{itl}^X = \log \frac{X_{i,t+1,l}}{X_{itl}}$ and $S_{jtl'}^Y = \log \frac{Y_{j,t+1,l'}}{Y_{jtl'}}$ for $1 \leq l \leq d_i$, $1 \leq i \leq p$, $1 \leq l' \leq g_j$ and $1 \leq j \leq q$.

The intuitive idea is, stock returns of transport companies should affect the stock returns of companies under the utilities sector, and here both the random vectors admit a natural grouping based on the countries. Table B.1 below shows the p-values corresponding to the different tests for independence between $\{S_t^X\}_{t=1}^{23}$ and $\{S_t^Y\}_{t=1}^{23}$. The tests based on the proposed dependence metrics considering the natural grouping deliver much smaller p-values compared to the tests based on the usual dCov and HSIC, as well as the projection correlation based test, which fail to reject the null hypothesis of independence between $\{S_t^X\}_{t=1}^{23}$ and $\{S_t^Y\}_{t=1}^{23}$. This makes intuitive sense as the dependence among financial asset returns is usually nonlinear in nature and thus cannot be fully characterized by the usual dCov and HSIC in the high dimensional setup.

Table B.1: p-values corresponding to the different tests for cross-sector independence of stock returns data considering the natural grouping based on countries.

I	II	III	IV	V	VI	VII
0.0008	0.0011	0.0004	0.1106	0.1129	0.4848	0.1120

Table B.2 below shows the p-values corresponding to the different tests for independence when we disregard the natural grouping and consider $d_i = 1$ and $g_j = 1$ for all $1 \leq i \leq p$ and $1 \leq j \leq q$. Considering unit group sizes makes our proposed statistics essentially equivalent to the marginal aggregation approach proposed by Zhu et al. (2020). In this case the proposed tests have higher p-values than when we consider the natural grouping, indicating that grouping or clustering might improve the power of testing as they are capable of detecting a wider range of dependencies.

Table B.2: p-values corresponding to the different tests for cross-sector independence of stock returns data considering unit group sizes.

I	II	III	IV	V	VI	VII
0.0067	0.0532	0.0796	0.1106	0.1129	0.4848	0.1120

B.4 Technical Appendix

Proof of Proposition 3.2.1. To prove (1), note that if d is a metric on a space $\mathcal{X}$, then so is $d^{1/2}$. It is easy to see that K^2 is a metric on $\mathbb{R}^{\tilde{p}}$. To prove (2), note that $(\mathbb{R}^{d_i}, \rho_i)$ has strong negative type for $1 \leq i \leq p$. The rest follows from Corollary 3.20 in Lyons (2013). $\diamondsuit$

Proof of Proposition B.1.1. It is easy to verify that $\mathcal{E}_{n,m}$ is an unbiased estimator of $\mathcal{E}$ and is a two-sample U-statistic with the kernel h. $\diamondsuit$

Proof of Theorem 12. The first part of the proof follows from Theorem 1 in Sen (1977) and the observation that $\mathbb{E}\,\left[|h| \log^+ |h|\right] \leq \mathbb{E}[h^2]$. The power mean inequality says that for $a_i \in \mathbb{R}$, $1 \leq i \leq n$, $n \geq 2$ and $r > 1$,

$$\left|\sum_{i=1}^{n} a_i\right|^r \;\leq\; n^{r-1} \sum_{i=1}^{n} |a_i|^r \,. \tag{B.2}$$

Using the power mean inequality, it is easy to see that the assumptions $\sup_{1\leq i\leq p} \mathbb{E}\rho_i(X_{(i)}, 0_{d_i}) < \infty$ and $\sup_{1\leq i\leq p} \mathbb{E}\rho_i(Y_{(i)}, 0_{d_i}) < \infty$ ensure that $\mathbb{E}[h^2] < \infty$. For proving the second part, define

$h_{1,0}(X) = \mathbb{E}\,[h(X,X';Y,Y')|X]$ and $h_{0,1}(Y) = \mathbb{E}\,[h(X,X';Y,Y')|Y]$ Clearly, when $X \overset{d}{=} Y$, $h_{1,0}(X)$ and $h_{0,1}(Y)$ are degenerate at 0 almost surely. Following Theorem 1.1 in Neuhaus (1977), we have

$$\frac{(m-1)(n-1)}{n+m}\,\mathcal{E}_{nm}(X,Y) \xrightarrow{d} \sum_{k=1}^{\infty} \sigma_k^2 \left[(a_k U_k + b_k V_k)^2 - (a_k^2 + b_k^2)\right] ,$$

where $\{U_k\}, \{V_k\}$ are two sequences of independent $N(0,1)$ variables, independent of each other, and (σ_k, a_k, b_k)'s depend on the distribution of (X,Y). The proof can be completed by some simple rearrangement of terms. ♢

Proof of Proposition 3.3.1. The proof is essentially similar to the proof of Proposition 2.1.1 in Zhu et al. (2020), replacing the Euclidean distance between, for example, X and X', viz. $\|X - X'\|_{\tilde{p}}$, by the new distance metric $K(X,X')$. To show that $R(X,X') = O_p(L^2(X,X'))$ if $L(X,X') = o_p(1)$, we define $f(x) = \sqrt{1+x}$. By the definition of the Lagrange's form of the remainder term from Taylor's expansion, we have

$$R(X,X') = \int_0^{L(X,X')} f''(t)\,(\,L(X,X') - t\,)\;dt\,.$$

Using R and L interchangeably with $R(X,X')$ and $L(X,X')$ respectively, we can write

$$\begin{aligned} |R| \;&\le\; |L| \left[\int_0^L f''(t)\,\mathbb{1}_{L>0}\,dt \;+\; \int_L^0 f''(t)\,\mathbb{1}_{L<0}\,dt\right] \\ &=\; \frac{|L|}{2}\,\Big|1 - \frac{1}{\sqrt{1+L}}\Big| \\ &=\; \frac{|L|}{2}\,\frac{|L|}{1+L+\sqrt{1+L}} \\ &\le\; \frac{L^2}{2(1+L)}\,. \end{aligned} \tag{B.3}$$

It is clear that $R(X,X') = O_p(L^2(X,X'))$ provided that $L(X,X') = o_p(1)$. ♢

Proof of Theorem 1. Observe that $\mathbb{E}\,L(X,Y) = \mathbb{E}\,L(X,X') = \mathbb{E}\,L(Y,Y') = 0$. By Proposition

3.3.1,

$$\begin{aligned}\mathcal{E}(X,Y) &= 2\,\mathbb{E}\,[\tau + \tau\,R(X,Y)] \;-\; \mathbb{E}\,[\tau_X + \tau_X\,R(X,X')] \;-\; \mathbb{E}\,[\tau_Y + \tau_Y\,R(Y,Y')] \\ &= 2\tau - \tau_X - \tau_Y \;+\; \mathcal{R}_{\mathcal{E}}\,.\end{aligned}$$

Clearly $|\mathcal{R}_{\mathcal{E}}| \;\leq\; 2\,\tau\,\mathbb{E}\,[\,|R(X,Y)|\,] \;+\; \tau_X\,\mathbb{E}\,[\,|R(X,X')|\,] \;+\; \tau_Y\,\mathbb{E}\,[\,|R(Y,Y')|\,]$. By (B.3) and Assumption 3.3.3, we have

$$\tau|R(X,Y)| \leq \frac{\tau L^2(X,Y)}{2(1+L(X,Y))} = O(\tau a_p^2) = o_p(1).$$

As $\{\sqrt{p}L^2(X,Y)/(1+L(X,Y))\}$ is uniformly integrable and $\tau \asymp \sqrt{p}$, we must have $\tau\mathbb{E}[|R(X,Y)|] = o(1)$. The other terms can be handled in a similar fashion. $\diamondsuit$

REMARK **B.4.1**. *Write* $L(X,Y) = \frac{1}{\tau^2}(A_p - \mathbb{E}\,A_p) = \frac{1}{\tau^2}\sum_{i=1}^{p}(Z_i - \mathbb{E}Z_i)$*, where* $A_p := \sum_{i=1}^{p} Z_i$ *and* $Z_i := \rho_i(X_i, Y_i)$ *for* $1 \leq i \leq p$*. Assume* $\sup_i \mathbb{E}\rho_i^8(X_i, 0_{d_i}) < \infty$ *and* $\sup_i \mathbb{E}\rho_i^8(X_i, 0_{d_i}) < \infty$*, which imply* $\sup_i \mathbb{E}Z_i^8 < \infty$*. Denote* $L(X,Y)$ *by* L *and* $R(X,Y)$ *by* R *for notational simplicities. Further assume that* $E\exp(tA_p) = O((1-\theta_1 t)^{-\theta_2 p})$ *for* $\theta_1, \theta_2 > 0$ *and* $\theta_2\, p > 4$ *uniformly over* $t < 0$ *(which is clearly satisfied when* Z_i*'s are independent and* $\mathbb{E}\exp(tZ_i) \leq a_1(1-a_2 t)^{-a_3}$ *uniformly over* $t < 0$ *and* $1 \leq i \leq p$ *for some* $a_1, a_2, a_3 > 0$ *with* $a_3\, p > 4$*). Under certain weak dependence assumptions, it can be shown that:*

1. $\{\sqrt{p}L^2/(1+L)\}$ *is uniformly integrable;*

2. $\mathbb{E}\,R^2 = O(\frac{1}{p^2})$*.*

Similar arguments hold for $L(X,X')$ *and* $R(X,X')$*, and,* $L(Y,Y')$ *and* $R(Y,Y')$ *as well.*

Proof of Remark B.4.1. To prove the first part, define $L_p := \sqrt{p}L^2/(1+L)$. Following Chapter 6 of Resnick (1999), it suffices to show that $\sup_p \mathbb{E}\,L_p^2 < \infty$. Towards that end, using Hölder's inequality we observe

$$\mathbb{E}\, L_p^2 \;\leq \big(\mathbb{E}(p^2 L^8)\big)^{1/2} \left(\mathbb{E}\Big[\frac{1}{(1+L)^4}\Big]\right)^{1/2} . \tag{B.4}$$

With $\sup_i \mathbb{E} Z_i^8 < \infty$ and under certain weak dependence assumptions, it can be shown that $\mathbb{E}(A_p - \mathbb{E}A_p)^8 = O(p^4)$ (see for example Theorem 1 in Doukhan et al. (1999)). Consequently we have $\mathbb{E}\, L^8 = O(\frac{1}{p^4})$, as $\tau \asymp \sqrt{p}$. Clearly this yields $\mathbb{E}\,(p^2 L^8) = O(\frac{1}{p^2})$.

Now note that

$$\mathbb{E}\Big[\frac{1}{(1+L)^4}\Big] \;=\; \tau^8\, \mathbb{E}\left(\frac{1}{A_p^4}\right) . \tag{B.5}$$

Equation (3) in Cressie et al. (1981) states that for a non-negative random variable U with moment-generating function $M_U(t) = \mathbb{E}\exp(tU)$, one can write

$$\mathbb{E}(U^{-k}) = (\Gamma(k))^{-1} \int_0^\infty t^{k-1} M_U(-t)\, dt\,, \tag{B.6}$$

for any positive integer k, provided both the integrals exist. Using equation (B.6), the assumptions stated in Remark B.4.1 and basic properties of beta integrals, some straightforward calculations yield

$$\mathbb{E}\left(\frac{1}{A_p^4}\right) \;\leq\; C_1 \int_0^\infty \frac{t^{4-1}}{(1+\theta_1 t)^{\theta_2 p}}\, dt \;=\; C_2\, \frac{\Gamma(\theta_2 p - 4)}{\Gamma(\theta_2 p)}\,, \tag{B.7}$$

where C_1, C_2 are positive constants, which clearly implies that $\mathbb{E}\left(\frac{1}{A_p^4}\right) = O(\frac{1}{p^4})$. This together with equation (B.5) implies that $\mathbb{E}\Big[\frac{1}{(1+L)^4}\Big] = O(1)$, as $\tau \asymp \sqrt{p}$.

Combining all the above, we get from (B.4) that $\mathbb{E}\, L_p^2 = O(\frac{1}{p})$ and therefore $\sup_p \mathbb{E}\, L_p^2 < \infty$, which completes the proof of the first part.

To prove the second part, note that following the proof of Proposition 3.3.1 and Hölder's inequality we can write

$$\mathbb{E}\, R^2 = O\left(\mathbb{E}\,\left[\frac{L^4}{(1+L)^2}\right]\right) = O\left(\big(\mathbb{E}(L^8)\big)^{1/2}\,\left(\mathbb{E}\Big[\frac{1}{(1+L)^4}\Big]\right)^{1/2}\right). \tag{B.8}$$

Following the arguments as in the proof of the first part, clearly we have $\mathbb{E}\, L^8 = O(\frac{1}{p^4})$ and $\mathbb{E}\Big[\frac{1}{(1+L)^4}\Big] = O(1)$. From this and equation (B.8), it is straightforward to verify that $\mathbb{E}\, R^2 = O(\frac{1}{p^2})$, which completes the proof of the second part. ♢

Proof of Lemma 3.3.1. To see (2), first observe that the sufficient part is straightforward from equation (3.8) in Chapter 3. For the necessary part, denote $a = \operatorname{tr}\Sigma_X$, $b = \operatorname{tr}\Sigma_Y$ and $c = \|\mu_X - \mu_Y\|^2$. Then we have $2\sqrt{a+b+c} = \sqrt{2a} + \sqrt{2b}$. Some straightforward calculations yield $(\sqrt{2a} - \sqrt{2b})^2 + 4\,c = 0$ which implies the rest.

To see (1), again the sufficient part is straightforward from equation (3.7) in Chapter 3 and the form of K given in equation (3.2) in Chapter 3. For the necessary part, first note that as $(\mathbb{R}^{d_i}, \rho_i)$ is a metric space of strong negative type for $1 \le i \le p$, there exists a Hilbert space $\mathcal{H}_i$ and an injective map $\phi_i : \mathbb{R}^{d_i} \to \mathcal{H}_i$ such that $\rho_i(z, z') = \|\phi_i(z) - \phi_i(z')\|^2_{\mathcal{H}_i}$, where $\langle\cdot,\cdot\rangle_{\mathcal{H}_i}$ is the inner product defined on $\mathcal{H}_i$ and $\|\cdot\|_{\mathcal{H}_i}$ is the norm induced by the inner product (see Proposition 3 in Sejdinovic et al. (2013) for detailed discussions). Further, if k_i is a distance-induced kernel induced by the metric ρ_i, then by Proposition 14 in Sejdinovic et al. (2013), $\mathcal{H}_i$ is the RKHS with the reproducing kernel k_i and $\phi_i(z)$ is essentially the canonical feature map for $\mathcal{H}_i$, viz. $\phi_i(z) : z \mapsto k_i(\cdot, z)$. It is easy to see that

$$\begin{aligned}
\tau_X^2 &= \mathbb{E}\sum_{i=1}^p \|\phi_i(X_{(i)}) - \phi_i(X'_{(i)})\|^2_{\mathcal{H}_i} \;=\; 2\,\mathbb{E}\sum_{i=1}^p \|\phi_i(X_{(i)}) - \mathbb{E}\,\phi_i(X_{(i)})\|^2_{\mathcal{H}_i},\\
\tau_Y^2 &= \mathbb{E}\sum_{i=1}^p \|\phi_i(Y_{(i)}) - \phi_i(Y'_{(i)})\|^2_{\mathcal{H}_i} \;=\; 2\,\mathbb{E}\sum_{i=1}^p \|\phi_i(Y_{(i)}) - \mathbb{E}\,\phi_i(Y_{(i)})\|^2_{\mathcal{H}_i},\\
\tau^2 &= \mathbb{E}\sum_{i=1}^p \|\phi_i(X_{(i)}) - \phi_i(Y_{(i)})\|^2_{\mathcal{H}_i} \;=\; \tau_X^2/2 + \tau_Y^2/2 + \zeta^2,
\end{aligned}$$

where $\zeta^2 = \sum_{i=1}^p \|\mathbb{E}\,\phi(X_{(i)}) - \mathbb{E}\,\phi(Y_{(i)})\|^2_{\mathcal{H}_i}$. Thus $2\tau - \tau_X - \tau_Y = 0$ is equivalent to

$$4(\tau_X^2/2 + \tau_Y^2/2 + \zeta^2) = (\tau_X + \tau_Y)^2 = \tau_X^2 + \tau_Y^2 + 2\tau_X\tau_Y.$$

which implies that

$$4\zeta^2 + (\tau_X - \tau_Y)^2 = 0.$$

Therefore, $2\tau - \tau_X - \tau_Y = 0$ holds if and only if (1) $\zeta = 0$, i.e., $\mathbb{E}\,\phi_i(X_{(i)}) = \mathbb{E}\,\phi_i(Y_{(i)})$ for all $1 \leq i \leq p$, and, (2) $\tau_X = \tau_Y$, i.e.,

$$\mathbb{E}\sum_{i=1}^{p} \|\phi_i(X_{(i)}) - \mathbb{E}\,\phi_i(X_{(i)})\|^2_{\mathcal{H}_i} \;=\; \mathbb{E}\sum_{i=1}^{p} \|\phi_i(Y_{(i)}) - \mathbb{E}\,\phi_i(Y_{(i)})\|^2_{\mathcal{H}_i}.$$

Now if $X \sim P$ and $Y \sim Q$, then note that

$$\mathbb{E}\,\phi_i(X_{(i)}) \;=\; \int_{\mathbb{R}^{d_i}} k_i(\cdot, z)\, dP_i(z) \;=\; \Pi_i(P_i) \quad \text{and} \quad \mathbb{E}\,\phi_i(Y_{(i)}) \;=\; \int_{\mathbb{R}^{d_i}} k_i(\cdot, z)\, dQ_i(z) \;=\; \Pi_i(Q_i)\,,$$

where Π_i is the mean embedding function (associated with the distance induced kernel k_i) defined in Section 1.2.1, P_i and Q_i are the distributions of $X_{(i)}$ and $Y_{(i)}$, respectively. As ρ_i is a metric of strong negative type on $\mathbb{R}^{d_i}$, the induced kernel k_i is characteristic to $\mathcal{M}_1(\mathbb{R}^{d_i})$ and hence the mean embedding function Π_i is injective. Therefore condition (1) above implies $X_{(i)} \overset{d}{=} Y_{(i)}$. $\diamondsuit$

