

STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL DATA

BY

CHANGBO ZHU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Doctoral Committee:

Professor Xiaofeng Shao, Chair and Director of Research
Associate Professor Xiaohui Chen
Assistant Professor Geogios Fellouris
Professor Emeritus John I. Marden

Abstract

Statistical inference is a procedure of using collected observations to deduce properties of the underlying data generating process. In this thesis, we investigate three important problems in high-dimensional statistics and develop some new methods and theory, which show the limitation of some existing approaches and motivate the use of our proposed methods.

In the first chapter, we study distance covariance, Hilbert-Schmidt covariance (aka Hilbert-Schmidt independence criterion [Gretton et al. (2008)]) and related independence tests under the high dimensional scenario. We show that the sample distance/Hilbert-Schmidt covariance between two random vectors can be approximated by the sum of squared componentwise sample cross-covariances up to an asymptotically constant factor, which indicates that the distance/Hilbert-Schmidt covariance based test can only capture linear dependence in high dimension. Under the assumption that the components within each high-dimensional vector are weakly dependent, the distance correlation based t test developed by Székely and Rizzo (2013) for independence is shown to have trivial limiting power when the two random vectors are nonlinearly dependent but component-wisely uncorrelated. This new and surprising phenomenon, which seems to be discovered for the first time, is further confirmed in our simulation study. As a remedy, we propose tests based on an aggregation of marginal sample distance/Hilbert-Schmidt covariances and show their superior power behavior against their joint counterparts in simulations. We further extend the distance correlation based t test to those based on Hilbert-Schmidt covariance and marginal distance/Hilbert-Schmidt covariance. A novel unified approach is developed to analyze the studentized sample distance/Hilbert-Schmidt covariance as well as the studentized sample marginal distance covariance under both null and alternative hypothesis. Our theoretical and simulation results shed light on the limitation of distance/Hilbert-Schmidt covariance when used jointly in the high dimensional setting and suggest the aggregation of marginal distance/Hilbert-Schmidt covariance as a useful alternative.

In the second chapter, we study a class of two sample test statistics based on inter-point distances in the high dimensional and low/medium sample size setting. Our test statistics include the well-known energy distance and maximum mean discrepancy with Gaussian and Laplacian kernels, and the critical values are obtained via permutations. We show that all these tests are inconsistent when the two high dimensional distributions correspond to the same marginal distributions but differ in other aspects of the distributions. The tests based on energy distance and maximum mean discrepancy mainly target the differences between marginal means and variances, whereas the test based on L^1 -distance can capture the difference in marginal distributions. Our theory sheds new light on the limitation of inter-point distance based tests, the impact of different distance metrics, and the behavior of permutation tests in high dimension. Some simulation results and a real data illustration are also presented to corroborate our theoretical findings.

In the third chapter, we propose a new methodology for change point detection of a high-dimensional time series. We extend the U -statistic based approach of Wang et al. (2019) by applying the trimming

technique and utilizing the self-normalization principle. Under the fixed- b asymptotics, where we fix the proportion of trimming parameter over the sample size, we derive the limiting distributions of our test statistic under both the null and local alternatives of a single mean change. Furthermore, we combine our test statistic with the wild binary segmentation procedure to perform the change-point estimation. Empirical simulations demonstrate that the trimming technique is effective and necessary for both testing and estimation when there is strong temporal dependence. As an important theoretical contribution, we derive the weak convergence of the U -statistic based processes for high-dimensional linear process and show the applicability of BN decomposition in high dimension.

To my wife, Shan Lu.

Acknowledgments

I own my appreciation to a number of people for their help, guidance and contributions in completing this thesis successfully. Foremost, I would like to express my deep and sincere gratitude to my advisor, Professor Xiaofeng Shao, for his endless support and priceless edification. He is not only a dedicated scholar but also a patient educator. His broad knowledge and passionate research attitude have always been my motives to push myself to the next level. I am lucky and honored to be his student.

I also want to extend my gratitude to my committee members, Professor Xiaohui Chen, Professor Geogios Fellouris and Professor John I. Marden for supporting my work. They are great educators, who can be amiable while going straight to the heart of the matter. Their kindness has made me feel more confident in presenting my research results and their sharp questions can always incite my critical thinking.

Special thanks to my collaborators Professor Xianyang Zhang and Professor Stanislav Volgushev for their detailed guidance and constructive comments. The research discussions and the regular meetings with them are the best lessons I can ever get. Without their help, many of my research work would be delayed, and moreover, the quality of these work would be compromised.