Now we introduce some notation before presenting the proof of Theorem 2. The key of our analysis is to study the variance of the leading term of $\mathcal{E}_{n,m}(X, Y)$ in the HDLSS setup, propose the variance estimator and study the asymptotic behavior of the variance estimator. It will be shown later (in the proof of Theorem 2) that the leading term in the Taylor's expansion of $\mathcal{E}_{n,m}(X, Y) -$

$(2\tau - \tau_X - \tau_Y)$ can be written as $L_1 + L_2$, where

$$\begin{aligned} L_1 \ &:= \ \frac{1}{nm\tau}\sum_{k=1}^{n}\sum_{l=1}^{m}\sum_{i=1}^{p} d_{kl}(i) - \frac{1}{n(n-1)\tau_X}\sum_{k<l}\sum_{i=1}^{p} d_{kl}^X(i) - \frac{1}{m(m-1)\tau_Y}\sum_{k<l}\sum_{i=1}^{p} d_{kl}^Y(i) \\ &:= \ L_1^1 - L_1^2 - L_1^3 \ , \end{aligned} \tag{B.9}$$

where L_1^i's are defined accordingly and

$$\begin{aligned} L_2 \ := & \ \frac{1}{nm\tau}\sum_{k=1}^{n}\sum_{l=1}^{m}\sum_{i=1}^{p}\Big(\mathbb{E}\,[\rho_i(X_{k(i)}, Y_{l(i)})|X_{k(i)}] + [\rho_i(X_{k(i)}, Y_{l(i)})|Y_{l(i)}] - 2\,\mathbb{E}\,\rho_i(X_{k(i)}, Y_{l(i)})\Big) \\ & - \frac{1}{n(n-1)\tau_X}\sum_{k<l}\sum_{i=1}^{p}\Big(\mathbb{E}\,[\rho_i(X_{k(i)}, X_{l(i)})|X_{k(i)}] + [\rho_i(X_{k(i)}, X_{l(i)})|X_{l(i)}] - 2\,\mathbb{E}\,\rho_i(X_{k(i)}, X_{l(i)})\Big) \\ & - \frac{1}{m(m-1)\tau_Y}\sum_{k<l}\sum_{i=1}^{p}\Big(\mathbb{E}\,[\rho_i(Y_{k(i)}, Y_{l(i)})|Y_{k(i)}] + [\rho_i(Y_{k(i)}, Y_{l(i)})|Y_{l(i)}] - 2\,\mathbb{E}\,\rho_i(Y_{k(i)}, Y_{l(i)})\Big)\,. \end{aligned} \tag{B.10}$$

By the double-centering properties, it is easy to see that L_1^i for $1 \le i \le 3$ are uncorrelated. Define

$$\begin{aligned} V :=& \frac{1}{nm\tau^2}\sum_{i,i'=1}^{p}\mathbb{E}\,[d_{kl}(i)\,d_{kl}(i')] \ + \ \frac{1}{2n(n-1)\tau_X^2}\sum_{i,i'=1}^{p}\mathbb{E}\,[d_{kl}^X(i)\,d_{kl}^X(i')] \\ & + \ \frac{1}{2m(m-1)\tau_Y^2}\sum_{i,i'=1}^{p}\mathbb{E}\,[d_{kl}^Y(i)\,d_{kl}^Y(i')] \\ :=& \, V_1 \ + \ V_2 \ + \ V_3, \end{aligned} \tag{B.11}$$

where V_i's are defined accordingly. Further let

$$\widetilde{V_1} \ := \ nmV_1 \ , \ \widetilde{V_2} \ := \ 2n(n-1)V_2 \ , \ \widetilde{V_3} \ := \ 2m(m-1)V_3 \,. \tag{B.12}$$

It can be verified that

$$\mathbb{E}\,[d_{kl}^X(i)\,d_{kl}^X(i')] \ = \ D^2_{\rho_i,\rho_{i'}}(X_{(i)}, X_{(i')})\,.$$

Thus we have

$$\widetilde{V_2} \;=\; \frac{1}{\tau_X^2}\sum_{i,i'=1}^{p} D^2_{\rho_i,\rho_{i'}}(X_{(i)},X_{(i')}) \quad \text{and} \quad \widetilde{V_3} \;=\; \frac{1}{\tau_Y^2}\sum_{i,i'=1}^{p} D^2_{\rho_i,\rho_{i'}}(Y_{(i)},Y_{(i')})\,. \tag{B.13}$$

We study the variances of L_1^i for $1 \leq i \leq 3$ and propose some suitable estimators. The variance for L_1^2 is given by

$$var(L_1^2) \;=\; \frac{1}{n^2(n-1)^2\tau_X^2}\sum_{i,i'=1}^{p}\sum_{k<l}\mathbb{E}\left[d^X_{kl}(i)\,d^X_{kl}(i')\right] \;=\; V_2\,.$$

Clearly

$$\frac{n(n-1)V_2}{2} \;=\; \frac{1}{4\tau_X^2}\sum_{i,i'=1}^{p} D^2_{\rho_i,\rho_j}(X_{(i)},X_{(i')})\,.$$

From Theorem 5 in Section 3.4.1, we know that for fixed n and growing p, $\widetilde{\mathcal{D}^2_n}(X,X)$ is asymptotically equivalent to $\frac{1}{4\tau_X^2}\sum_{i,i'=1}^{p}\widetilde{D^2_{n\,;\,\rho_i,\rho_j}}(X_{(i)},X_{(i')})$. Therefore an estimator of $\widetilde{V_2}$ is given by $4\,\widetilde{\mathcal{D}^2_n}(X,X)$. Note that the computational cost of $\widetilde{\mathcal{D}^2_n}(X,X)$ is linear in p while direct calculation of its leading term $\frac{1}{4\tau_X^2}\sum_{i,i'=1}^{p}\widetilde{D^2_{n\,;\,\rho_i,\rho_j}}(X_{(i)},X_{(i')})$ requires computation in the quadratic order of p. Similarly it can be shown that the variance of L_1^3 is V_3 and $\widetilde{V_3}$ can be estimated by $4\,\widetilde{\mathcal{D}^2_m}(Y,Y)$. Likewise some easy calculations show that the variance of L_1^1 is V_1. Define

$$\begin{aligned}\hat{\rho}_i(X_{k(i)},Y_{l(i)}) \;:=\rho_i(X_{k(i)},Y_{l(i)}) \;-\; \frac{1}{n}\sum_{a=1}^{n}\rho_i(X_{a(i)},Y_{l(i)}) \;-\; \frac{1}{m}\sum_{b=1}^{m}\rho_i(X_{k(i)},Y_{b(i)}) \\ +\;\frac{1}{nm}\sum_{a=1}^{n}\sum_{b=1}^{m}\rho_i(X_{a(i)},Y_{b(i)})\;,\end{aligned} \tag{B.14}$$

and

$$\hat{R}(X_k,Y_l) \;:=\; R(X_k,Y_l) - \frac{1}{n}\sum_{a=1}^{n}R(X_a,Y_l) - \frac{1}{m}\sum_{b=1}^{m}R(X_k,Y_b) + \frac{1}{nm}\sum_{a=1}^{n}\sum_{b=1}^{m}R(X_a,Y_b)\,. \tag{B.15}$$

It can be verified that

$$\hat{\rho}_i(X_{k(i)}, Y_{l(i)}) \;=\; d_{kl}(i) \;-\; \frac{1}{n}\sum_{a=1}^{n} d_{al}(i) \;-\; \frac{1}{m}\sum_{b=1}^{m} d_{kb}(i) \;+\; \frac{1}{nm}\sum_{a=1}^{n}\sum_{b=1}^{m} d_{ab}(i).$$

Observe that

$$\mathbb{E}\left[\hat{\rho}_i(X_{k(i)}, Y_{l(i)})\rho_{i'}(X_{k(i')}, Y_{l(i')})\right] \;=\; (1-1/n)(1-1/m)\,\mathbb{E}\left[d_{kl}(i)\, d_{kl}(i')\right]. \tag{B.16}$$

Let $\hat{\mathbf{A}}_i = (\hat{\rho}_i(X_{k(i)}, Y_{l(i)}))_{k,l},\; \mathbf{A}_i = (\rho_i(X_{k(i)}, Y_{l(i)}))_{k,l} \in \mathbb{R}^{n\times m}$. Note that

$$\begin{aligned}
&\frac{1}{(n-1)(m-1)}\,\mathbb{E}\sum_{k=1}^{n}\sum_{l=1}^{m}\hat{\rho}_i(X_{k(i)}, Y_{l(i)})\hat{\rho}_i(X_{k(i')}, Y_{l(i')})\\
&= \frac{1}{(n-1)(m-1)}\,\mathbb{E}\,\mathrm{tr}(\hat{\mathbf{A}}_i\hat{\mathbf{A}}_{i'}^{\top})\\
&= \frac{1}{(n-1)(m-1)}\,\mathbb{E}\,\mathrm{tr}(\hat{\mathbf{A}}_i\mathbf{A}_{i'}^{\top})\\
&= \frac{1}{(n-1)(m-1)}\,\mathbb{E}\sum_{k=1}^{n}\sum_{l=1}^{m}\;\rho_i(X_{k(i')}, Y_{l(i')})\,\hat{\rho}_i(X_{k(i)}, Y_{l(i)})\\
&= \mathbb{E}\left[d_{kl}(i)\, d_{kl}(i')\right],
\end{aligned} \tag{B.17}$$

which suggests that

$$\breve{V}_1 = \frac{1}{nm\tau^2}\sum_{i,i'=1}^{p}\frac{1}{(n-1)(m-1)}\sum_{k=1}^{n}\sum_{l=1}^{m}\hat{\rho}_i(X_{k(i)}, Y_{l(i)})\,\hat{\rho}_i(X_{k(i')}, Y_{l(i')})$$

is an unbiased estimator for V_1. However, the computational cost for $\breve{V}_1$ is linear in p^2 which is prohibitive for large p. We aim to find a joint metric whose computational cost is linear in p whose leading term is proportional to $\breve{V}_1$. It can be verified that $cdCov^2_{n,m}(X,Y)$ is asymptotically equivalent to

$$\frac{1}{4\tau^2}\sum_{i,i'=1}^{p}\frac{1}{(n-1)(m-1)}\sum_{k=1}^{n}\sum_{l=1}^{m}\hat{\rho}_i(X_{k(i)}, Y_{l(i)})\hat{\rho}_i(X_{k(i')}, Y_{l(i')})\;.$$

This can be seen from the observation that

$$
\begin{aligned}
4\,cdCov^2_{n,m}(X,Y) \;&=\; \frac{1}{\tau^2}\sum_{i,i'=1}^{p}\frac{1}{(n-1)(m-1)}\sum_{k=1}^{n}\sum_{l=1}^{m}\hat{\rho}_i(X_{k(i)},Y_{l(i)})\,\hat{\rho}_{i'}(X_{k(i')},Y_{l(i')}) \\
&\;+\; \frac{\tau^2}{(n-1)(m-1)}\sum_{k=1}^{n}\sum_{l=1}^{m}\hat{R}^2(X_k,Y_l) \\
&\;+\; \frac{1}{(n-1)(m-1)}\sum_{k=1}^{n}\sum_{l=1}^{m}\frac{1}{\tau}\sum_{i=1}^{p}\hat{\rho}_i(X_{k(i)},Y_{(li)})\;\tau\hat{R}(X_k,Y_l).
\end{aligned}
\tag{B.18}
$$

Using the Hölder's inequality as well as the fact that $\tau^2\,\hat{R}^2(X_k,Y_l)$ is $O_p(\tau^2 a_p^4) = o_p(1)$ under Assumption 3.3.3. Therefore, we can estimate $\widetilde{V}_1$ by $4cdCov^2_{n,m}(X,Y)$. Thus the variance of L_1 is V which can be estimated by

$$
\begin{aligned}
\hat{V} \;&:=\; \frac{1}{nm}\,4\,cdCov^2_{n,m}(X,Y) \;+\; \frac{1}{2n(n-1)}\,4\,\widetilde{\mathcal{D}^2_n}(X,X) \;+\; \frac{1}{2m(m-1)}\,4\,\widetilde{\mathcal{D}^2_m}(Y,Y) \\
&:=\; \hat{V}_1 \;+\; \hat{V}_2 \;+\; \hat{V}_3\,.
\end{aligned}
\tag{B.19}
$$

Proof of Theorem 2. Using Proposition 3.3.1, some algebraic calculations yield

$$
\begin{aligned}
&\mathcal{E}_{nm}(X,Y) - (2\tau - \tau_X - \tau_Y) \\
&= \frac{\tau}{nm}\sum_{k=1}^{n}\sum_{l=1}^{m}L(X_k,Y_l) - \frac{\tau_X}{2n(n-1)}\sum_{k\neq l}^{n}L(X_k,X_l) - \frac{\tau_Y}{2m(m-1)}\sum_{k\neq l}^{m}L(Y_k,Y_l) \;+\; R_{n,m} \\
&= \frac{1}{nm\tau}\sum_{k=1}^{n}\sum_{l=1}^{m}\sum_{i=1}^{p}\big(\rho_i(X_{k(i)},Y_{l(i)}) - \mathbb{E}\,\rho_i(X_{k(i)},Y_{l(i)})\big) \\
&\quad - \frac{1}{2n(n-1)\tau_X}\sum_{k\neq l}^{n}\sum_{i=1}^{p}\big(\rho_i(X_{k(i)},X_{l(i)}) - \mathbb{E}\,\rho_i(X_{k(i)},X_{l(i)})\big) \\
&\quad - \frac{1}{2m(m-1)\tau_Y}\sum_{k\neq l}^{m}\sum_{i=1}^{p}\big(\rho_i(Y_{k(i)},Y_{l(i)}) - \mathbb{E}\,\rho_i(Y_{k(i)},Y_{l(i)})\big) \;+\; R_{n,m},
\end{aligned}
$$

where

$$R_{n,m} \;=\; \frac{2\tau}{nm}\sum_{k=1}^{n}\sum_{l=1}^{m} R(X_k, Y_l) - \frac{\tau_X}{n(n-1)}\sum_{k\neq l}^{n} R(X_k, X_l) - \frac{\tau_Y}{m(m-1)}\sum_{k\neq l}^{m} R(Y_k, Y_l)\;. \tag{B.20}$$

By Assumption 3.3.3, $R_{n,m} = O_p(\tau a_p^2 + \tau_X b_p^2 + \tau_Y c_p^2) = o_p(1)$ as $p \to \infty$. Denote the leading term above by L. We can rewrite L as $L_1 + L_2$, where L_1 and L_2 are defined in equations (B.9) and (B.10), respectively. Some calculations yield that

$$\begin{aligned} L_2 \;&= \frac{1}{n}\sum_{k=1}^{n}\left[\frac{1}{\tau}\sum_{i=1}^{p}\mathbb{E}\left[\rho_i(X_{k(i)}, Y_{(i)})|X_{k(i)}\right] \;-\; \frac{1}{\tau_X}\sum_{i=1}^{p}\mathbb{E}\left[\rho_i(X_{k(i)}, X'_{(i)})|X_{k(i)}\right]\right] \;-\; (\tau - \tau_X) \\ &\quad + \frac{1}{m}\sum_{l=1}^{m}\left[\frac{1}{\tau}\sum_{i=1}^{p}\mathbb{E}\left[\rho_i(X_{(i)}, Y_{l(i)})|Y_{l(i)}\right] \;-\; \frac{1}{\tau_Y}\sum_{i=1}^{p}\mathbb{E}\left[\rho_i(Y_{l(i)}, Y'_{(i)})|Y_{l(i)}\right]\right] \;-\; (\tau - \tau_Y) \\ &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\,\left[\tau L(X_k, Y) - \tau_X L(X_k, X')\,|\,X_k\right] \;+\; \frac{1}{m}\sum_{l=1}^{m}\mathbb{E}\,\left[\tau L(X, Y_l) - \tau_X L(Y_l, Y')\,|\,Y_l\right]\;. \end{aligned} \tag{B.21}$$

For $(P_X, P_Y) \in \mathcal{P}$, we have $L_2 = o_p(1)$.