In addition, I want to thank my academic brothers and sisters Chung Eun Lee, Shun Yao, Runmin Wang, Teng Wu and Yangfan Zhang for their timely help and cheerful encouragement. They share their research and learning experience with me generously. Many strategies in doing research and learning statistics that I get from them turn out to be extremely useful.

Life as a graduate student can be difficult. I am grateful to have Yujia Deng, Wei Han, Trevor Austin Harris, Sarah Elizabeth Formentini, Danielle Kaye Sass, Yihe Wang, Yan, Liu, Yubai Yuan as my peers. It is a wonderful and enjoyable experience to learn and grow with them. Besides, I could never forget the joy and laughter we had at the occasional parties. Also, I want to thank my friends for their unconditional help and companionship. Thank you to Jialin Song for teaching me how to drive. Thank you to Linlin Liu for giving me a hand in times of need. Thank you to Yufan Jiang for being a great listener.

Then, I would love to express my gratitude to my parents, Jinfen Liu and Zhongxia Zhu, for their selfish love and countless cares. Their encouragement can always help me regain confidence and strength. Even though I have been studying outside my hometown for more than ten years to chase my dream, I can never stop myself from missing my parents. They are the warmth in my heart and the source of my strength.

Last but not least, I want to thank my wife, Shan Lu, as it is her love that made me who I am.

Table of Contents

List of Tables	viii
List of Figures	ix
Chapter 1 Distance-based and RKHS-based Dependence Metrics in High Dimension . .	1
1.1 Introduction	1
1.1.1 Notations	4
1.2 High Dimension Low Sample Size	5
1.2.1 Distance Covariance and Variants	5
1.2.2 Studentized Test Statistics	11
1.3 Numerical Results	17
1.4 Conclusion	22
1.5 High Dimension Medium Sample Size	22
1.5.1 Distance Covariance and Variants	23
1.5.2 Studentized Test Statistics	24
1.6 Additional Simulation Examples and Comparisons	26
1.7 Technical Details	32
1.7.1 Proof of Proposition 2	32
1.7.2 Proof of Remark 4	33
1.7.3 Proof of Theorem 1	33
1.7.4 Proof of Theorem 2	35
1.7.5 Proof of Proposition 6	37
1.7.6 Proof of Theorem 3	37
1.7.7 Proof of Proposition 10	41
1.7.8 Proof of Proposition 12	42
1.7.9 Proof of Proposition 14	47
1.7.10 Proof of Proposition 13	47
1.7.11 Proof of Corollary 1	49
1.7.12 Proof of Remark 17	49
1.7.13 Proof of Theorem 4	50
1.7.14 Proof of Theorem 5	52
1.7.15 Proof of Remark 20	53
1.7.16 Proof of Theorem 6	55
1.7.17 Proof of Proposition 21	56
1.7.18 Proof of Proposition 22	57
1.7.19 Proof of Corollary 2	57
Chapter 2 Interpoint Distance Based Two Sample Tests in High Dimension	58
2.1 Introduction	58
2.1.1 Notation	60
2.2 Interpoint Distance Based Two Sample Tests	60
2.3 Power Analysis for Permutation Test	61

2.3.1	Local Alternatives	63
2.3.2	High Dimensional Low Sample Size (HDLSS)	65
2.3.3	High Dimensional Medium Sample Size (HDMSS)	69
2.4	Numerical Studies	72
2.4.1	Performance on simulated data	72
2.4.2	Performance on real data	75
2.5	Discussions & Conclusion	76
2.6	Technical Details	78
2.6.1	Proof of Sufficient Conditions for Local Alternatives	78
2.6.2	Proof of Theorem 7	79
2.6.3	Proof of Proposition 35	81
2.6.4	Proof of Theorem 8	82
2.6.5	Proof of Theorem 9	84
2.6.6	Proof of Theorem 10	90
2.6.7	Proof of Corollary 3	93
2.6.8	Proof of Theorem 11	93
Chapter 3	Change Point Detection for High-dimensional Time Series	94
3.1	Introduction	94
3.2	Change Point Detection for High-dimensional Time Series	95
3.2.1	Review of Wang et al. (2019)	95
3.2.2	Trimmed U -statistic	96
3.2.3	Wild Binary Segmentation	96
3.3	Asymptotic Theory	97
3.4	Simulation Study	101
3.4.1	Size and Power for Change Point Testing	101
3.4.2	Change Point Estimation	101
3.5	Conclusion	104
3.6	Technical Details	104
3.6.1	Properties of Linear Process	104
3.6.2	Proof of Theorem 12	105
3.6.3	Proof of Theorem 13	113
3.6.4	Proof of Auxiliary Lemmas	114
3.6.5	Proof of Lemma 55	118
References	124

List of Tables

1.1	Size comparison from Example 1	18
1.2	Power comparison under H_{A_s} from Example 2	20
1.3	Power comparison under H_{A_s} from Example 3	21
1.4	Power Comparison on Earthquake data	22
1.5	Size comparison from Example 5	27
1.6	Power comparison from Example 6	29
1.7	Size comparison from Example 1	30
1.8	Power comparison from Example 6	31
1.9	Power comparison from Example 2	31
1.10	Power comparison from Example 3	32
1.11	Power Comparison on Earthquake data	32
2.1	Correspondence between different choices of ψ, φ and existing distance metrics as well as two sample test statistics in the literature.	61
2.2	Characterization of H_{A_c} for some specific metrics.	64
2.3	Sufficient conditions for H_{A_t} with respect to some specific metrics.	64
2.4	Sufficient conditions for H_{A_t} with respect to some specific metrics.	65
2.5	Size comparison from Example 7 for $p = 500$	72
3.1	Simulated critical values of T_n	99
3.2	Simulation results for Example 11 with $p = 2n$ and significance level 0.05.	102
3.3	Simulation results for Example 12 with $n = 250$ and significance level 0.05. WBS ¹ and WBS ² are with respect to trimming $\eta = 0.01$ and $\eta = 0.02$ respectively.	103

List of Figures

2.1	Illustration of the Permutation Procedure.	62
2.2	Power comparison for example 8 and $n = 70, m = 30, p = 500$, where in the top 3 figures Z_1, Z_2 are generated from normal distribution and in the bottom 3 figures, Z_1, Z_2 are generated from exponential distribution.	73
2.3	Power comparison for Example 9 and $n = 70, m = 30$. For the top two figures, the dimension p is equal to 500 and we plot the power as β ranges from 0 to 1. For the bottom two figures, β is fixed to be 0.5 and the power is plotted with respect to p	75
2.4	Power comparison for Example 10 and $n = 70, m = 30$. For the top two figures, the dimension p is equal to 500 and we plot the power as β ranges from 0 to 1. For the bottom two figures, β is fixed to be 1 and the power is plotted with respect to p	76
2.5	A glance of the data in Section 2.4.2, where we plot one point from each of the two classes for each data set.	77
2.6	Power comparison for real data examples in Section 2.4.2.	77
3.1	Illustration of $Q_{X_n}^u(w_1, w_2; h_1, h_2)$ (green region) and $\tilde{S}_{X_n}^u(w_1, h_1 - 1 \tau)$ (yellow region)	107

Chapter 1

Distance-based and RKHS-based Dependence Metrics in High Dimension

1.1 Introduction

Testing for independence between two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ is a fundamental problem in statistics. There is a huge literature in the low dimensional context. Here we mention rank correlation coefficients based tests and nonparametric Cramér-von Mises type statistics in Hoeffding (1948), Blum et al. (1961), De Wet (1980); tests based on signs or empirical characteristic functions, see Sinha and Wieand (1977), Deheuvels (1981), Csörgő (1985), Hettmansperger and Oja (1994), Gieser and Randles (1997), Taskinen et al. (2003) among others; tests based on recently developed nonlinear dependence metrics that target at non-linear and non-monotone dependence include distance covariance [Székely et al. (2007)], Hilbert-Schmidt independence criterion (HSIC) [Gretton et al. (2008)] (aka Hilbert-Schmidt covariance in this work) and sign covariance [Bergsma and Dassios (2014)]. Also see Berrett and Samworth (2019) for some recent work on independence testing via mutual information.