Under Assumption 3.3.4, the asymptotic distribution of L_1 as $p \to \infty$ is given by

$$L_1 \xrightarrow{d} N\Big(0\,,\, \frac{\sigma^2}{nm} + \frac{\sigma_X^2}{2n(n-1)} + \frac{\sigma_Y^2}{2m(m-1)}\Big).$$

Define the vector $d_{\text{vec}} := \big(\frac{1}{\tau}\sum_{i=1}^{p} d_{kl}(i)\big)_{1\leq k\leq n,\, 1\leq l\leq m}$. It can be verified that

$$4(n-1)(m-1)\, cdCov_{n,m}^2(X, Y) \;=\; d_{\text{vec}}^\top\, A\, d_{\text{vec}} \tag{B.22}$$

where $A = A_1 + A_2 + A_3 + A_4$ with $A_1 = I_n \otimes I_m$, $A_2 = -I_n \otimes \frac{1}{m}1_m 1_m^\top$, $A_3 = -\frac{1}{n}1_n 1_n^\top \otimes I_m$ and $A_4 = \frac{1}{nm}1_{nm}1_{nm}^\top$. Here $\otimes$ denotes the Kronecker product. It is not hard to see that $A^2 = A$

and rank$(A) = (n-1)(m-1)$. Therefore by Assumption 3.3.4, we have as $p \to \infty$,

$$4(n-1)(m-1)\, cdCov^2_{n,m}(X,Y) \overset{d}{\to} \sigma^2 \chi^2_{(n-1)(m-1)}.$$

By Theorem 7, we have as $p \to \infty$,

$$4\,\widetilde{\mathcal{D}}^2_n(X,X) \overset{d}{\to} \frac{\sigma_X^2}{v_n}\chi^2_{v_n}\ , \ \text{ i.e., }\ 4\,v_n\,\widetilde{\mathcal{D}}^2_n(X,X) \overset{d}{\to} \sigma_X^2\,\chi^2_{v_n}\,,$$

and similarly

$$4\,v_m\,\widetilde{\mathcal{D}}^2_m(Y,Y) \overset{d}{\to} \sigma_Y^2\,\chi^2_{v_m}\,.$$

By Assumption 3.3.4, $\chi^2_{(n-1)(m-1)}, \chi^2_{v_n}$ and $\chi^2_{v_m}$ are mutually independent. The proof can be completed by combining all the arguments above and using the continuous mapping theorem. ♢

Proof of Proposition 3.3.2. Note that as $n, m \to \infty$,

$$\mathbb{E}\left[(M-m_0)^2\right] = \frac{2(n-1)(m-1)\sigma^4 + 2v_n\sigma_X^4 + 2v_m\sigma_Y^4}{\{\,(n-1)(m-1) + v_n + v_m\,\}^2} = o(1),$$

where $m_0 = \mathbb{E}[M]$. Therefore by Chebyshev's inequality, $M - m_0 = o_p(1)$ as $n, m \to \infty$. As a consequence, we have $M \xrightarrow{p} m_0^*$ as $n, m \to \infty$. Observing that Φ is a bounded function, the rest follows from Lebesgue's Dominated Convergence Theorem. ♢

Under H_0, without any loss of generality define $U_1 = X_1, \ldots, U_n = X_n, U_{n+1} := Y_1, \ldots,$

$U_{n+m} := Y_m$. Further define

$$\phi_{i_1 i_2} := \phi(U_{i_1}, U_{i_2}) = \begin{cases} -\frac{1}{n(n-1)}\, H(U_{i_1}, U_{i_2}) & \text{if } i_1, i_2 \in \{1, \dots, n\}\,, \\ \frac{1}{nm}\, H(U_{i_1}, U_{i_2}) & \text{if } i_1 \in \{1, \dots, n\}, i_2 \in \{n+1, \dots, n+m\}\,, \\ -\frac{1}{m(m-1)}\, H(U_{i_1}, U_{i_2}) & \text{if } i_1, i_2 \in \{n+1, \dots, n+m\}\,. \end{cases} \tag{B.23}$$

It can be verified that $\mathrm{cov}(\phi_{i_1 i_2},\, \phi_{i'_1 i'_2}) = 0$ if the cardinality of the set $\{i_1, i_2\} \cap \{i'_1, i'_2\}$ is less than 2. Define

$$\breve{T}_{n,m} \;=\; \frac{\mathcal{E}_{n,m}(X,Y)}{\sqrt{V}}.$$

LEMMA **B.4.1**. *Under H_0 and Assumptions 3.3.5, B.2.1 and B.2.2, as n, m and $p \to \infty$, we have*

$$\breve{T}_{n,m} \;\xrightarrow{d}\; N(0,1)\,.$$

Proof of Lemma B.4.1. Set $N = n + m$. Define $V_{Nj} := \sum_{i=1}^{j-1} \phi_{ij}$ for $2 \leq j \leq N$, $S_{Nr} := \sum_{j=2}^{r} V_{Nj} = \sum_{j=2}^{r} \sum_{i=1}^{j-1} \phi_{ij}$ for $2 \leq r \leq N$, and $\mathcal{F}_{N,r} := \sigma(X_1, \dots, X_r)$. Then the leading term of $\mathcal{E}_{nm}(X,Y)$, viz., L_1 (see equation (B.9)) can be expressed as

$$L_1 \;=\; S_{NN} = \sum_{j=2}^{N} V_{Nj} \;=\; \sum_{j=2}^{N} \sum_{i=1}^{j-1} \phi_{ij} \;=\; \sum_{1 \leq i_1 < i_2 \leq n} \phi_{i_1 i_2} + \sum_{i_1=1}^{n} \sum_{i_2=n+1}^{N} \phi_{i_1 i_2} + \sum_{n+1 \leq i_1 < i_2 \leq N} \phi_{i_1 i_2}\,.$$

By Corollary 3.1 of Hall and Heyde (1980), it suffices to show the following :

1. For each N, $\{S_{Nr}, \mathcal{F}_{N,r}\}_{r=1}^{N}$ is a sequence of zero mean and square integrable martingales,
2. $\frac{1}{V} \sum_{j=2}^{N} \mathbb{E}\left[V_{Nj}^2 \,|\, \mathcal{F}_{N,j-1}\right] \;\xrightarrow{P}\; 1\,,$
3. $\frac{1}{V} \sum_{j=2}^{N} \mathbb{E}\left[V_{Nj}^2\, \mathbb{1}(|V_{Nj}| > \epsilon\sqrt{V}) \,|\, \mathcal{F}_{N,j-1}\right] \;\xrightarrow{P}\; 0\,, \quad \forall\, \epsilon > 0.$

To show (1), it is easy to see that S_{Nr} is square integrable, $\mathbb{E}(S_{Nr}) = \sum_{j=2}^{r} \sum_{i=1}^{j-1} \mathbb{E}(\phi_{ij}) = 0$, and, $\mathcal{F}_{N,1} \subseteq \mathcal{F}_{N,2} \subseteq \dots \subseteq \mathcal{F}_{N,N}$. We only need to show $\mathbb{E}(S_{Nq} \,|\, \mathcal{F}_{N,r}) = S_{Nr}$ for $q > r$. Now

$\mathbb{E}(S_{Nq}\,|\,\mathcal{F}_{N,r}) \;=\; \sum_{j=2}^{q}\sum_{i=1}^{j-1}\mathbb{E}(\phi_{ij}\,|\,\mathcal{F}_{N,r})$. If $j \;\leq\; r \;<\; q$ and $i \;<\; j$, then $\mathbb{E}(\phi_{ij}\,|\,\mathcal{F}_{N,r}) \;=\; \phi_{ij}$. If $r < j \leq q$, then :

(i) if $r < i < j \leq q$, then $\mathbb{E}(\phi_{ij}\,|\,\mathcal{F}_{N,r}) = \mathbb{E}(\phi_{ij}) = 0$,

(ii) if $i \leq r < j \leq q$, then $\mathbb{E}(\phi_{ij}\,|\,\mathcal{F}_{N,r}) = 0$ (due to $\mathcal{U}$-centering).

Therefore $\mathbb{E}(S_{Nq}\,|\,\mathcal{F}_{N,r}) = S_{Nr}$ for $q > r$. This completes the proof of (1).

To show (2), define $L_j(i,k) := \mathbb{E}\left[\phi_{ij}\,\phi_{kj}\,|\,\mathcal{F}_{N,j-1}\right]$ for $i,k < j \leq N,$ and

$$\eta_N := \sum_{j=2}^{N}\mathbb{E}\,\left[V_{Nj}^2\,|\,\mathcal{F}_{N,j-1}\right] = \sum_{j=2}^{N}\sum_{i,k=1}^{j-1}\mathbb{E}[\phi_{ij}\,\phi_{kj}\,|\,\mathcal{F}_{N,j-1}] = \sum_{j=2}^{N}\sum_{i,k=1}^{j-1}L_j(i,k)\,.$$

Note that $\mathbb{E}\left[L_j(i,k)\right] = 0$ for $i \neq k$. Clearly

$$\mathbb{E}[\eta_N] \;=\; \sum_{j=2}^{N}\mathbb{E}[V_{Nj}^2] \;=\; \sum_{j=2}^{N}\sum_{i,k=1}^{j-1}\mathbb{E}[\phi_{ij}\,\phi_{kj}] \;=\; \sum_{j=2}^{N}\sum_{i=1}^{j-1}\mathbb{E}[\phi_{ij}^2] = V\,. \tag{B.24}$$

By virtue of Chebyshev's inequality, it will suffice to show var$(\frac{\eta_N}{V}) = o(1)$. Note that

$$\begin{aligned}&\mathbb{E}\left[L_j(i,k)\,L_{j'}(i',k')\right]\\ &= \begin{cases}\mathbb{E}\,\left[\phi^2(U_i,U_j)\phi^2(U_i,U'_{j'})\right] & i = k = i' = k'\,,\\ \mathbb{E}\,\left[\phi(U_i,U_j)\phi(U_k,U_j)\phi(U_i,U'_{j'})\phi(U_k,U'_{j'})\right] & i = i' \neq k = k' \text{ or } i = k' \neq k = i'\,,\\ \mathbb{E}\,\left[\phi^2(U_i,U_j)\right]\mathbb{E}\,\left[\phi^2(U_{i'},U_{j'})\right] & i = k \neq i' = k'\,.\end{cases}\end{aligned} \tag{B.25}$$

In view of equation (B.23), it can be verified that the above expression for $\mathbb{E}\,L_j(i,k)\,L_{j'}(i',k')$

holds true for $j = j'$ as well. Therefore

$$\begin{aligned}
\text{var}\left(\eta_N^2\right) \;=\; & \sum_{j,j'=2}^{N} \sum_{i,k=1}^{j-1} \sum_{i',k'=1}^{j'-1} \text{cov}\left(L_j(i,k)\,, L_{j'}(i',k')\right) \\
=\; & \sum_{j=j'} \Bigg\{ \sum_{i=1}^{j-1} \text{cov}\, \left(\phi^2(U_i,U_j), \phi^2(U_i,U_j')\right) \\
& +\; 2 \sum_{i\neq k}^{j-1} \mathbb{E}\, \left[\phi(U_i,U_j)\phi(U_k,U_j)\phi(U_i,U_j')\phi(U_k,U_j')\right] \Bigg\} \\
& +\; 2 \sum_{2\leq j<j'\leq N} \Bigg\{ \sum_{i=1}^{j-1} \text{cov}\, \left(\phi^2(U_i,U_j), \phi^2(U_i,U_{j'}')\right) \\
& +\; 2 \sum_{i\neq k}^{j-1} \mathbb{E}\, \left[\phi(U_i,U_j)\phi(U_k,U_j)\phi(U_i,U_{j'}')\phi(U_k,U_{j'}')\right] \Bigg\} .
\end{aligned}$$

Under Assumption 3.3.5 and H_0, it can be verified that

$$\text{var}(\eta_N) = O\Big(\frac{1}{N^5}\,\mathbb{E}\,\left[H^2(X,X'')H^2(X',X'')\right] + \frac{1}{N^4}\mathbb{E}\left[H(X,X'')H(X',X'')H(X,X''')H(X',X''')\right]\Big), \tag{B.26}$$

and

$$V^2 \;\asymp\; \frac{1}{N^4}\,\left(\mathbb{E}\,\left[H^2(X,X')\right]\right)^2 . \tag{B.27}$$

Therefore under Assumption B.2.1 and H_0, we have

$$\text{var}\left(\frac{\eta_N}{V}\right) = o(1),$$

which completes the proof of (2). To show (3), note that it suffices to show

$$\frac{1}{V^2}\sum_{j=2}^{N}\mathbb{E}\left[\,V_{Nj}^4\,|\,\mathcal{F}_{N,j-1}\,\right] \;\xrightarrow{P}\; 0 \,.$$

Observe that

$$\begin{aligned}\sum_{j=2}^{N} \mathbb{E}\left[V_{Nj}^4\right] &= \sum_{j=2}^{N} \mathbb{E}\left(\sum_{i=1}^{j-1} \phi_{ij}\right)^4 \\ &= \sum_{j=2}^{N}\sum_{i=1}^{j-1} \mathbb{E}[\phi^4(U_i, U_j)] \;+\; 3\sum_{j=2}^{N}\sum_{i_1 \neq i_2}^{j-1} \mathbb{E}[\phi^2(U_{i_1}, U_j)\, \phi^2(U_{j_2}, U_j)]\,.\end{aligned}$$

Under Assumption 3.3.5, we have

$$\sum_{j=2}^{N} \mathbb{E}\left[V_{Nj}^4\right] \;=\; O\Big(\frac{1}{N^6}\, \mathbb{E}\,\left[H^4(X, X')\right] \;+\; \frac{1}{N^5}\, \mathbb{E}\,\left[H^2(X, X'')H^2(X', X'')\right]\Big)\,.$$

This along with the observation from equation (B.26) and Assumption B.2.1 complete the proof of (3).

Finally to see that $\frac{R_{n,m}}{\sqrt{V}} = o_p(1)$, note that from equation (B.20) we can derive using power mean inequality that $\mathbb{E}\, R_{n,m}^2 \;\leq\; C\, \tau^2\, \mathbb{E}\,\left[R^2(X, X')\right]$ for some positive constant C. Using this, equation (B.27), Chebyshev's inequality and Hölder's inequality, we have for any $\epsilon > 0$

$$P\left(\left|\frac{R_{n,m}}{\sqrt{V}}\right| > \epsilon\right) \;\leq\; \frac{\mathbb{E}\, R_{n,m}^2}{\epsilon^2\, V} \;\leq\; C'\, \frac{N^2\, \tau^2\, \mathbb{E}\,\left[R^2(X, X')\right]}{\epsilon^2\, \mathbb{E}\,\left[H^2(X, X')\right]} \;\leq\; \frac{C'}{\epsilon^2}\, \left(\frac{N^4\, \tau^4\, \mathbb{E}\,\left[R^4(X, X')\right]}{\left(\mathbb{E}\,\left[H^2(X, X')\right]\right)^2}\right)^{1/2}, \tag{B.28}$$

for some positive constant C'. From this and Assumptions 3.3.5 and B.2.2, we get $\frac{R_{n,m}}{\sqrt{V}} = o_p(1)$, as $N \asymp n$. This completes the proof of the lemma. ♢

LEMMA **B.4.2**. *Under H_0 and Assumptions 3.3.5 and B.2.2, as n, m and $p \to \infty$, we have*

$$\frac{\left|\mathbb{E}\,[\hat{V}_i] - V_i\right|}{V_i} = o(1) \;\;,\;\; 1 \leq i \leq 3\,,$$

where V_i and $\hat{V}_i$, $1 \leq i \leq 3$ are defined in equations (B.11) and (B.19), respectively.

Proof of Lemma B.4.2. We first deal with $\hat{V}_2$. Note that

$$\widetilde{\mathcal{D}^2_n}(X, X) \;=\; \frac{1}{n(n-3)} \sum_{k \neq l} \left(\widetilde{D}^X_{kl} \right)^2 ,$$

where

$$\begin{aligned} \widetilde{D}^X_{kl} \; &= K(X_k, X_l) \;-\; \frac{1}{n-2} \sum_{b=1}^{n} K(X_k, X_b) \;-\; \frac{1}{n-2} \sum_{a=1}^{n} K(X_a, X_l) \\ &\quad +\; \frac{1}{(n-1)(n-2)} \sum_{a,b=1}^{n} K(X_a, X_b) \\ &=\; \frac{1}{2\tau} \sum_{i=1}^{p} \widetilde{\rho}_i(X_{k(i)}, X_{l(i)}) \;+\; \tau \widetilde{R}(X_k, X_l) \,, \end{aligned} \tag{B.29}$$

using Proposition 3.3.1. As a consequence, we can write

$$\begin{aligned} \widetilde{\mathcal{D}^2_n}(X, X) \;&=\; \frac{1}{4\tau^2} \sum_{i,i'=1}^{p} \widetilde{D^2_{n\,;\,\rho_i,\rho_{i'}}}(X_{(i)}, X_{(i')}) \;+\; \frac{\tau^2}{n(n-3)} \sum_{k \neq l} \widetilde{R}^2(X_k, X_l) \\ &\quad +\; \frac{1}{n(n-3)} \sum_{k \neq l} \frac{1}{\tau} \sum_{i=1}^{p} \widetilde{\rho}_i(X_{k(i)}, X_{(li)}) \; \tau \widetilde{R}(X_k, X_l) \,. \end{aligned} \tag{B.30}$$

Note that following Step 3 in Section 1.6 in the supplementary material of Zhang et al. (2018), we can write

$$\begin{aligned} \widetilde{R}(X_k, X_l) \;&=\; \frac{n-3}{n-1} \bar{R}(X_k, X_l) \;-\; \frac{n-3}{(n-1)(n-2)} \sum_{b \notin \{k,l\}} \bar{R}(X_k, X_b) \\ &\quad -\; \frac{n-3}{(n-1)(n-2)} \sum_{a \notin \{k,l\}} \bar{R}(X_a, X_l) \;+\; \frac{1}{(n-1)(n-2)} \sum_{a,b \notin \{k,l\}} \bar{R}(X_a, X_b) \,, \end{aligned}$$

where $\bar{R}(X, X') = R(X, X') - E[R(X, X')|X] - E[R(X, X')|X'] + E[R(X, X')]$. Using the power mean inequality, it can be verified that $\mathbb{E}\,[\widetilde{R}^2(X_k, X_l)] \leq C\,\mathbb{E}\,[\bar{R}^2(X_k, X_l)]$ for some positive constant C. Using this and the Hölder's inequality, the expectation of the third term in the summation in equation (B.30) can be bounded as follows

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{n(n-3)}\sum_{k\neq l}\frac{1}{\tau}\sum_{i=1}^{p}\widetilde{\rho}_i(X_{k(i)},X_{l(i)})\ \tau\widetilde{R}(X_k,X_l)\right]\right| \\
&\leq \frac{1}{n(n-3)}\sum_{k\neq l}\left(\mathbb{E}\left[\left(\frac{1}{\tau}\sum_{i=1}^{p}\widetilde{\rho}_i(X_{k(i)},X_{l(i)})\right)^2\right]\tau^2\,\mathbb{E}\left[\bar{R}^2(X_k,X_l)\right]\right)^{1/2} \\
&\leq C'\left(\left(\frac{1}{\tau^2}\sum_{i,i'=1}^{p}D^2_{\rho_i,\rho_{i'}}(X_{(i)},X_{(i')})\right)\tau^2\,\mathbb{E}\left[\bar{R}^2(X,X')\right]\right)^{1/2}
\end{aligned}$$

for some positive constant C'. Combining all the above, we get

$$\begin{aligned}
|\mathbb{E}\,(\hat{V}_2)-V_2| \leq &\frac{C_1}{n(n-1)}\,\tau^2\,\mathbb{E}\,\bar{R}^2(X,X') \\
&+\frac{C_2}{n(n-1)}\left(\left(\frac{1}{\tau^2}\sum_{i,i'=1}^{p}D^2_{\rho_i,\rho_{i'}}(X_{(i)},X_{(i')})\right)\tau^2\,\mathbb{E}\left[\bar{R}^2(X,X')\right]\right)^{1/2},
\end{aligned}$$

for some positive constants C_1 and C_2. As $V_2 = \frac{1}{2n(n-1)}E[H^2(X,X')]$,

$$\frac{\left|\mathbb{E}[\hat{V}_2]-V_2\right|}{V_2} = o(1) \quad \text{is satisfied if} \quad \frac{\tau^2\,\mathbb{E}\left[\bar{R}^2(X,X')\right]}{\mathbb{E}[H^2(X,X')]} \;=\; o(1)\,.$$