In the high dimensional setting, the literature is scarce. Székely and Rizzo (2013) extended the distance correlation proposed in Székely et al. (2007) to the problem of testing independence of two random vectors under the setting that the dimensions p and q grow while sample size n is fixed. This setting is known as high dimension, low sample size (HDLSS) in the literature and has been adopted in Hall et al. (2005), Ahn et al. (2007), Jung and Marron (2009), and Wei et al. (2016) etc. A closely related asymptotic framework is the high dimension medium sample size (HDMSS) [Aoshima et al. (2018)], where $n \wedge p \wedge q \rightarrow \infty$ with p, q growing more rapidly. Among the recent work that is related to independence testing in the high dimensional setting, Pan et al. (2014) proposed tests of independence among a large number of high dimensional random vectors using insights from random matrix theory; Yang and Pan (2015) proposed a new statistic based on the sum of regularized sample canonical correlation coefficients of X and Y , which is limited to testing for uncorrelatedness due to the use of canonical correlation. Leung and Drton (2018) proposed to test for mutual independence of high dimensional vectors using sum of pairwise rank correlations and sign covariances; Yao et al. (2018) addressed the mutual independence testing problem in the high dimensional context by using sum of pairwise squared sample distance covariances; Zhang et al. (2018) proposed a L^2 type test for conditional mean/quantile dependence of a univariate response variable given a high dimensional covariate vector based on martingale difference divergence [Shao and Zhang (2014)], which is an extension of distance covariance to quantify (conditional) mean dependence.

Distance covariance/correlation was first introduced in Székely et al. (2007) and has received much attention since then. Owing to its notable ability to quantify any types of dependence including non-monotone, non-linear dependence and also the flexibility to be applicable to two random vectors in arbitrary, not necessarily equal dimensions, a lot of research work has been done to extend and apply distance covariance into many modern statistical problems; see e.g. Kong et al. (2012), Li et al. (2012), Zhou (2012), Lyons

(2013), Székely and Rizzo (2014), Dueck et al. (2014), Shao and Zhang (2014), Park et al. (2015), Matteson and Tsay (2017), Zhang et al. (2018), Edelmann et al. (2017), Yao et al. (2018) among others. In this paper, we shall revisit the test proposed by Székely and Rizzo (2013), which seems to be the only test in the high dimensional setting that captures nonlinear and nonmonotonic dependence. Unlike the positive finding reported in Székely and Rizzo (2013), we obtained some negative and shocking results that show the limitation of distance covariance/correlation in the high dimensional context.

Specifically, we show that for two random vectors $X = (x_1, \dots, x_p)^T \in \mathbb{R}^p$ and $Y = (y_1, \dots, y_q)^T \in \mathbb{R}^q$ with finite component-wise second moments, as $p, q \rightarrow \infty$ and n can either be fixed or grows to infinity at a slower rate,

$$dCov_n^2(\mathbf{X}, \mathbf{Y}) \approx \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j), \quad (1.1)$$

where $X_k \stackrel{d}{=} X$ and $Y_k \stackrel{d}{=} Y$ are independent samples, \mathcal{X}_i and \mathcal{Y}_j are the component-wise samples, $\mathbf{X} = (X_1, X_2, \dots, X_n)^T = (\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_p)$ and $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T = (\mathcal{Y}_1, \mathcal{Y}_2, \dots, \mathcal{Y}_q)$ denote the sample matrices, $dCov_n^2(\mathbf{X}, \mathbf{Y})$ is the unbiased sample distance covariance, τ is a constant quantity depending on the marginal distributions of X and Y as well as p and q , $cov_n^2(\mathcal{X}_i, \mathcal{Y}_j)$ is an unbiased sample estimate of $cov^2(x_i, y_j)$ to be defined later. To the best of our knowledge, this is the first work in the literature uncovering the connection between sample distance covariance and sample covariance, the latter of which can only measure the linear dependence between two random variables. This approximation suggests that the distance covariance can only measure linear dependence in the high dimensional setting although it is well-known to be capable of capturing non-linear dependence in the fixed dimensional case.

Gretton et al. (2008) proposed Hilbert-Schmidt independence criterion (aka Hilbert-Schmidt covariance in this paper), which can be seen as a generalization of distance covariance by kernelizing the L^2 distance as shown by Sejdinovic et al. (2013). Despite the kernelization process, we show that the Hilbert-Schmidt covariance ($hCov$) enjoys similar approximation property under high dimension low/medium sample size setting, i.e.

$$hCov_n^2(\mathbf{X}, \mathbf{Y}) \approx A_p B_q \times \frac{1}{\tau^2} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j), \quad (1.2)$$

where $hCov_n^2(\mathbf{X}, \mathbf{Y})$ is the unbiased sample Hilbert-Schmidt covariance, A_p and B_q both converge in probability to constants that depend on the pre-chosen kernels. This approximation also suggests that when the dimension is high, the Hilbert-Schmidt covariance ($hCov$) applied to the whole components of the vectors also exhibits the loss of power when X and Y are non-linearly dependent, but component-wisely uncorrelated or weakly correlated.

As a natural remedy, we propose a distance covariance based marginal test statistic, i.e.,

$$mdCov_n^2(\mathbf{X}, \mathbf{Y}) = \sqrt{\binom{n}{2}} \sum_{i=1}^p \sum_{j=1}^q dCov_n^2(\mathcal{X}_i, \mathcal{Y}_j).$$

This test statistic is an aggregate of the componentwise sample distance covariances and captures the component by component nonlinear dependence. Similarly, the marginal Hilbert-Schmidt covariance ($mhCov$)

is defined as

$$mhCov_n^2(\mathbf{X}, \mathbf{Y}) = \sqrt{\binom{n}{2}} \sum_{i=1}^p \sum_{j=1}^q hCov_n^2(\mathcal{X}_i, \mathcal{Y}_j).$$

The distance covariance, Hilbert-Schmidt covariance, marginal distance covariance and marginal Hilbert-Schmidt covariance based tests can be carried out by standard permutation procedures. The superiority of $mdCov$ and $mhCov$ based tests over its joint counterparts in power is demonstrated in the simulation studies. On the other hand, Székely and Rizzo (2013) discussed the distance correlation ($dCor$) based t -test under HDLSS and derived the limiting null distribution of the test statistic under suitable assumptions. We consider the same t -test statistic and further extend to Hilbert-Schmidt covariance ($hCov$), marginal distance covariance ($mdCov$) and marginal Hilbert-Schmidt covariance ($mhCov$). To derive the asymptotic distribution of studentized version of $dCov$, $hCov$, $mdCov$ and $mhCov$ under both the null of independence (for HDLSS and HDMSS setting) and some specific alternative classes (for HDLSS setting), we develop a novel unified approach. In particular, we define a unified quantity ($uCov$) based on the bivariate kernel k and show that under HDLSS setting, properly scaled $dCov_n^2$, $hCov_n^2$ and $mdCov_n^2$ are all asymptotically equal to $uCov_n^2$ up to different choices of kernels, i.e.

$$\left. \begin{aligned} dCov_n^2(\mathbf{X}, \mathbf{Y}) &\approx a \times uCov_n^2(\mathbf{X}, \mathbf{Y}) \\ hCov_n^2(\mathbf{X}, \mathbf{Y}) &\approx A_p B_q \times uCov_n^2(\mathbf{X}, \mathbf{Y}) \end{aligned} \right\} \quad \text{when } k(x, y) = |x - y|^2, \quad (1.3)$$

$$mdCov_n^2(\mathbf{X}, \mathbf{Y}) = b \times uCov_n^2(\mathbf{X}, \mathbf{Y}) \quad \left. \right\} \quad \text{when } k(x, y) = |x - y|,$$

where a, b are constants and A_p, B_p both converge in probability to constants. Next, we show that

$$\left\{ \begin{aligned} uCov_n^2(\mathbf{X}, \mathbf{Y}) &\xrightarrow{d} \frac{2}{n(n-3)} \mathbf{c}^T \mathbf{M} \mathbf{d}, \quad \text{under HDLSS,} \\ C_{n,p,q} uCov_n^2(\mathbf{X}, \mathbf{Y}) &\xrightarrow{d} N(0, 1), \quad \text{under HDMSS,} \end{aligned} \right.$$

where \mathbf{c}, \mathbf{d} are jointly Gaussian, \mathbf{M} is a projection matrix and $C_{n,p,q}$ is a normalizing constant. Thus, we can easily apply the above results to $dCov$, $hCov$ and $mdCov$ -based t -test statistics using (1.3). The unified approach still works for $mhCov$ -based t -test if we consider the bandwidth parameters appeared in the kernel distance to be fixed constants. However, we encounter technical difficulties if the bandwidth parameters along each dimension depends on the whole component-wise samples, since this makes the pair-wise sample distance correlated with each other and complicates the asymptotic analysis.