Using power mean inequality and Jensen's inequality, it is not hard to verify that $\mathbb{E}\left[\bar{R}^4(X,X')\right] = O\left(\mathbb{E}\left[R^4(X,X')\right]\right)$. Using this and Hölder's inequality, we have

$$\frac{\tau^2\,\mathbb{E}\left[\bar{R}^2(X,X')\right]}{\mathbb{E}[H^2(X,X')]} \;=\; O\left(\left(\frac{\tau^4\,\mathbb{E}\left[R^4(X,X')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2}\right)^{1/2}\right).$$

Clearly Assumption B.2.2 implies $\frac{\tau^4\,\mathbb{E}\left[R^4(X,X')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} = o(1)$, which in turn implies

$$\frac{\tau^2\,\mathbb{E}\left[\bar{R}^2(X,X')\right]}{\mathbb{E}[H^2(X,X')]} \;=\; o(1)\,.$$

Similar expressions can be derived for $\hat{V}_3$ as well. For the term involving $\hat{V}_1$, in the similar fashion, we can write

$$\begin{aligned}\mathbb{E}\left[4\,cdCov^2_{n,m}(X,Y)\right] \;=\; & \frac{1}{\tau^2}\sum_{i,i'=1}^{p}\frac{1}{(n-1)(m-1)}\sum_{k=1}^{n}\sum_{l=1}^{m}\mathbb{E}\left[\hat{\rho}_i(X_{k(i)},Y_{l(i)})\,\hat{\rho}_{i'}(X_{k(i')},Y_{l(i')})\right]\\ & +\;\tau^2\frac{1}{(n-1)(m-1)}\sum_{k=1}^{n}\sum_{l=1}^{m}\mathbb{E}\left[\hat{R}^2(X_k,Y_l)\right]\\ & +\;\frac{1}{(n-1)(m-1)}\sum_{k=1}^{n}\sum_{l=1}^{m}\frac{1}{\tau}\sum_{i=1}^{p}\mathbb{E}\left[\hat{\rho}_i(X_{k(i)},Y_{(li)})\;\tau\hat{R}(X_k,Y_l)\right]\,,\end{aligned} \tag{B.31}$$

where the expression for $\hat{R}(X_k,Y_l)$ is given in equation (B.15). Following equation (B.17) we can write

$$\frac{1}{\tau^2}\sum_{i,i'=1}^{p}\frac{1}{(n-1)(m-1)}\sum_{k=1}^{n}\sum_{l=1}^{m}\mathbb{E}\left[\hat{\rho}_i(X_{k(i)},Y_{l(i)})\,\hat{\rho}_{i'}(X_{k(i')},Y_{l(i')})\right] \;=\; \mathbb{E}\left[H^2(X,Y)\right]\,.$$

Therefore in view of equations (B.11), (B.16) and (B.19), using the power mean inequality we can write

$$|\mathbb{E}\,(\hat{V}_1)-V_1| \le \frac{C_1'}{nm}\,\tau^2\,\mathbb{E}\,\bar{R}^2(X,Y) \;+\; \frac{C_2'}{nm}\left(\left(\frac{1}{\tau^2}\sum_{i,i'=1}^{p}\mathbb{E}\left[d_{kl}(i)d_{kl}(i')\right]\right)\tau^2\,\mathbb{E}\left[\bar{R}^2(X,Y)\right]\right)^{1/2},$$

for some positive constants C_1' and C_2'. Then under H_0 and Assumptions 3.3.5 and B.2.2, we have

$$\frac{\left|\mathbb{E}\,(\hat{V}_1)-V_1\right|}{V_1} = o(1)\,.$$

$\diamondsuit$

LEMMA **B.4.3**. *Under H_0 and Assumptions 3.3.5, B.2.1 and B.2.2, as n, m and $p \to \infty$, we have*

$$\frac{var(\hat{V}_i)}{V_i^2} = o(1), \;\; 1 \le i \le 3\,.$$

Proof of Lemma B.4.3. Again we deal with $\hat{V}_2$ first. To simplify the notations, denote A_{ij} =

$K(X_i, X_j)$ and $\widetilde{A}_{ij} = \widetilde{D}^X_{ij}$ for $1 \le i \ne j \le n$. Observe that

$$\begin{aligned} \operatorname{var}\left(\widetilde{\mathcal{D}^2_n}(X,X)\right) &= \operatorname{var}\left(\frac{1}{n(n-3)}\sum_{i\neq j}\widetilde{A}^2_{ij}\right) \\ &\asymp \frac{1}{n^4}\left[\sum_{i<j}\operatorname{var}(\widetilde{A}^2_{ij}) \;+\; \sum_{i<j<j'}\operatorname{cov}(\widetilde{A}^2_{ij},\widetilde{A}^2_{jj'}) \;+\; \sum_{\substack{i<j,i'<j' \\ \{i,j\}\cap\{i',j'\}=\phi}}\operatorname{cov}(\widetilde{A}^2_{ij},\widetilde{A}^2_{i'j'})\right]. \end{aligned} \tag{B.32}$$

As in the proof of Lemma B.4.2, we can write

$$\begin{aligned} \widetilde{A}_{ij} &= \frac{n-3}{n-1}\bar{A}_{ij} - \frac{n-3}{(n-1)(n-2)}\sum_{l\notin\{i,j\}}\bar{A}_{il} - \frac{n-3}{(n-1)(n-2)}\sum_{k\notin\{i,j\}}\bar{A}_{kj} \\ &\quad + \frac{1}{(n-1)(n-2)}\sum_{k,l\notin\{i,j\}}\bar{A}_{kl}, \end{aligned} \tag{B.33}$$

where the four summands are uncorrelated with each other. Using the power mean inequality, it can be shown that

$$\mathbb{E}\,(\widetilde{A}^4_{ij}) \;\le\; C\,\mathbb{E}\,(\bar{A}^4_{ij}) \;=\; C\,\mathbb{E}\,\left[\bar{K}^4(X,X')\right],$$

for some positive constant C, where $\bar{K}(X,X') = K(X,X') - E[K(X,X')|X] - E[K(X,X')|X'] + E[K(X,X')]$ (similarly define $\bar{L}(X,X')$). Therefore the first summand in equation (B.32) scaled by $\widetilde{V_2}^2$ is $o(1)$ as $n, p \to \infty$, provided

$$\frac{1}{n^2}\,\frac{\mathbb{E}\,\left[\bar{K}^4(X,X')\right]}{\widetilde{V_2}^2} \;=\; o(1),$$

where $\widetilde{V_2}$ is defined in equations (B.12) and (B.13). Note that

$$\bar{K}(X,X') \;=\; \frac{\tau_X}{2}\,\bar{L}(X,X') \;+\; \tau_X\,\bar{R}(X,X').$$

Using the power mean inequality we can write

$$\frac{1}{n^2}\,\frac{\mathbb{E}\,\left[\bar{K}^4(X,X')\right]}{\left(\mathbb{E}\,\left[H^2(X,X')\right]\right)^2} \,\leq\, C_0\,\frac{1}{n^2}\,\frac{\tau_X^4\,\mathbb{E}\,\left[\bar{L}^4(X,X')\right]}{\left(\mathbb{E}\,\left[H^2(X,X')\right]\right)^2} \,+\, C_0'\,\frac{1}{n^2}\,\frac{\tau_X^4\,\mathbb{E}\,\left[\bar{R}^4(X,X')\right]}{\left(\mathbb{E}\,\left[H^2(X,X')\right]\right)^2}$$

for some positive constants C_0 and C_0'. It is easy to see that

$$\bar{L}(X_k,X_l) \,=\, \frac{1}{\tau_X^2}\,\bar{K}^2(X_k,X_l) \,=\, \frac{1}{\tau_X^2}\,\sum_{i=1}^{p} d_{kl}^X(i) \,=\, \frac{1}{\tau_X}\,H(X_k,X_l)\,. \tag{B.34}$$

From equation (B.34) it is easy to see that the condition

$$\frac{1}{n^2}\,\frac{\tau_X^4\,\mathbb{E}\,\left[\bar{L}^4(X,X')\right]}{\left(\mathbb{E}\,\left[H^2(X,X')\right]\right)^2} = o(1) \quad \text{is equivalent to} \quad \frac{1}{n^2}\,\frac{\mathbb{E}\,\left[H^4(X,X')\right]}{\left(\mathbb{E}\,\left[H^2(X,X')\right]\right)^2} = o(1).$$

For the third summand in equation (B.32), observe that

$$\begin{aligned}\widetilde{A}_{ij}^2 =& O(1)\bar{A}_{ij}^2 + O\left(\frac{1}{n^2}\right)\sum_{l,l'\notin\{i,j\}}\bar{A}_{il}\bar{A}_{il'} + O\left(\frac{1}{n^2}\right)\sum_{k,k'\notin\{i,j\}}\bar{A}_{kj}\bar{A}_{k'j} + O\left(\frac{1}{n^4}\right)\sum_{k,k',l,l'\notin\{i,j\}}\bar{A}_{kl}\bar{A}_{k'l'}\\ &+ O\left(\frac{1}{n}\right)\bar{A}_{ij}\sum_{l\notin\{i,j\}}\bar{A}_{il} \,+\, O\left(\frac{1}{n}\right)\bar{A}_{ij}\sum_{k\notin\{i,j\}}\bar{A}_{kj} \,+\, O\left(\frac{1}{n^2}\right)\bar{A}_{ij}\sum_{k,l\notin\{i,j\}}\bar{A}_{kl}\\ &+ O\left(\frac{1}{n^2}\right)\sum_{k,l\notin\{i,j\}}\bar{A}_{il}\bar{A}_{kj} + O\left(\frac{1}{n^3}\right)\sum_{k,l,l'\notin\{i,j\}}\bar{A}_{il}\bar{A}_{kl'} \,+\, O\left(\frac{1}{n^3}\right)\sum_{k,k',l\notin\{i,j\}}\bar{A}_{kl}\bar{A}_{k'j}\,.\end{aligned} \tag{B.35}$$

Likewise $\widetilde{A}_{i'j'}^2$ admits a similar expression as in equation (B.35). We claim that when $\{i,j\}\cap\{i',j'\}=\phi$, the leading term of $\mathrm{cov}(\widetilde{A}_{ij}^2,\widetilde{A}_{i'j'}^2)$ is $O\left(\frac{1}{n^2}\,\mathbb{E}\,(\bar{A}_{ij}^4)\right)$. To see this first note that $\bar{A}_{ij}$ is independent of $\bar{A}_{i'j'}$ when $\{i,j\}\cap\{i',j'\}=\phi$. Using the double-centering properties, it can be verified that

$$\mathrm{cov}\left(\bar{A}_{i'j'}^2,\bar{A}_{ij}\sum_{l\notin\{i,j\}}\bar{A}_{il}\right) \,=\, \mathrm{cov}\left(\bar{A}_{i'j'}^2,\bar{A}_{ij}\sum_{k\notin\{i,j\}}\bar{A}_{kj}\right) \,=\, \mathrm{cov}\left(\bar{A}_{i'j'}^2,\bar{A}_{ij}\sum_{k,l\notin\{i,j\}}\bar{A}_{kl}\right) = 0.$$

To compute the quantity $\operatorname{cov}\left(\bar{A}^2_{i'j'}\ ,\ O\left(\frac{1}{n^2}\right)\sum\limits_{l,l'\notin\{i,j\}}\bar{A}_{il}\bar{A}_{il'}\right)$, consider the following cases:

Case 1 . When $l = l' = i'$ or $l = l' = j'$ or $l = i', l' = j'$, cov $\left(\bar{A}^2_{i'j'}, \bar{A}_{il}\bar{A}_{il'}\right)$ boils down to cov$(\bar{A}^2_{i'j'}, \bar{A}^2_{ii'})$ or cov$(\bar{A}^2_{i'j'}, \bar{A}^2_{ij'})$ or cov$(\bar{A}^2_{i'j'}, \bar{A}_{ii'}\bar{A}_{ij'})$.

Case 2 . When $l = i, l' \notin \{i, j, i', j'\}$ or $l = j', l' \notin \{i, j, i', j'\}$, cov $\left(\bar{A}^2_{i'j'}, \bar{A}_{il}\bar{A}_{il'}\right)$ boils down to cov$(\bar{A}^2_{i'j'}, \bar{A}_{ii'}\bar{A}_{il'})$ or cov$(\bar{A}^2_{i'j'}, \bar{A}_{ij'}\bar{A}_{il'})$, which can be easily verified to be zero.

Case 3 . When $\{l, l'\} \cap \{i', j'\} = \phi$, cov $\left(\bar{A}^2_{i'j'}, \bar{A}_{il}\bar{A}_{il'}\right)$ is again zero.

Similar arguments can be made about

$$\operatorname{cov}\left(\bar{A}^2_{i'j'}\ ,\ O\left(\frac{1}{n^2}\right)\sum_{k,k'\notin\{i,j\}}\bar{A}_{kj}\bar{A}_{k'j}\right) \quad \text{and} \quad \operatorname{cov}\left(\bar{A}^2_{i'j'}\ ,\ O\left(\frac{1}{n^2}\right)\sum_{k,l\notin\{i,j\}}\bar{A}_{il}\bar{A}_{kj}\right).$$

With this and using Hölder's inequality, it can be verified that when $\{i, j\}\cap\{i', j'\} = \phi$, the leading term of cov$(\widetilde{A}^2_{ij}, \widetilde{A}^2_{i'j'})$ is $O\left(\frac{1}{n^2}\,\mathbb{E}\left(\bar{A}^4_{ij}\right)\right)$. Therefore the third summand in equation (B.32) scaled by $\widetilde{V_2}^2$ can be argued to be $o(1)$ in similar lines of the argument for the first summand in equation (B.32).

For the second summand in equation (B.32), in the similar line we can argue that the leading term of cov$(\widetilde{A}^2_{ij}, \widetilde{A}^2_{jj'})$ is

$$O\left(\frac{1}{n}\right)\mathbb{E}\left[\bar{A}^4_{ij}\right] \;+\; O(1)\,\mathbb{E}\left[\bar{A}^2_{ij}\bar{A}^2_{jj'}\right]\ .$$

Therefore the leading term of $\frac{1}{n^4}\sum\limits_{i<j<j'}$ cov$(\widetilde{A}^2_{ij}, \widetilde{A}^2_{jj'})$ is

$$O\left(\frac{1}{n^2}\right)\mathbb{E}\left[\bar{A}^4_{ij}\right] \;+\; O\left(\frac{1}{n}\right)\,\mathbb{E}\left[\bar{A}^2_{ij}\bar{A}^2_{jj'}\right]\ .$$

For the second term above, using the power mean inequality we can write

$$\begin{aligned}
\frac{1}{n}\frac{\mathbb{E}\left[\bar{A}_{ij}^2\,\bar{A}_{jj'}^2\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} \;&\leq\; C_3\,\frac{1}{n}\,\frac{\tau^4\,\mathbb{E}\left[\bar{L}^2(X,X')\,\bar{L}^2(X',X'')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} \;+\; C_3'\frac{1}{n}\,\frac{\tau^4\,\mathbb{E}\left[\bar{L}^2(X,X')\,\bar{R}^2(X',X'')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} \\
&\quad+\; C_3''\,\frac{1}{n}\,\frac{\tau^4\,\mathbb{E}\left[\bar{R}^2(X,X')\,\bar{R}^2(X',X'')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} \\
&=\; C_3\,\frac{1}{n}\,\frac{\mathbb{E}\left[H^2(X,X')\,H^2(X',X'')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} \;+\; C_3'\frac{1}{n}\,\frac{\tau^2\,\mathbb{E}\left[H^2(X,X')\,\bar{R}^2(X',X'')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2} \\
&\quad+\; C_3''\,\frac{1}{n}\,\frac{\tau^4\,\mathbb{E}\left[\bar{R}^2(X,X')\,\bar{R}^2(X',X'')\right]}{\left(\mathbb{E}\left[H^2(X,X')\right]\right)^2}
\end{aligned}$$

for some positive constants C_3, C_3' and C_3''. Using Hölder's inequality it can be seen that the second summand in equation (B.32) scaled by $\widetilde{V_2}^2$ is $o(1)$ as $n, p \to \infty$ under Assumptions B.2.1 and B.2.2. This completes the proof that

$$\frac{\text{var}(\hat{V}_2)}{V_2^2} = o(1)\,.$$

A similar line of argument and the simple observation that

$$\begin{aligned}
\hat{K}(X_k, Y_l) \;&=\; K(X_k,Y_l) - \frac{1}{n}\sum_{a=1}^{n} K(X_a, Y_l) - \frac{1}{m}\sum_{b=1}^{m} K(X_k, Y_b) + \frac{1}{nm}\sum_{a=1}^{n}\sum_{b=1}^{m} K(X_a, Y_b) \\
&=\; \bar{K}(X_k,Y_l) - \frac{1}{n}\sum_{a=1}^{n} \bar{K}(X_a, Y_l) - \frac{1}{m}\sum_{b=1}^{m} \bar{K}(X_k, Y_b) + \frac{1}{nm}\sum_{a=1}^{n}\sum_{b=1}^{m} \bar{K}(X_a, Y_b)
\end{aligned}$$

will show that under Assumptions 3.3.5, B.2.1 and B.2.2,

$$\frac{\text{var}(\hat{V}_1)}{V_1^2} = o(1) \qquad \text{and} \qquad \frac{\text{var}(\hat{V}_3)}{V_3^2} = o(1)\,.$$

◇

LEMMA **B.4.4**. *Under H_0 and Assumptions 3.3.5, B.2.1 and B.2.2, as n, m and $p \to \infty$, we have* $\hat{V}/V \xrightarrow{P} 1\,.$

Proof. It is enough to show that

$$\mathbb{E}\left[\left(\frac{\hat{V}}{V}-1\right)^2\right]=o(1)\ ,\ \text{ i.e., }\ \frac{\text{var}(\hat{V})+\left(\mathbb{E}\,[\hat{V}]-V\right)^2}{V^2}=o(1)\,.$$

It suffices to show the following

$$\frac{\text{var}(\hat{V}_i)}{V_i^2}=o(1)\quad \text{and}\quad \frac{\left(\mathbb{E}\,[\hat{V}_i]-V_i\right)^2}{V_i^2}=o(1),\quad 1\le i\le 3.$$

The proof can be completed using Lemmas B.4.2 and B.4.3. ♢

Proof of Theorem 15. The proof essentially follows from Lemma B.4.1 and B.4.4.