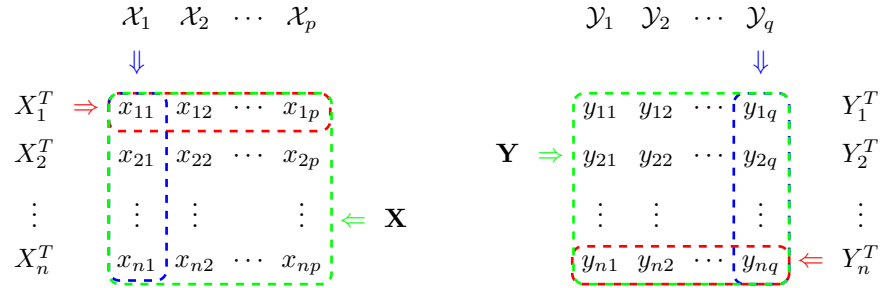
We obtain the same limiting null distribution as Székely and Rizzo (2013) and further show that this test statistic has a trivial power against the alternative where X and Y are non-linearly dependent, but component-wisely uncorrelated. This clearly demonstrates that the distance covariance/correlation based joint independence test (i.e., treating all components of a vector as a whole jointly) fails to capture the non-linear dependence in high dimension. This phenomenon is new and was not reported in Székely and Rizzo (2013). It shows that there might be some intrinsic difficulties for distance covariance to capture the non-linear dependence when the dimension is high and provide a cautionary note on the use of distance covariance/correlation directly to the whole components of high dimensional data. Besides, we have the following additional contributions relative to Székely and Rizzo (2013): (i) we relax the component-wise i.i.d. assumption used for asymptotic analysis; (ii) the limiting distributions are derived under both the null and certain classes of alternative hypothesis for the HDLSS framework; (iii) our unified approach holds for any bivariate kernel that has continuous second order derivative in a neighborhood containing 1; (iv)

our approach is built on some new technical arguments which reveal some insights on \mathcal{U} -centering; (v) the limiting null distribution is also derived under the HDMSS setting. It is worth noting that the phenomenon of decreasing power with higher dimension for Hilbert-Schmidt covariance (with Gaussian/Laplacian kernels) based independence test has been observed in Ramdas et al. (2015), but they did not provide a complete theoretical explanation. In this sense, our theory to a large extent settles their conjecture and offers a deeper understanding about the dimension's impact on the behavior of Hilbert-Schmidt covariance.

Distance and Hilbert-Schmidt (with Gaussian kernel or Laplacian kernel) covariance have been frequently applied to testing dependence between high dimensional vectors in biological science, see Kroupi et al. (2014), Kroupi et al. (2012), Hua and Ghosh (2015), Yang (2017), etc. In particular, Kroupi et al. (2014) use Hilbert-Schmidt (with Gaussian or Laplacian kernel) covariance to test the dependence between EEG signals for the perception of pleasant and unpleasant odors across subjects, where the data are collected for 5 subjects, each with 18 trials and dimension 250. Hua and Ghosh (2015) use Hilbert-Schmidt covariance with Gaussian kernel to examine the association between phenotype variable and genotype variable. The Alzheimers Disease Neuroimaging Initiative (ADNI) data is used in their simulation studies, where phenotype variable has dimension 119 and genotype variable has dimension 141. Finally, Hilbert-Schmidt covariance with Gaussian kernel is used by Yang (2017) to conduct independence test between neural responses and visual features, where the dimensions are of several hundreds. In the machine learning community, the application of Hilbert-Schmidt (with linear, Gaussian or Laplacian kernel) covariance for high dimensional data involves multi-label dimension reduction [Zhang and Zhou (2010), Xu et al. (2016), Mikalsen et al. (2019) etc] and unsupervised feature selection [Bedo (2008)], among others. In all the above-mentioned applications, the dimensions of the vectors involved are at hundreds or thousands.

1.1.1 Notations

In this paper, random data samples are denoted as, for each $i = 1, 2, \dots, n$, $X_i \stackrel{d}{=} X = (x_1, \dots, x_p)^T \in \mathbb{R}^p$, $Y_i \stackrel{d}{=} Y = (y_1, \dots, y_q)^T \in \mathbb{R}^q$. Next, let $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ and $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ denote the random sample matrices. In addition, the random component-wise samples are denoted as $\mathcal{X}_1, \dots, \mathcal{X}_p$ and $\mathcal{Y}_1, \dots, \mathcal{Y}_q$, which are illustrated in the following table,