♢

Proof of Proposition B.1.2. The proof of the first part follows similar lines of the proof of Proposition 1 in Székely et al. (2014), replacing the Euclidean distance between X and X', viz. $\|X-X'\|_{\tilde{p}}$, by $K(X,X')$. The second part of the proposition has a proof similar to Lemma 2.1 in Yao et al. (2018) and Section 1.1 in the Supplement of Yao et al. (2018). ♢

Proof of Theorem 14. The first two parts of the theorem immediately follow from Proposition 2.6 and Theorem 2.7 in Lyons (2013), respectively and the parallel U-statistics theory (see for example Serfling (1980)). The third part follows from the first part and the fact that $\mathcal{D}$ is non-zero for two dependent random vectors. ♢

Proof of Theorem 3. Following the definition of $\mathcal{D}(X,Y)$ and applying Proposition 3.3.1, we can write

$$
\begin{aligned}
\frac{1}{\tau_{XY}}\,\mathcal{D}^2(X,Y) \;&=\; \mathbb{E}\,\frac{K(X,X')}{\tau_X}\,\frac{K(Y,Y')}{\tau_Y} \;+\; \mathbb{E}\,\frac{K(X,X')}{\tau_X}\,\mathbb{E}\,\frac{K(Y,Y')}{\tau_Y} \\
&\quad -\; 2\,\mathbb{E}\,\frac{K(X,X')}{\tau_X}\,\frac{K(Y,Y'')}{\tau_Y} \\
&=\; \mathbb{E}\,\left(1+\frac{1}{2}L(X,X')+R(X,X')\right)\,\left(1+\frac{1}{2}L(Y,Y')+R(Y,Y')\right) \\
&\quad +\; \mathbb{E}\,\left(1+\frac{1}{2}L(X,X')+R(X,X')\right)\,\mathbb{E}\,\left(1+\frac{1}{2}L(Y,Y')+R(Y,Y')\right) \\
&\quad -\; 2\,\mathbb{E}\,\left(1+\frac{1}{2}L(X,X')+R(X,X')\right)\,\left(1+\frac{1}{2}L(Y,Y'')+R(Y,Y'')\right) \\
&=\; L \;+\; R,
\end{aligned}
$$

where

$$
L \;=\; \frac{1}{4}\,\left[\,\mathbb{E}\,L(X,X')L(Y,Y') \;+\; \mathbb{E}\,L(X,X')\,\mathbb{E}\,L(Y,Y') \;-\; 2\,\mathbb{E}\,L(X,X')L(Y,Y'')\,\right],
$$

and

$$
\begin{aligned}
R \;=&\;\mathbb{E}\,\left[\frac{1}{2}L(X,X')R(Y,Y') \;+\; \frac{1}{2}R(X,X')L(Y,Y') \;+\; R(X,X')R(Y,Y')\right] \\
&-\; 2\,\mathbb{E}\,\left[\frac{1}{2}L(X,X')R(Y,Y'') \;+\; \frac{1}{2}R(X,X')L(Y,Y'') \;+\; R(X,X')R(Y,Y'')\right] \\
&+\mathbb{E}\,R(X,X')\,\mathbb{E}\,R(Y,Y').
\end{aligned}
$$

Some simple calculations yield

$$
\begin{aligned}
L \;&=\; \frac{1}{4\tau_{XY}^2}\,\big\{\,\mathbb{E}\,[K^2(X,X')K^2(Y,Y')] \;+\; \mathbb{E}\,[K^2(X,X')]\,\mathbb{E}\,[K^2(Y,Y')] \\
&\quad -\; 2\,\mathbb{E}\,[K^2(X,X')K^2(Y,Y'')] \\
&=\; \frac{1}{4\tau_{XY}^2}\,\sum_{i=1}^{p}\sum_{j=1}^{q}\Big\{\,\mathbb{E}\,[\rho_i(X_{(i)},X'_{(i)})\,\rho_j(Y_{(j)},Y'_{(j)})] \;+\; \mathbb{E}\,[\rho_i(X_{(i)},X'_{(i)})]\,\mathbb{E}\,[\rho_j(Y_{(j)},Y'_{(j)})] \\
&\qquad\qquad -\; 2\,\mathbb{E}\,[\rho_i(X_{(i)},X'_{(i)})\,\rho_j(Y_{(j)},Y''_{(j)})]\Big\} \\
&=\; \frac{1}{4\tau_{XY}^2}\,\sum_{i=1}^{p}\sum_{j=1}^{q} D^2_{\rho_i,\,\rho_j}(X_{(i)},Y_{(j)})\,.
\end{aligned}
$$

To observe that the remainder term is negligible, note that under Assumption 3.4.2,

$$
\begin{aligned}
\mathbb{E}\,[L(X,X')R(Y,Y')] \;&\leq\; \big(\,\mathbb{E}\,[L(X,X')^2]\,\mathbb{E}\,[R(Y,Y')^2]\,\big)^{1/2} \;=\; O(a'_p b'^2_q)\;, \\
\mathbb{E}\,[R(X,X')L(Y,Y')] \;&\leq\; \big(\,\mathbb{E}\,[R(X,X')^2]\,\mathbb{E}\,[L(Y,Y')^2]\,\big)^{1/2} \;=\; O(a'^2_p b'_q)\;, \\
\mathbb{E}\,[R(X,X')R(Y,Y')] \;&\leq\; \big(\,\mathbb{E}\,[R(X,X')^2]\,\mathbb{E}\,[R(Y,Y')^2]\,\big)^{1/2} \;=\; O(a'^2_p b'^2_q)\;,
\end{aligned}
$$

Clearly, $\mathcal{R} = \tau_{XY} R = O(\tau_{XY}\, a'^2_p b'_q + \tau_{XY}\, a'_p b'^2_q)$. ♢

Proof of Theorem 4. The proof is essentially similar to the proof of Theorem 3. Note that using Proposition 3.3.1, we can write

$$
\begin{aligned}
\frac{1}{\tau_Y}\,\mathcal{D}^2(X,Y) \;&=\; \mathbb{E}\,K(X,X')\,\frac{K(Y,Y')}{\tau_Y} \;+\; \mathbb{E}\,K(X,X')\,\mathbb{E}\,\frac{K(Y,Y')}{\tau_Y} \;-\; 2\,\mathbb{E}\,K(X,X')\,\frac{K(Y,Y'')}{\tau_Y} \\
&=\; \mathbb{E}\,K(X,X')\,\left(1+\frac{1}{2}L(Y,Y')+R(Y,Y')\right) \\
&\quad +\; \mathbb{E}\,K(X,X')\,\mathbb{E}\,\left(1+\frac{1}{2}L(Y,Y')+R(Y,Y')\right) \\
&\quad -\; 2\,\mathbb{E}\,K(X,X')\,\left(1+\frac{1}{2}L(Y,Y'')+R(Y,Y'')\right) \\
&=\; L \;+\; R,
\end{aligned}
$$

where

$$\begin{aligned} L \;&=\; \frac{1}{2\tau_Y^2}\sum_{j=1}^{q}\Big\{\,\mathbb{E}\,[K(X,X')\,\rho_j(Y_{(j)},Y'_{(j)})] \;+\; \mathbb{E}\,[K(X,X')\,\mathbb{E}\,[\rho_j(Y_{(j)},Y'_{(j)})] \\ &\quad -\; 2\,\mathbb{E}\,[K(X,X')\,\rho_j(Y_{(j)},Y''_{(j)})]\Big\} \\ &=\; \frac{1}{2\tau_Y^2}\sum_{j=1}^{q} D^2_{K,\,\rho_j}(X,Y_{(j)})\,, \end{aligned}$$

and

$$R \;= \mathbb{E}\;[\,K(X,X')R(Y,Y')\,] \;+\; \mathbb{E}\;[K(X,X')]\;\mathbb{E}\,[R(Y,Y')] \;-\; 2\,\mathbb{E}\;[\,K(X,X')R(Y,Y'')\,]\;.$$

Under the assumption that $\mathbb{E}\,[R^2(Y,Y')] = O(b_q'^4)$, using Hölder's inequality it is easy to see that $\tau_Y R = O(\tau_Y\, b_q'^2) = o(1)$.

$\diamondsuit$

Proof of Theorem 5. Following equation (B.29), we have for $1 \le k \ne l \le n$

$$\begin{aligned} \widetilde{D}^X_{kl} \;&=\; \frac{\tau_X}{2}\widetilde{L}(X_k,X_l) \;+\; \tau_X\widetilde{R}(X_k,X_l) \;=\; \frac{1}{2\tau_X}\sum_{i=1}^{p}\widetilde{\rho}_i(X_{k(i)},X_{l(i)}) \;+\; \tau_X\widetilde{R}(X_k,X_l)\,, \\ \widetilde{D}^Y_{kl} \;&=\; \frac{\tau_Y}{2}\widetilde{L}(Y_k,Y_l) \;+\; \tau_Y\widetilde{R}(Y_k,Y_l) \;=\; \frac{1}{2\tau_Y}\sum_{j=1}^{q}\widetilde{\rho}_i(Y_{k(j)},Y_{l(j)}) \;+\; \tau_Y\widetilde{R}(Y_k,Y_l)\,. \end{aligned}$$

From equation (1.14) in Chapter 3 it is easy to check that

$$\begin{aligned} \widetilde{\mathcal{D}^2_n}(X,Y) \;&=\; \frac{1}{4\tau_{XY}}\sum_{i=1}^{p}\sum_{j=1}^{q}\widetilde{D^2_{n\,;\,\rho_i,\rho_j}}(X_{(i)},Y_{(j)}) \;+\; \frac{\tau_{XY}}{2n(n-3)}\sum_{k\ne l}\widetilde{L}(X_k,X_l)\widetilde{R}(Y_k,Y_l) \\ &\quad +\; \frac{\tau_{XY}}{2n(n-3)}\sum_{k\ne l}\widetilde{L}(Y_k,Y_l)\widetilde{R}(X_k,X_l) \;+\; \frac{\tau_{XY}}{n(n-3)}\sum_{k\ne l}\widetilde{R}(X_k,X_l)\,\widetilde{R}(Y_k,Y_l)\,. \end{aligned}$$

Under Assumption 3.4.3, using Hölder's inequality and power mean inequality, it can be verified that

$$\begin{aligned}
\sum_{k\neq l} \widetilde{L}(X_k,X_l)\widetilde{R}(Y_k,Y_l) \;&\leq\; \left(\sum_{k\neq l} \widetilde{L}(X_k,X_l)^2 \sum_{k\neq l} \widetilde{R}(Y_k,Y_l)^2\right)^{1/2} \;=\; O_p(a_p b_q^2)\;, \\
\sum_{k\neq l} \widetilde{L}(Y_k,Y_l)\widetilde{R}(X_k,X_l) \;&\leq\; \left(\sum_{k\neq l} \widetilde{L}(Y_k,Y_l)^2 \sum_{k\neq l} \widetilde{R}(X_k,X_l)^2\right)^{1/2} \;=\; O_p(a_p^2 b_q)\;, \\
\sum_{k\neq l} \widetilde{R}(X_k,X_l)\widetilde{R}(Y_k,Y_l) \;&\leq\; \left(\sum_{k\neq l} \widetilde{R}(X_k,X_l)^2 \sum_{k\neq l} \widetilde{R}(Y_k,Y_l)^2\right)^{1/2} \;=\; O_p(a_p^2 b_q^2)\;.
\end{aligned}$$

This completes the proof of the theorem. ♢

Proof of Theorem 6. Following equation (B.29), we have for $1 \leq k \neq l \leq n$

$$\widetilde{D}^Y_{kl} \;=\; \frac{1}{2\tau_Y}\sum_{j=1}^{q} \widetilde{\rho}_j(Y_{k(j)}, Y_{l(j)}) \;+\; \tau_Y \widetilde{R}(Y_k,Y_l)\,,$$

and therefore

$$\widetilde{\mathcal{D}^2_n}(X,Y) \;=\; \frac{1}{2\tau_Y}\sum_{j=1}^{q} \widetilde{D^2_{n\,;\,K,\rho_j}}(X,Y_{(j)}) \;+\; \frac{\tau_Y}{n(n-3)}\sum_{k\neq l} \widetilde{K}(X_k,X_l)\widetilde{R}(Y_k,Y_l)\,.$$

Using power mean inequality, it can be verified that $\sum_{k\neq l} \widetilde{K}(X_k,X_l)\widetilde{R}(Y_k,Y_l) \;=\; O_p(b_q^2)$. This completes the proof of the theorem. ♢

Proof of Theorem 7. The proof follows similar lines of the proof Theorem 2.2.1 in Zhu et al. (2020), with the distance metric being the one from the class of metrics we proposed in equation (3.2). ♢

Proof of Theorem 8. The proof of the theorem follows similar lines of the proof of Proposition 2.2.2 in Zhu et al. (2020). ♢

Proof of Theorem 16. The decomposition into the leading term follows the similar lines of the proof of Theorem 5. The negligibility of the remainder term can be shown by mimicking the proof of Theorem 3.1.1 in Zhu et al. (2020). ♢

Proof of Theorem 17. It essentially follows similar lines of Proposition 3.2.1 in Zhu et al. (2020).

$\diamondsuit$

APPENDIX C

This is the Appendix for Chapter 4.

C.1 Sketch of the proof of Theorem 9

From the proof of Lemma D.1 in the supplementary materials of Chakraborty and Zhang (2019), we can write under H_0

$$\mathcal{E}_{1,n,k} \;=\; L_{n,k} \,+\, R_{n,k} \tag{C.1}$$

where

$$\begin{aligned}
L_{n,k} \;&=\; \frac{1}{k(n-k)} \sum_{i_1=1}^{k} \sum_{i_2=k+1}^{n} H(X_{i_1}, X_{i_2}) \,-\, \frac{1}{k(k-1)} \sum_{1 \le i_1 < i_2 \le k} H(X_{i_1}, X_{i_2}) \\
&\quad -\, \frac{1}{(n-k)(n-k-1)} \sum_{k+1 \le i_1 < i_2 \le n} H(X_{i_1}, X_{i_1})\,, \\
\text{and}\;\; R_{n,k} \;&=\; \frac{2\tau}{k(n-k)} \sum_{i_1=1}^{k} \sum_{i_2=k+1}^{n} R(X_{i_1}, X_{i_2}) \,-\, \frac{\tau}{k(k-1)} \sum_{1 \le i_1 \neq i_2 \le k} R(X_{i_1}, X_{i_2}) \\
&\quad -\, \frac{\tau}{(n-k)(n-k-1)} \sum_{k+1 \le i_1 \neq i_2 \le n} R(X_{i_1}, X_{i_1})\,.
\end{aligned} \tag{C.2}$$

Following the discussions in Section D in the supplementary materials of Chakraborty and Zhang (2019), the variance of $L_{n,k}$ is given by

$$\begin{aligned}
V_{n,k} \;&:=\; \Big(\frac{1}{k(n-k)} + \frac{1}{2\,k(k-1)} + \frac{1}{2\,(n-k)(n-k-1)}\Big)\, \mathbb{E}\, H^2(X, X') \\
&=:\; V_{n,k\,;\,1} \,+\, V_{n,k\,;\,2} \,+\, V_{n,k\,;\,3}\,,
\end{aligned} \tag{C.3}$$

which can be estimated by

$$\begin{aligned}\hat{V}_{n,k} \;:=\; & \frac{1}{k(n-k)}\, 4\, cdCov^2_{k,n-k} \;+\; \frac{1}{2\,k(k-1)}\, 4\, \widetilde{\mathcal{D}^2_k} \;+\; \frac{1}{2\,(n-k)(n-k-1)}\, 4\, \widetilde{\mathcal{D}^2_{n-k}} \\ =:\; & \hat{V}_{n,k\,;\,1} \;+\; \hat{V}_{n,k\,;\,2} \;+\; \hat{V}_{n,k\,;\,3}\,.\end{aligned} \tag{C.4}$$

Define

$$\breve{T}_{1,n,k} \;=\; \frac{\mathcal{E}_{1,n,k}}{\sqrt{V_{n,k}}}\,. \tag{C.5}$$

For $1 \le l \le k < m \le n$, define $\widetilde{S}_n(k,m) := \sum_{i_2=k+1}^{m} \sum_{i_1=k}^{i_2-1} H(X_{i_1}, X_{i_2})$ and

$$\begin{aligned}\widetilde{L}_n(k\,; l, m) \;:=\; & \frac{1}{(k-l+1)(m-k)} \sum_{i_2=k+1}^{m} \sum_{i_1=l}^{k} H(X_{i_1}, X_{i_2}) \\ & -\; \frac{1}{(k-l+1)(k-l)} \sum_{l \le i_1 < i_2 \le k} H(X_{i_1}, X_{i_2}) \\ & -\; \frac{1}{(m-k)(m-k-1)} \sum_{k+1 \le i_1 < i_2 \le m} H(X_{i_1}, X_{i_1})\,.\end{aligned} \tag{C.6}$$

From (C.2) and (C.6), it is easy to see that $L_{n,k} = \widetilde{L}_n(k\,;1,n)$. With the definition of $\widetilde{S}_n(k,m)$ as above, we can write

$$\begin{aligned}\widetilde{L}_n(k\,; l, m) \;=\; & -\frac{1}{(k-l)(k-l+1)}\, \widetilde{S}_n(l,k) \;-\; \frac{1}{(m-k)(m-k-1)}\, \widetilde{S}_n(k+1,m) \\ & +\; \frac{1}{(k-l+1)(m-k)} \Big(\widetilde{S}_n(l,m) - \widetilde{S}_n(l,k) - \widetilde{S}_n(k+1,m) \Big)\,.\end{aligned} \tag{C.7}$$

Denote $S_n(a,b) := \widetilde{S}_n(\lfloor na \rfloor + 1, \lfloor nb \rfloor)$ for any $0 \le a < b \le 1$. Further let $l = \lfloor na \rfloor + 1$, $k = \lfloor nr \rfloor$, and $m = \lfloor nb \rfloor$ for $0 \le a < r < b \le 1$. Therefore from (C.7) we have

$$\begin{aligned}\widetilde{L}_n(k\,; l, m) \;=\; & -\frac{1}{(k-l)(k-l+1)}\, S_n(a,r) \;-\; \frac{1}{(m-k)(m-k-1)}\, S_n(r,b) \\ & +\; \frac{1}{(k-l+1)(m-k)} \Big(S_n(a,b) - S_n(a,r) - S_n(r,b) \Big)\,.\end{aligned} \tag{C.8}$$

Also define

$$\widetilde{V}_n(k\,;l,m) \;:=\; \Big(\frac{1}{(k-l+1)(m-k)} + \frac{1}{2(k-l)(k-l+1)} + \frac{1}{2(m-k)(m-k-1)}\Big)\,V_0\,, \tag{C.9}$$

where $V_0 := \mathbb{E}\,H^2(X,X')$. From (C.3) and (C.9), it is easy to check that $V_{n,k} = \widetilde{V}_n(k\,;1,n)$.

THEOREM **18**. *Under Assumption 4.3.2, as $n,p\to\infty$,*

$$\Big\{\frac{\sqrt{2}}{n\sqrt{V_0}}\,S_n(a,b)\Big\}_{a,b\,\in\,[0,1]} \;\xrightarrow{d}\; Q \qquad \textit{in}\;\; L^\infty\left([0,1]^2\right)\,,$$

where Q is a centered gaussian process with covariance function given by

$$\textit{cov}\,\big(Q(a_1,b_1)\,,\,Q(a_2,b_2)\big) \;=\; \big(b_1\wedge b_2 \,-\, a_1\vee a_2\big)^2\;\mathbb{1}\big(b_1\wedge b_2 > a_1\vee a_2\big)\,.$$

In particular, $var\,\big(Q(a,b)\big) = (b-a)^2\,\mathbb{1}(b>a)$.

The proof of Theorem 18 is given in Section C.2. From (C.8) and (C.9), we can write

$$\begin{aligned}\frac{\widetilde{L}_n(k\,;l,m)}{\sqrt{\widetilde{V}_n(k\,;l,m)}} \;&=\; \frac{1}{\sqrt{\frac{1}{(k-l+1)(m-k)}+\frac{1}{2(k-l)(k-l+1)}+\frac{1}{2(m-k)(m-k-1)}}\,\sqrt{V_0}} \times \Big[-\,\frac{1}{(k-l)(k-l+1)}\,S_n(a,r) \\ &\qquad -\,\frac{1}{(m-k)(m-k-1)}\,S_n(r,b) \,+\, \frac{1}{(k-l+1)(m-k)}\,\Big(S_n(a,b)-S_n(a,r)-S_n(r,b)\Big)\Big] \\ &= \frac{1}{n\sqrt{\frac{2}{(k-l+1)(m-k)}+\frac{1}{(k-l)(k-l+1)}+\frac{1}{(m-k)(m-k-1)}}\,\sqrt{V_0}} \times \Big[-\,\frac{n^2}{(k-l)(k-l+1)}\,\frac{\sqrt{2}\,S_n(a,r)}{n\sqrt{V_0}} \\ &\quad -\,\frac{n^2}{(m-k)(m-k-1)}\,\frac{\sqrt{2}\,S_n(r,b)}{n\sqrt{V_0}} + \frac{n^2}{(k-l+1)(m-k)}\Big(\frac{\sqrt{2}\,S_n(a,b)}{n\sqrt{V_0}} - \frac{\sqrt{2}\,S_n(a,r)}{n\sqrt{V_0}} - \frac{\sqrt{2}\,S_n(r,b)}{n\sqrt{V_0}}\Big)\Big]\,.\end{aligned}$$

Combining the above with Theorem 18, it is not hard to see that as $n,p\to\infty$,

$$\left\{\frac{\widetilde{L}_n(\lfloor nr\rfloor\,;\lfloor na\rfloor+1,\lfloor nb\rfloor)}{\sqrt{\widetilde{V}_n(\lfloor nr\rfloor\,;\lfloor na\rfloor+1,\lfloor nb\rfloor)}}\right\}_{a,r,b\,\in\,[0,1]} \;\xrightarrow{d}\; G \qquad \text{in}\;\; L^\infty\left([0,1]^3\right)\,, \tag{C.10}$$

where

$$G(r\,;\,a,b) \;:=\; \frac{1}{\sqrt{\frac{2}{(r-a)(b-r)}+\frac{1}{(r-a)^2}+\frac{1}{(b-r)^2}}} \times \Big[-\,\frac{1}{(r-a)^2}\,Q(a,r)\;-\;\frac{1}{(b-r)^2}\,Q(r,b) \\ +\;\frac{1}{(r-a)(b-r)}\Big(Q(a,b)-Q(a,r)-Q(r,b)\Big)\Big]$$

for $0 \le a < r < b \le 1$ and zero otherwise. As a further consequence, when $a = 0$ and $b = 1$, we have

$$\left\{\frac{\widetilde{L}_n(\lfloor nr\rfloor\,;1,n))}{\sqrt{\widetilde{V}_n(\lfloor nr\rfloor\,;1,n)}}\right\}_{r\,\in\,[0,1]} \xrightarrow{d} G_0 \qquad \text{in } L^\infty([0,1])\;, \tag{C.11}$$

where

$$\begin{aligned} G_0(r) &:= \frac{1}{\sqrt{\frac{2}{r(1-r)}+\frac{1}{r^2}+\frac{1}{(1-r)^2}}} \times \Big[-\,\frac{1}{r^2}\,Q(0,r)\;-\;\frac{1}{(1-r)^2}\,Q(r,1)\,+ \\ &\qquad \frac{1}{r(1-r)}\Big(Q(0,1)-Q(0,r)-Q(r,1)\Big)\Big] \\ &=\; Q(0,1)\;-\;\frac{1}{r}\,Q(0,r)\;-\;\frac{1}{1-r}\,Q(r,1) \end{aligned} \tag{C.12}$$

for $0 < r < 1$ and zero otherwise. The second equality in (C.12) follows from some straightforward calculations.

Now for $1 \le l \le k < m \le n$, define $\widetilde{R}_n(k,m) := \sum_{i_2=k+1}^{m}\sum_{i_1=k}^{i_2-1}\tau\,R(X_{i_1},X_{i_2})$ and

$$\widetilde{Q}_n(k\,;l,m) \;:=\; \frac{2\tau}{(k-l+1)(m-k)}\sum_{i_2=k+1}^{m}\sum_{i_1=l}^{k}R(X_{i_1},X_{i_2}) \;-\; \frac{\tau}{(k-l+1)(k-l)}\sum_{l\le i_1\ne i_2\le k}R(X_{i_1},X_{i_2}) \\ -\;\frac{\tau}{(m-k)(m-k-1)}\sum_{k+1\le i_1\ne i_2\le m}R(X_{i_1},X_{i_1})\,. \tag{C.13}$$

Comparing (C.2) and (C.13), it is easy to verify that $R_{n,k} = \widetilde{Q}_n(k\,;1,n)$. With the definition of

$\widetilde{R}_n(k,m)$ as above, clearly we have

$$\begin{aligned}\widetilde{Q}_n(k\,;l,m) \;=\; & \frac{2}{(k-l+1)(m-k)}\Big(\widetilde{R}_n(l,m)-\widetilde{R}_n(l,k)-\widetilde{R}_n(k+1,m)\Big)\\ & -\;\frac{2}{(k-l+1)(k-l)}\,\widetilde{R}_n(l,k)\;-\;\frac{2}{(m-k)(m-k-1)}\,\widetilde{R}_n(k+1,m)\,.\end{aligned}\tag{C.14}$$

Denote $R_n(a,b) := \widetilde{R}_n(\lfloor na\rfloor+1, \lfloor nb\rfloor)$ for any $0 \le a < b \le 1$. Therefore we have from (C.15)

$$\begin{aligned}\widetilde{Q}_n(k\,;l,m) \;=\; & \frac{2}{(k-l+1)(m-k)}\Big(R_n(a,b)-R_n(a,r)-R_n(r,b)\Big)\;-\;\frac{2}{(k-l+1)(k-l)}\,R_n(a,r)\\ & -\;\frac{2}{(m-k)(m-k-1)}\,R_n(r,b)\,.\end{aligned}\tag{C.15}$$

Define $G_n(a,b) := \frac{1}{n\sqrt{V_0}}\,R_n(a,b)$.

THEOREM **19**. *Under Assumption 4.3.3, as* $n,p\to\infty$, $\sup\limits_{a,b\,\in\,[0,1]^2}|G_n(a,b)| \;=\; o_p(1)$.

The proof of Theorem 19 is given in Section C.2. From (C.9) and (C.15), we have

$$\begin{aligned}\frac{\widetilde{Q}_n(k\,;l,m)}{\sqrt{\widetilde{V}_n(k\,;l,m)}} \;=\; & \frac{1}{\sqrt{\frac{1}{(k-l+1)(m-k)}+\frac{1}{2(k-l)(k-l+1)}+\frac{1}{2(m-k)(m-k-1)}}\,\sqrt{V_0}}\;\times\\ & \Big[\frac{2}{(k-l+1)(m-k)}\Big(R_n(a,b)-R_n(a,r)-R_n(r,b)\Big)\\ & -\;\frac{2}{(k-l)(k-l+1)}\,R_n(a,r)\;+\;\frac{2}{(m-k)(m-k-1)}\,R_n(r,b)\Big]\,.\end{aligned}$$

Multiplying both the numerator and denominator above by n^2, it is not hard to see that as a consequence of Theorem 19 we have

$$\sup_{a,r,b\,\in\,[0,1]}\left|\frac{\widetilde{Q}_n(\lfloor nr\rfloor\,;\lfloor na\rfloor+1,\lfloor nb\rfloor)}{\sqrt{\widetilde{V}_n(\lfloor nr\rfloor\,;\lfloor na\rfloor+1,\lfloor nb\rfloor)}}\right| \;=\; o_p(1) \qquad \text{as } \; n,p\to\infty\,. \tag{C.16}$$

As a special case, putting $a = 0$ and $b = 1$, we get from (C.16)

$$\sup_{r \in [0,1]} \left| \frac{\widetilde{Q}_n(\lfloor nr \rfloor\,; 1, n)}{\sqrt{\widetilde{V}_n(\lfloor nr \rfloor\,; 1, n)}} \right| \;=\; o_p(1) \qquad \text{as } \; n, p \to \infty\,. \tag{C.17}$$

THEOREM **20**. *Under Assumptions 4.3.2 and 4.3.3, as $n, p \to \infty$,*

$$\sup_{r \in [0,1]} \left| \frac{\hat{V}_{n,\lfloor nr \rfloor}}{V_{n,\lfloor nr \rfloor}} - 1 \right| \;=\; o_p(1)\,.$$

With all the above, the proof of Theorem 9 can be completed as below.

Proof of Theorem 9. Combining (C.1) and (C.5) with (C.11) and (C.17) yields

$$\left\{ \breve{T}_{1,n,\lfloor nr \rfloor} \right\}_{r \in [0,1]} \;\xrightarrow{d}\; G_0 \qquad \text{in } \; L^{\infty}\left([0,1]\right)\,,$$

as $n, p \to \infty$. This equipped with Theorem 20 completes the proof of Theorem 9. ♢

C.2 Technical Appendix

Proof of Theorem 18. To establish the uniform weak convergence of $\frac{\sqrt{2}}{n\sqrt{V_0}}\, S_n(a, b)$, by Theorem 10.2 in Pollard (1990) we need to show :

T1. the finite dimensional convergence, viz.

$$\Big(\frac{\sqrt{2}}{n\sqrt{V_0}}\, S_n(a_1, b_1)\,,\, \ldots\,,\, \frac{\sqrt{2}}{n\sqrt{V_0}}\, S_n(a_s, b_s) \Big) \;\xrightarrow{d}\; \Big(Q(a_1, b_1),\, \ldots\,, Q(a_s, b_s) \Big)$$

as $\; n, p \to \infty \;$ for fixed $\; 0 \le a_i < b_i \le 1,\, 1 \le i \le s$, and

T2. asymptotic stochastic equicontinuity of $\frac{\sqrt{2}}{n\sqrt{V_0}}\, S_n(a, b)$ on $[0, 1]^2$, viz. for any $x > 0$,

$$\lim_{\delta \downarrow 0}\, \limsup_{n,p \to \infty}\; P\Big(\sup_{\|(a,b)-(c,d)\| \le \delta} \Big| \frac{\sqrt{2}}{n\sqrt{V_0}}\, S_n(a, b) \;-\; \frac{\sqrt{2}}{n\sqrt{V_0}}\, S_n(c, d) \Big| \Big) \;=\; 0\,.$$

To prove T1, we will consider the case $s = 2$ and the general case can be proved in a similar fashion). By Cramér-Wold theorem, it is equivalent to prove

$$\alpha_1 \frac{\sqrt{2}}{n\sqrt{V_0}} S_n(a_1, b_1) + \alpha_2 \frac{\sqrt{2}}{n\sqrt{V_0}} S_n(a_2, b_2) \xrightarrow{d} \alpha_1 Q(a_1, b_1) + \alpha_2 Q(a_2, b_2) \tag{C.18}$$

for any fixed $\alpha_1, \alpha_2 \in \mathbb{R}$, as $n, p \to \infty$. As $0 \le a_i < b_i \le 1$, $i = 1, 2$, we consider the following three cases : i) $a_1 \le a_2 \le b_2 \le b_1$, ii) $a_1 \le a_2 \le b_1 \le b_2$, and iii) $a_1 \le b_1 \le a_2 \le b_2$.

Consider case (ii). We will prove T1 and T2 for this case and similar arguments can prove them for the other two cases.

Proof of T1. We can write

$$\begin{aligned} &\alpha_1 \frac{\sqrt{2}}{n\sqrt{V_0}} S_n(a_1, b_1) + \alpha_2 \frac{\sqrt{2}}{n\sqrt{V_0}} S_n(a_2, b_2) \\ &= \frac{\sqrt{2}}{n\sqrt{V_0}} \Big[\alpha_1 \sum_{i=\lfloor na_1 \rfloor+2}^{\lfloor nb_1 \rfloor} \sum_{j=\lfloor na_1 \rfloor+1}^{i-1} H(X_i, X_j) + \alpha_2 \sum_{i=\lfloor na_2 \rfloor+2}^{\lfloor nb_2 \rfloor} \sum_{j=\lfloor na_2 \rfloor+1}^{i-1} H(X_i, X_j)\Big] \\ &= \sum_{i=\lfloor na_1 \rfloor+2}^{\lfloor nb_2 \rfloor} \widetilde{\xi}_{n,i}, \end{aligned} \tag{C.19}$$

where

$$\widetilde{\xi}_{n,i} := \frac{\sqrt{2}}{n\sqrt{V_0}} \begin{cases} \alpha_1 \xi_{1,i} & \text{if } \lfloor na_1 \rfloor + 2 \le i \le \lfloor na_2 \rfloor + 1 \\ \alpha_1 \xi_{1,i} + \alpha_2 \xi_{2,i} & \text{if } \lfloor na_2 \rfloor + 2 \le i \le \lfloor nb_1 \rfloor \\ \alpha_2 \xi_{2,i} & \text{if } \lfloor nb_1 \rfloor + 1 \le i \le \lfloor nb_2 \rfloor \end{cases}, \tag{C.20}$$

$$\text{with} \quad \xi_{1,i} := \sum_{j=\lfloor na_1 \rfloor+1}^{i-1} H(X_i, X_j), \quad \text{and} \quad \xi_{2,i} := \sum_{j=\lfloor na_2 \rfloor+1}^{i-1} H(X_i, X_j).$$

Define $\mathcal{F}_i := \sigma(X_i, X_{i-1}, \dots)$. By Theorem 3.2 and Corollary 3.1 in Hall and Heyde (1980), it suffices to show :

P1. For each $n \ge 1$, $\{\sum_{l=2}^{i} \widetilde{\xi}_{n,\lfloor na_1 \rfloor+l}, \mathcal{F}_l\}_{i=2}^{\lfloor nb_2 \rfloor - \lfloor na_1 \rfloor}$ is a sequence of zero mean, square integrable martingales.

P2. $V_n := \sum_{i=2}^{\lfloor nb_2 \rfloor - \lfloor na_1 \rfloor} \mathbb{E}\left[\widetilde{\xi}^2_{n,\lfloor na_1 \rfloor + i} \,|\, \mathcal{F}_{\lfloor na_1 \rfloor + i - 1}\right] \xrightarrow{P} \alpha_1^2\,(b_1 - a_1)^2 + \alpha_2^2\,(b_2 - a_2)^2 + 2\,\alpha_1\,\alpha_2\,(b_1 - a_2)^2$, as $n, p \to \infty$.

P3. $\sum_{i=\lfloor na_1 \rfloor + 2}^{\lfloor nb_2 \rfloor} \mathbb{E}\left[\widetilde{\xi}^4_{n,i}\right] \longrightarrow 0$, as $n, p \to \infty$.

From Theorem 3.2 in Hall and Heyde (1980),the variance of $\alpha_1\,Q(a_1,b_1) + \alpha_2\,Q(a_2,b_2)$ should be $\operatorname{plim}_{n,p\to\infty} V_n$ as in P2. From there it is intuitive that

$$\text{cov}\left(Q(a_1,b_1)\,,\,Q(a_2,b_2)\right) \;=\; \left(b_1 \wedge b_2 \,-\, a_1 \vee a_2\right)^2\, \mathbb{1}\left(b_1 \wedge b_2 > a_1 \vee a_2\right).$$

To show P1, it is easy to see that $\widetilde{\xi}_{n,\lfloor na_1 \rfloor + l}$ is square integrable, $\mathbb{E}\left(\widetilde{\xi}_{n,\lfloor na_1 \rfloor + l}\right) = 0$ and $\mathcal{F}_2 \subseteq \mathcal{F}_l$.

To prove P3, note that using the power mean inequality

$$\left|\sum_{i=1}^{n} a_i\right|^r \;\le\; n^{r-1}\,\sum_{i=1}^{n} |a_i|^r \tag{C.21}$$

for $a_i \in \mathbb{R}$, $1 \le i \le n$, $n \ge 2$ and $r > 1$, we can write

$$\begin{aligned}
\sum_{i=\lfloor na_1 \rfloor+2}^{\lfloor nb_2 \rfloor} \mathbb{E}\left[\widetilde{\xi}^4_{n,i}\right] \;&=\; \sum_{i=\lfloor na_1 \rfloor+2}^{\lfloor na_2 \rfloor+1} \mathbb{E}\left[\widetilde{\xi}^4_{n,i}\right] \,+\, \sum_{i=\lfloor na_2 \rfloor+2}^{\lfloor nb_1 \rfloor} \mathbb{E}\left[\widetilde{\xi}^4_{n,i}\right] \,+\, \sum_{i=\lfloor nb_1 \rfloor+1}^{\lfloor nb_2 \rfloor} \mathbb{E}\left[\widetilde{\xi}^4_{n,i}\right] \\
&=\; \sum_{i=\lfloor na_1 \rfloor+2}^{\lfloor na_2 \rfloor+1} \mathbb{E}\left[\alpha_1\,\frac{\sqrt{2}}{n\sqrt{V_0}}\,\xi_{1,i}\right]^4 \,+\, \sum_{i=\lfloor na_2 \rfloor+2}^{\lfloor nb_1 \rfloor} \mathbb{E}\left[\alpha_1\,\frac{\sqrt{2}}{n\sqrt{V_0}}\,\xi_{1,i} \,+\, \alpha_2\,\frac{\sqrt{2}}{n\sqrt{V_0}}\,\xi_{2,i}\right]^4 \\
&\quad+\, \sum_{i=\lfloor nb_1 \rfloor+1}^{\lfloor nb_2 \rfloor} \mathbb{E}\left[\alpha_2\,\frac{\sqrt{2}}{n\sqrt{V_0}}\,\xi_{2,i}\right]^4 \\
&\lesssim\; \frac{1}{n^4\,V_0^2}\Big(\alpha_1^4 \sum_{i=\lfloor na_1 \rfloor+2}^{\lfloor nb_1 \rfloor} \mathbb{E}\left[\xi^4_{1,i}\right] \,+\, \alpha_2^4 \sum_{i=\lfloor na_2 \rfloor+2}^{\lfloor nb_2 \rfloor} \mathbb{E}\left[\xi^4_{2,i}\right]\Big).
\end{aligned} \tag{C.22}$$

We have essentially used the definitions in (C.20) in the above calculations. Now for the first summand in the right hand side of (C.22), note that using (C.20) we have

$$
\begin{aligned}
&\frac{1}{n^4\,V_0^2} \sum_{i=\lfloor na_1\rfloor+2}^{\lfloor nb_1\rfloor} \mathbb{E}\left[\xi_{1,i}^4\right] \;=\; \frac{1}{n^4\,V_0^2} \sum_{i=\lfloor na_1\rfloor+2}^{\lfloor nb_1\rfloor} \mathbb{E}\Big[\sum_{j=\lfloor na_1\rfloor+1}^{i-1} H(X_i,X_j)\Big]^4 \\
&= \frac{1}{n^4\,V_0^2} \sum_{i=\lfloor na_1\rfloor+2}^{\lfloor nb_1\rfloor} \Big[\sum_{j=\lfloor na_1\rfloor+1}^{i-1} \mathbb{E}\,H^4(X_i,X_j) \;+\; 3 \sum_{\lfloor na_1\rfloor+1\le j_1\neq j_2\le i-1} \mathbb{E}\,H^2(X_i,X_{j_1})\,H^2(X_i,X_{j_2})\Big] \\
&= \frac{1}{n^4}\,O\Big(\frac{n^2\,\mathbb{E}\,H^4(X,X') \;+\; n^3\,\mathbb{E}\,H^2(X,X')H^2(X,X'')}{\big[\mathbb{E}\,H^2(X,X')\big]^2}\Big)\,.
\end{aligned}
\tag{C.23}
$$

This is because $\lfloor na\rfloor \asymp n$ for $0 < a \le 1$. In fact it is easy to see that $\lim_{n\to\infty} \frac{\lfloor na\rfloor}{n} = \lim_{n\to\infty} \frac{na - \{na\}}{n} = a$, as $0 \le \{na\} < 1$.

Similar expressions hold for the second summand in the right hand side of (C.22). With this, it is easy to see that under Assumption 4.3.2

$$
\sum_{i=\lfloor na_1\rfloor+2}^{\lfloor nb_2\rfloor} \mathbb{E}\left[\widetilde{\xi}_{n,i}^4\right] \;=\; o(1) \qquad \text{as } \; n,p\to\infty\,,
$$

which completes the proof of P3.

To prove P2, write

$$
V_n \;=\; \sum_{i=2}^{\lfloor nb_2\rfloor-\lfloor na_1\rfloor} \mathbb{E}\left[\widetilde{\xi}_{n,\lfloor na_1\rfloor+i}^2 \,|\, \mathcal{F}_{\lfloor na_1\rfloor+i-1}\right] \;=\; \sum_{l=\lfloor na_1\rfloor+2}^{\lfloor nb_2\rfloor} \mathbb{E}\left[\widetilde{\xi}_{n,l}^2 \,|\, \mathcal{F}_{l-1}\right] \tag{C.24}
$$

where we have simply substituted $l = \lfloor na_1 \rfloor + i$. From (C.24) we have

$$\begin{aligned} V_n \;&=\; \sum_{l=\lfloor na_1 \rfloor+2}^{\lfloor na_2 \rfloor+1} \mathbb{E}\Big[\Big(\frac{\sqrt{2}}{n\sqrt{V_0}}\;\alpha_1\,\xi_{1,l}\Big)^2 \,\big|\, \mathcal{F}_{l-1}\Big] \;+\; \sum_{l=\lfloor na_2 \rfloor+2}^{\lfloor nb_1 \rfloor} \mathbb{E}\Big[\Big\{\frac{\sqrt{2}}{n\sqrt{V_0}}\;\big(\alpha_1\,\xi_{1,l} \;+\; \alpha_2\,\xi_{2,l}\big)\Big\}^2 \,\big|\, \mathcal{F}_{l-1}\Big] \\ &\quad+\; \sum_{l=\lfloor nb_1 \rfloor+1}^{\lfloor nb_2 \rfloor} \mathbb{E}\Big[\Big(\frac{\sqrt{2}}{n\sqrt{V_0}}\;\alpha_2\,\xi_{2,l}\Big)^2 \,\big|\, \mathcal{F}_{l-1}\Big] \\ &=\; \frac{2}{n^2 V_0}\Big(\alpha_1^2 \sum_{l=\lfloor na_1 \rfloor+2}^{\lfloor nb_1 \rfloor} \mathbb{E}\left[\xi_{1,l}^2 \,\big|\, \mathcal{F}_{l-1}\right] \;+\; \alpha_2^2 \sum_{l=\lfloor na_2 \rfloor+2}^{\lfloor nb_2 \rfloor} \mathbb{E}\left[\xi_{2,l}^2 \,\big|\, \mathcal{F}_{l-1}\right] + \\ &\qquad 2\,\alpha_1\,\alpha_2 \sum_{l=\lfloor na_2 \rfloor+2}^{\lfloor nb_1 \rfloor} \mathbb{E}\left[\xi_{1,l}\,\xi_{2,l} \,\big|\, \mathcal{F}_{l-1}\right]\Big) \\ &=\; \alpha_1^2\, V_{1n} \;+\; \alpha_2^2\, V_{2n} \;+\; 2\,\alpha_1\alpha_2\, V_{3n}\,, \end{aligned} \tag{C.25}$$

where

$$\begin{aligned} V_{1n} \;&=\; \frac{2}{n^2 V_0} \sum_{l=\lfloor na_1 \rfloor+2}^{\lfloor nb_1 \rfloor} \mathbb{E}\left[\xi_{1,l}^2 \,\big|\, \mathcal{F}_{l-1}\right], \\ V_{2n} \;&=\; \frac{2}{n^2 V_0} \sum_{l=\lfloor na_2 \rfloor+2}^{\lfloor nb_2 \rfloor} \mathbb{E}\left[\xi_{2,l}^2 \,\big|\, \mathcal{F}_{l-1}\right], \\ V_{3n} \;&=\; \frac{2}{n^2 V_0} \sum_{l=\lfloor na_2 \rfloor+2}^{\lfloor nb_1 \rfloor} \mathbb{E}\left[\xi_{1,l}\,\xi_{2,l} \,\big|\, \mathcal{F}_{l-1}\right]. \end{aligned} \tag{C.26}$$

Using the definition of $\xi_{1,l}$ from (C.20), we can write

$$V_{1n} \;=\; \frac{2}{n^2 V_0} \sum_{l=\lfloor na_1 \rfloor+2}^{\lfloor nb_1 \rfloor} \;\sum_{j_1,j_2=\lfloor na_1 \rfloor+1}^{l-1} \mathbb{E}\left[H(X_l, X_{j_1})\, H(X_l, X_{j_2}) \,\big|\, \mathcal{F}_{l-1}\right], \tag{C.27}$$

and therefore

$$\mathbb{E}\, V_{1n} \;=\; \frac{2}{n^2 V_0} \sum_{l=\lfloor na_1 \rfloor+2}^{\lfloor nb_1 \rfloor} \;\sum_{j=\lfloor na_1 \rfloor+1}^{l-1} \mathbb{E}\left[H^2(X_l, X_j)\right], \tag{C.28}$$

as $\mathbb{E}\left[H(X_l, X_{j_1})\, H(X_l, X_{j_2})\right] \;=\; 0$ for $j_1 \neq j_2$. Using the fact that $V_0 = \mathbb{E}\left[H^2(X, X')\right]$, some straightforward calculations yield

$$
\begin{aligned}
\mathbb{E}\, V_{1n} &= \frac{2}{n^2 V_0} \sum_{\lfloor na_1 \rfloor + 1 \leq j < l \leq \lfloor nb_1 \rfloor} \mathbb{E}\left[H^2(X, X')\right] = \frac{2}{n^2} \binom{\lfloor nb_1 \rfloor - \lfloor na_1 \rfloor}{2} \\
&= \frac{1}{n^2} \left(\lfloor nb_1 \rfloor - \lfloor na_1 \rfloor\right) \left(\lfloor nb_1 \rfloor - \lfloor na_1 \rfloor - 1\right) \\
&\to (b_1 - a_1)^2 \,,
\end{aligned}
\tag{C.29}
$$

as $n \to \infty$. Define $L_l(j_1, j_2) := \mathbb{E}\left[H(X_l, X_{j_1})\, H(X_l, X_{j_2}) \,\middle|\, \mathcal{F}_{l-1}\right]$. Then from (C.27) we can write

$$
V_{1n} = \frac{2}{n^2 V_0} \sum_{l = \lfloor na_1 \rfloor + 2}^{\lfloor nb_1 \rfloor} \sum_{j_1, j_2 = \lfloor na_1 \rfloor + 1}^{l-1} L_l(j_1, j_2) \,,
$$

and therefore

$$
\operatorname{var}(V_{1n}) = \frac{4}{n^4 V_0^2} \sum_{l, l' = \lfloor na_1 \rfloor + 2}^{\lfloor nb_1 \rfloor} \sum_{j_1, j_2 = \lfloor na_1 \rfloor + 1}^{l-1} \sum_{j_1', j_2' = \lfloor na_1 \rfloor + 1}^{l'-1} \operatorname{cov}\left(L_l(j_1, j_2), L_{l'}(j_1', j_2')\right) .
$$

Following the proof of Lemma D.1 in the supplementary materials of Chakraborty and Zhang (2019), we have $\mathbb{E}\, L_l(j_1, j_2) = 0$ for $j_1 \neq j_2$, and

$$
\begin{aligned}
&\mathbb{E}\left[L_l(j_1, j_2)\, L_{l'}(j_1', j_2')\right] \\
&= \begin{cases}
\mathbb{E}\left[H^2(X_l, X_{j_1})\, H^2(X_{l'}', X_{j_1})\right] & \text{if } \; j_1 = j_2 = j_1' = j_2' \,, \\
\mathbb{E}\left[H(X_l, X_{j_1})\, H(X_l, X_{j_2})\, H(X_{l'}', X_{j_1})\, H(X_{l'}', X_{j_2})\right] & \text{if } \; j_1 = j_1' \neq j_2 = j_2' \\
 & \text{or } \; j_1 = j_2' \neq j_1' = j_2 \,, \\
\mathbb{E}\left[H^2(X_l, X_{j_1})\right] \mathbb{E}\left[H^2(X_{l'}, X_{j_1'})\right] & \text{if } \; j_1 = j_2 \neq j_1' = j_2' \,.
\end{cases}
\end{aligned}
$$

where the above expression holds for $l = l'$ as well. Therefore

$$\begin{aligned}\mathrm{var}\,(V_{1n}) \;&=\; \frac{4}{n^4 V_0^2}\Bigg[\sum_{l=l'}\Big\{\sum_{j_1=\lfloor na_1\rfloor+1}^{l-1}\mathrm{cov}\left(H^2(X_l, X_{j_1}), H^2(X'_l, X_{j_1})\right)\\ &\qquad\qquad +\; 2\sum_{\lfloor na_1\rfloor+1\le j_1\ne j_2\le l-1}\mathbb{E}\left[H(X_l, X_{j_1})\,H(X_l, X_{j_2})\,H(X'_l, X_{j_1})\,H(X'_l, X_{j_2})\right]\Big\}\\ &\qquad +\; 2\sum_{\lfloor na_1\rfloor+2\le l<l'\le\lfloor nb_1\rfloor}\Big\{\sum_{j_1=\lfloor na_1\rfloor+1}^{l-1}\mathrm{cov}\left(H^2(X_l, X_{j_1}), H^2(X'_{l'}, X_{j_1})\right)\\ &\qquad\qquad +\; 2\sum_{\lfloor na_1\rfloor+1\le j_1\ne j_2\le l-1}\mathbb{E}\left[H(X_l, X_{j_1})\,H(X_l, X_{j_2})\,H(X'_{l'}, X_{j_1})\,H(X'_{l'}, X_{j_2})\right]\Big\}\Bigg]\,.\end{aligned}$$

This implies

$$\begin{aligned}\mathrm{var}\,(V_{1n}) \;&=\; \frac{1}{n^4 V_0^2}\,O\Big(n^3\,\mathbb{E}\left[H^2(X, X')\,H^2(X, X'')\right]\\ &\qquad +\; n^4\,\mathbb{E}\left[H(X, X'')\,H(X', X'')\,H(X, X''')\,H(X', X''')\right]\Big)\\ &=\; o(1)\,,\end{aligned} \tag{C.30}$$

as $n, p \to \infty$, under Assumption 4.3.2. Combining (C.29) and (C.30), we get

$$\begin{aligned}\mathbb{E}\left(V_{1n} - (b_1 - a_1)^2\right)^2 \;&=\; \mathrm{var}\,(V_{1n}) \;+\; \left(\mathbb{E}\,V_{1n} - (b_1 - a_1)^2\right)^2\\ &=\; o(1)\,,\end{aligned}$$

which combined with Chebyshev's inequality implies

$$V_{1n} \xrightarrow{P} (b_1 - a_1)^2 \qquad \text{as} \quad n, p \to \infty\,. \tag{C.31}$$

Likewise it can be shown that as $n, p \to \infty$,

$$V_{2n} \xrightarrow{P} (b_2 - a_2)^2 \qquad \text{and} \qquad V_{3n} \xrightarrow{P} (b_1 - a_2)^2\,. \tag{C.32}$$

Combining (C.31) and (C.32), we get from (C.25)

$$V_n \xrightarrow{P} \alpha_1^2 (b_1 - a_1)^2 + \alpha_2^2 (b_2 - a_2)^2 + 2\,\alpha_1\,\alpha_2\,(b_1 - a_2)^2 \,. \tag{C.33}$$

This completes the proof of P2, and thereby the proof of T1, i.e., the finite dimensional convergence. ♢

Proof of T2. Denote $u = (a, b)$ and $v = (c, d)$. Also define $W_n(u) := \frac{\sqrt{2}}{n\sqrt{V_0}}\, S_n(u)$ for $u \in [0, 1]^2$. To prove the stochastic equicontinuity of $W_n(u)$ for $u \in [0, 1]^2$, we need to show for any $\epsilon > 0$

$$\lim_{\delta \downarrow 0} \limsup_{n,p \to \infty} P\Big(\sup_{\substack{u,v \,\in\, [0,1]^2 \\ d(u,v) < \delta}} \big|W_n(u) - W_n(v)\big| \Big) = 0 \,,$$

where $\big([0, 1]^2, d\big)$ is compact.

By Theorem A.8 in Li and Racine (2007), it suffices to show that $\forall\, u, v \in [0, 1]^2$

$$\mathbb{E}\,\big|W_n(u) - W_n(v)\big|^{\alpha} \lesssim d^{\gamma}(u, v) \tag{C.34}$$

for some $\alpha > 0$ and $\gamma > 1$. For our purpose, we choose $d(u, v) = \|u - v\|_1^{1/2}$ for $u, v \in [0, 1]^2$. Note that $[0, 1]^2 \subseteq \mathbb{R}^2$ is compact (closed and bounded) with respect to the metric $\rho(u, v) = \|u - v\|_1$. It is easy to verify that $[0, 1]^2$ is closed and bounded (and hence compact) with respect to the metric $d(u, v) = \rho^{1/2}(u, v)$ as well.

Choosing $\alpha = 2$ and $\gamma = 2$, we will prove that $\forall\, u, v \in [0, 1]^2$

$$\mathbb{E}\,\big|W_n(u) - W_n(v)\big|^{2} \lesssim d^{2}(u, v) \,, \tag{C.35}$$

which will complete the proof.

Towards that end, consider the case $a < c < d < b$. We will show that (C.35) holds in this case, and similar arguments will do the job for other cases.

Observe that

$$\begin{aligned}
W_n(u) \,-\, W_n(v) \;&=\; \frac{\sqrt{2}}{n\sqrt{V_0}}\, S_n(a,b) \,-\, \frac{\sqrt{2}}{n\sqrt{V_0}}\, S_n(c,d) \\
&=\; \frac{\sqrt{2}}{n\sqrt{V_0}} \Big[\sum_{i=\lfloor na\rfloor+2}^{\lfloor nb\rfloor} \sum_{j=\lfloor na\rfloor+1}^{i-1} H(X_i,X_j) \,-\, \sum_{i=\lfloor nc\rfloor+2}^{\lfloor nd\rfloor} \sum_{j=\lfloor nc\rfloor+1}^{i-1} H(X_i,X_j) \Big] \\
&=\; \frac{\sqrt{2}}{n\sqrt{V_0}} \Big[\sum_{i=\lfloor na\rfloor+2}^{\lfloor nc\rfloor} \sum_{j=\lfloor na\rfloor+1}^{i-1} H(X_i,X_j) \,+\, \sum_{i=\lfloor nc\rfloor+1}^{\lfloor nd\rfloor} \sum_{j=\lfloor na\rfloor+1}^{\lfloor nc\rfloor} H(X_i,X_j) \\
&\qquad \sum_{i=\lfloor nd\rfloor+1}^{\lfloor nb\rfloor} \sum_{j=\lfloor na\rfloor+1}^{\lfloor nc\rfloor} H(X_i,X_j) \,+\, \sum_{i=\lfloor nd\rfloor+1}^{\lfloor nb\rfloor} \sum_{j=\lfloor nc\rfloor+1}^{\lfloor nd\rfloor} H(X_i,X_j) \\
&\qquad \sum_{i=\lfloor nd\rfloor+2}^{\lfloor nb\rfloor} \sum_{j=\lfloor nd\rfloor+1}^{i-1} H(X_i,X_j) \Big] \\
&=:\; I \,+\, II \,+\, III \,+\, IV \,+\, V\,.
\end{aligned} \tag{C.36}$$

By power mean inequality,

$$(I \,+\, II \,+\, III \,+\, IV \,+\, V)^2 \;\lesssim\; I^2 \,+\, II^2 \,+\, III^2 \,+\, IV^2 \,+\, V^2\,. \tag{C.37}$$

Now

$$\mathbb{E}\,(I^2) \;=\; \frac{2}{n^2 V_0} \sum_{i_1,i_2=\lfloor na\rfloor+2}^{\lfloor nc\rfloor} \sum_{j_1=\lfloor na\rfloor+1}^{i_1-1} \sum_{j_2=\lfloor na\rfloor+1}^{i_2-1} \mathbb{E}\left[H(X_{i_1},X_{j_1})\, H(X_{i_2},X_{j_2})\right].$$

Clearly $\mathbb{E}\left[H(X_{i_1},X_{j_1})\, H(X_{i_1},X_{j_1})\right] = 0$ if the cardinality of the set $\{i_1,j_1\} \cap \{i_2,j_2\}$ is 0 or 1. Therefore we have

$$\begin{aligned}
\mathbb{E}\,(I^2) \;&=\; \frac{2}{n^2 V_0} \sum_{i=\lfloor na\rfloor+2}^{\lfloor nc\rfloor} \sum_{j=\lfloor na\rfloor+1}^{i-1} \mathbb{E}\left[H^2(X_i,X_j)\right] \;=\; \frac{2}{n^2 V_0} \sum_{\lfloor na\rfloor+1 \le j < i \le \lfloor nc\rfloor} V_0 \\
&=\; \frac{1}{n^2} \big(\lfloor nc\rfloor - \lfloor na\rfloor\big) \big(\lfloor nc\rfloor - \lfloor na\rfloor - 1\big)\,.
\end{aligned} \tag{C.38}$$

Note that

$$\begin{aligned} \lfloor nc \rfloor - \lfloor na \rfloor - 1 \;&\leq\; nc - na + na - \lfloor na \rfloor - 1 \\ &=\; n(c-a) \,+\, (\{na\} - 1) \\ &\leq\; n(c-a)\,, \end{aligned} \tag{C.39}$$

as $\{na\} \leq 1$. Therefore we have from (C.38) and (C.39)

$$\mathbb{E}\,(I^2) \;\lesssim\; c - a\,. \tag{C.40}$$

Likewise it can be shown that

$$\mathbb{E}\,(V^2) \;\lesssim\; b - d\,. \tag{C.41}$$

Now

$$\begin{aligned} \mathbb{E}\,(II^2) \;&=\; \frac{2}{n^2 V_0} \sum_{i_1,i_2=\lfloor nc \rfloor+1}^{\lfloor nd \rfloor} \; \sum_{j_1,j_2=\lfloor na \rfloor+1}^{\lfloor nc \rfloor} \mathbb{E}\left[H(X_{i_1},X_{j_1})\,H(X_{i_2},X_{j_2})\right] \\ &=\; \frac{2}{n^2 V_0} \sum_{i=\lfloor nc \rfloor+1}^{\lfloor nd \rfloor} \; \sum_{j=\lfloor na \rfloor+1}^{\lfloor nc \rfloor} \mathbb{E}\left[H^2(X_i,X_j)\right] \\ &=\; \frac{2}{n^2}\big(\lfloor nd \rfloor - \lfloor nc \rfloor\big)\big(\lfloor nc \rfloor - \lfloor na \rfloor\big) \\ &\lesssim\; \frac{1}{n}\left[n(c-a)+1\right] \\ &\lesssim\; c - a\,. \end{aligned} \tag{C.42}$$

Similarly it can be shown that

$$\mathbb{E}\,(III^2) \;\lesssim\; c - a \qquad \text{and} \qquad \mathbb{E}\,(IV^2) \;\lesssim\; b - d\,. \tag{C.43}$$

Combining (C.40)-(C.43) with (C.36) and (C.37), we get

$$\mathbb{E}\left|W_n(u) \,-\, W_n(v)\right|^2 \;\lesssim\; (c-a) \,+\, (b-d) \;=\; \|u-v\|_1 \;=\; d^2(u,v)\,.$$

This proves (C.35) and thereby completes the proof of T2.

$\diamondsuit$

This completes the proof of Theorem 18.

$\diamondsuit$

Proof of Theorem 19. Again consider the subset $[0,1]^2 \subseteq \mathbb{R}^2$ equipped with the metric $d(u,v) = \|u-v\|_1^{1/2}$ for $u,v \in [0,1]^2$. By Theorem 1 in Andrews (1992), we essentially need to show :

A1. $[0,1]^2$ is totally bounded with respect to the metric d.

A2. Pointwise convergence : $G_n(u) \overset{P}{\to} 0 \;\; \forall\, u \in [0,1]^2$ as $n,p \to \infty$.

A3. Asymptotic stochastic equicontinuity : for any $\epsilon > 0$,

$$\lim_{\delta\downarrow 0}\,\limsup_{n,p\to\infty}\, P\Big(\sup_{\substack{u,v\,\in\,[0,1]^2\\ d(u,v)\le\delta}} \left|G_n(u) \,-\, G_n(v)\right|\Big) \;=\; 0\,.$$

A1 is easy to see. $[0,1]^2$ is compact with respect to the metric d, and therefore totally bounded. To see A2, note that for fixed $u \in [0,1]^2$, using Chebyshev's inequality we have for any $\epsilon > 0$

$$P\big(|G_n(u)| > \epsilon\big) \;\le\; \frac{1}{\epsilon^2}\,\mathbb{E}\,G_n^2(u) \;=\; \frac{1}{n^2\epsilon^2 V_0}\,\mathbb{E}\,R_n^2(a,b)\,. \tag{C.44}$$

Recalling that $R_n(a,b) = \widetilde{R}_n(\lfloor an\rfloor+1, \lfloor bn\rfloor)$ and the definition of $\widetilde{R}_n(k,m)$, it is not hard to verify that

$$R_n^2(a,b) \;=\; \sum_{\lfloor na\rfloor+1\le i_1<i_2\le\lfloor nb\rfloor}\;\sum_{\lfloor na\rfloor+1\le i_1'<i_2'\le\lfloor nb\rfloor} \tau^2\, R(X_{i_1},X_{i_2})\, R(X_{i_1'},X_{i_2'})\,.$$

Therefore by Hölder's inequality, we have

$$\begin{aligned}\mathbb{E}\, R_n^2(a,b) \;&\le\; \sum_{\lfloor na\rfloor+1\le i_1<i_2\le\lfloor nb\rfloor}\;\sum_{\lfloor na\rfloor+1\le i_1'<i_2'\le\lfloor nb\rfloor} \tau^2 \left(\mathbb{E}\, R^2(X_{i_1},X_{i_2})\right)^{1/2}\left(\mathbb{E}\, R^2(X_{i_1'},X_{i_2'})\right)^{1/2}\\ &=\; \tau^2\left[\frac{1}{2}\left(\lfloor nb\rfloor-\lfloor na\rfloor\right)\left(\lfloor nb\rfloor-\lfloor na\rfloor-1\right)\left(\mathbb{E}\, R^2(X,X')\right)^{1/2}\right]^2\\ &=\; O\Big(n^4\,\tau^2\;\mathbb{E}\, R^2(X,X')\Big)\\ &=\; O\Big(n^4\left[\tau^4\;\mathbb{E}\, R^4(X,X')\right]^{1/2}\Big)\,.\end{aligned} \tag{C.45}$$

Combining (C.44) and (C.45), we get

$$\begin{aligned}P\big(|G_n(u)|>\epsilon\big) \;&=\; O\Big(\frac{n^2}{\epsilon^2\;\mathbb{E}\, H^2(X,X')}\left[\tau^4\;\mathbb{E}\, R^4(X,X')\right]^{1/2}\Big)\\ &=\; O\Big(\frac{1}{\epsilon^2}\left[\frac{n^4\,\tau^4\;\mathbb{E}\, R^4(X,X')}{\left(\mathbb{E}\, H^2(X,X')\right)^2}\right]^{1/2}\Big)\,.\end{aligned} \tag{C.46}$$

Under Assumption 4.3.3, it is easy to see from (C.46) that

$$P\big(|G_n(u)|>\epsilon\big) \;=\; o(1)\,,$$

which implies $G_n(u) \xrightarrow{P} 0$ for any fixed $u\in[0,1]^2$ as $n,p\to\infty$. This proves A2.

Finally to prove A3, again by Theorem A.8 in Li and Racine (2007) it will suffice to show that $\forall\, u,v\in[0,1]^2$

$$\mathbb{E}\left|G_n(u)\;-\;G_n(v)\right|^2 \;\lesssim\; d^2(u,v)\,. \tag{C.47}$$

Similar to the proof of T2 before in the proof of Theorem 18, we will show that (C.47) holds in the case $a<c<d<b$. Similar arguments can prove (C.47) for other cases.

Similar to the proof of T2, now we have

$$
\begin{aligned}
G_n(u) \,-\, G_n(v) \;&=\; \frac{1}{n\sqrt{V_0}}\, R_n(a,b) \,-\, \frac{1}{n\sqrt{V_0}}\, R_n(c,d) \\
&=\; \frac{\tau}{n\sqrt{V_0}} \Big[\sum_{i=\lfloor na \rfloor+2}^{\lfloor nb \rfloor} \sum_{j=\lfloor na \rfloor+1}^{i-1} R(X_i, X_j) \,-\, \sum_{i=\lfloor nc \rfloor+2}^{\lfloor nd \rfloor} \sum_{j=\lfloor nc \rfloor+1}^{i-1} R(X_i, X_j) \Big] \\
&=\; \frac{\tau}{n\sqrt{V_0}} \Big[\sum_{i=\lfloor na \rfloor+2}^{\lfloor nc \rfloor} \sum_{j=\lfloor na \rfloor+1}^{i-1} R(X_i, X_j) \,+\, \sum_{i=\lfloor nc \rfloor+1}^{\lfloor nd \rfloor} \sum_{j=\lfloor na \rfloor+1}^{\lfloor nc \rfloor} R(X_i, X_j) \\
&\qquad\quad \sum_{i=\lfloor nd \rfloor+1}^{\lfloor nb \rfloor} \sum_{j=\lfloor na \rfloor+1}^{\lfloor nc \rfloor} R(X_i, X_j) \,+\, \sum_{i=\lfloor nd \rfloor+1}^{\lfloor nb \rfloor} \sum_{j=\lfloor nc \rfloor+1}^{\lfloor nd \rfloor} R(X_i, X_j) \\
&\qquad\quad \sum_{i=\lfloor nd \rfloor+2}^{\lfloor nb \rfloor} \sum_{j=\lfloor nd \rfloor+1}^{i-1} R(X_i, X_j) \Big] \\
&=:\; I_G \,+\, II_G \,+\, III_G \,+\, IV_G \,+\, V_G\,.
\end{aligned}
\tag{C.48}
$$

By power mean inequality,

$$
(I_G \,+\, II_G \,+\, III_G \,+\, IV_G \,+\, V_G)^2 \;\lesssim\; I_G^2 \,+\, II_G^2 \,+\, III_G^2 \,+\, IV_G^2 \,+\, V_G^2\,. \tag{C.49}
$$

Now

$$
\mathbb{E}\,(I_G^2) \;=\; \frac{\tau^2}{n^2 V_0} \sum_{i_1, i_2=\lfloor na \rfloor+2}^{\lfloor nc \rfloor} \sum_{j_1=\lfloor na \rfloor+1}^{i_1-1} \sum_{j_2=\lfloor na \rfloor+1}^{i_2-1} \mathbb{E}\left[R(X_{i_1}, X_{j_1})\, R(X_{i_2}, X_{j_2})\right]. \tag{C.50}
$$

Again using Hölder's inequality and similar arguments as used in deriving (C.45), we get from (C.50)

$$
\begin{aligned}
\mathbb{E}\,(I_G^2) \;&=\; \frac{\tau^2}{n^2 V_0} \Big(\sum_{i=\lfloor na \rfloor+2}^{\lfloor nc \rfloor} \sum_{j=\lfloor na \rfloor+1}^{i-1} \left(\mathbb{E}\, R^2(X_i, X_j)\right)^{1/2} \Big)^2 \\
&=\; \frac{\tau^2}{n^2 V_0}\; \frac{\left(\lfloor nc \rfloor - \lfloor na \rfloor\right)^2 \left(\lfloor nc \rfloor - \lfloor na \rfloor - 1\right)^2}{4\, n^2}\; n^2\, \mathbb{E}\, R^2(X, X').
\end{aligned}
\tag{C.51}
$$

Using the fact that $\lfloor nc \rfloor - \lfloor na \rfloor - 1 \leq n(c-a)$, $(c-a)^2 \leq (c-a)$ and Hölder's inequality, we get from (C.51)

$$\begin{aligned} \mathbb{E}\,(I_G^2) \;&\lesssim\; (c-a)\left(\frac{n^2\,\tau^2\;\mathbb{E}\,R^2(X,X')}{\mathbb{E}\,H^2(X,X')}\right) \;\leq\; (c-a)\left(\frac{n^2\,\tau^2\;\Big(\mathbb{E}\,R^4(X,X')\Big)^{1/2}}{\mathbb{E}\,H^2(X,X')}\right) \\ &\leq\; (c-a)\left(\frac{n^4\,\tau^4\;\mathbb{E}\,R^4(X,X')}{\Big[\mathbb{E}\,H^2(X,X')\Big]^2}\right)^{1/2}. \end{aligned} \tag{C.52}$$

Under Assumption 4.3.3, $\frac{n^4\,\tau^4\;\mathbb{E}\,R^4(X,X')}{\Big[\mathbb{E}\,H^2(X,X')\Big]^2} = o(1)$ as $n, p \to \infty$, and hence $\frac{n^4\,\tau^4\;\mathbb{E}\,R^4(X,X')}{\Big[\mathbb{E}\,H^2(X,X')\Big]^2}$ must be a bounded sequence in n and p. Therefore we have from (C.52)

$$\mathbb{E}\,(I_G^2) \;\lesssim\; (c-a)\,. \tag{C.53}$$

Likewise it can be shown that

$$\mathbb{E}\,(II_G^2) \;\lesssim\; (c-a)\;,\; \mathbb{E}\,(III_G^2) \;\lesssim\; (c-a)\;,\; \mathbb{E}\,(IV_G^2) \;\lesssim\; (b-d)\;,\; \mathbb{E}\,(V_G^2) \;\lesssim\; (b-d)\,. \tag{C.54}$$

Combining (C.53)-(C.54) with (C.48) and (C.49), we get

$$\mathbb{E}\left|G_n(u) \,-\, G_n(v)\right|^2 \;\lesssim\; (c-a) \,+\, (b-d) \;=\; \|u-v\|_1 \;=\; d^2(u,v)\,.$$

This proves (C.47) and thereby completes the proof of A3 and hence the theorem.

◇