Furthermore, matrices are denoted by upper case boldface letters (e.g. \mathbf{A} , \mathbf{B}). For any matrix $\mathbf{A} = (a_{st}) \in \mathbb{R}^{n \times n}$, we use $\tilde{\mathbf{A}} = (\tilde{a}_{st}) \in \mathbb{R}^{n \times n}$ to denote the \mathcal{U} -centered version of \mathbf{A} , i.e.,

$$\tilde{a}_{st} = \begin{cases} a_{st} - \frac{1}{n-2} \sum_{v=1}^n a_{sv} - \frac{1}{n-2} \sum_{u=1}^n a_{ut} + \frac{1}{(n-1)(n-2)} \sum_{u,v=1}^n a_{uv}, & s \neq t, \\ 0, & s = t. \end{cases}$$

Following Székely and Rizzo (2014), the inner product between two \mathcal{U} -centered matrices $\tilde{\mathbf{A}} = (\tilde{a}_{st}) \in \mathbb{R}^{n \times n}$ and $\tilde{\mathbf{B}} = (\tilde{b}_{st}) \in \mathbb{R}^{n \times n}$ is defined as

$$(\tilde{\mathbf{A}} \cdot \tilde{\mathbf{B}}) := \frac{1}{n(n-3)} \sum_{s \neq t} \tilde{a}_{st} \tilde{b}_{st}.$$

Next, we use $\mathbf{1}_n$ to denote the n dimensional column vector whose entries are all equal to 1. Similarly, we use $\mathbf{0}_n$ to denote the n dimensional column vector whose entries are all equal to 0. Finally, we use $|\cdot|$ to denote the L^2 norm of a vector, (X', Y') and (X'', Y'') to be independent copies of (X, Y) and $X \perp Y$ to indicate that X and Y are independent.

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_p , which is defined as follows: for a sequence of random variables $\{Z_s\}_{s \in \mathbb{Z}}$ and a sequence of numbers $\{a_s\}_{s \in \mathbb{Z}}$, $Z_s \asymp_p a_s$ if and only if $Z_s/a_s = O_p(1)$ and $a_s/Z_s = O_p(1)$ as $s \rightarrow \infty$. For more details about these notations, please see DasGupta (2008).

1.2 High Dimension Low Sample Size

The analyses in this section are conducted under the HDLSS setting, i.e., the sample size n is fixed and the dimensions $p \wedge q \rightarrow \infty$.

1.2.1 Distance Covariance and Variants

In this section, we introduce the following test statistics based on distance covariance ($dCov$), marginal distance covariance ($mdCov$), Hilbert-Schmidt covariance ($hCov$) and marginal Hilbert-Schmidt covariance ($mhCov$). In addition, their asymptotic behaviors under the HDLSS setting are derived. The following moment conditions will be used throughout the paper.

Assumption 1. *D1 For any p, q , the variance and the second moment of any coordinate of $X = (x_1, x_2, \dots, x_p)^T$ and $Y = (y_1, y_2, \dots, y_q)^T$ is uniformly bounded below and above, i.e.,*

$$\begin{aligned} 0 < a &\leq \inf_i \text{var}(x_i) \leq \sup_i E(x_i^2) \leq b < \infty, \\ 0 < a' &\leq \inf_j \text{var}(y_j) \leq \sup_j E(y_j^2) \leq b' < \infty, \end{aligned}$$

for some constants a, b, a', b' .

Next, denote $\tau_X^2 = E|X - X'|^2$, $\tau_Y^2 = E|Y - Y'|^2$ and $\tau^2 := \tau_X^2 \tau_Y^2 = E|X - X'|^2 E|Y - Y'|^2$. Notice that under Assumption 1, it can be easily seen that

$$\tau_X \asymp \sqrt{p}, \tau_Y \asymp \sqrt{q} \text{ and } \tau \asymp \sqrt{pq}.$$

The statistics we study in this work use the pair-wise L^2 distance between data points. The following proposition presents an expansion formula on the normalized L^2 distance when the dimension is high, which plays a key role in our theoretical analysis.

Proposition 2. *Under Assumption 1, we have*

$$\frac{|X - X'|}{\tau_X} = 1 + \frac{1}{2}L_X(X, X') + R_X(X, X'),$$

where

$$L_X(X, X') := \frac{|X - X'|^2 - \tau_X^2}{\tau_X^2},$$

and $R_X(X, X')$ is the remainder term. If we further assume that as $p \wedge q \rightarrow \infty$, $L_X(X, X') = o_p(1)$, then $R_X(X, X') = O_p(L_X(X, X')^2)$. Similar result holds for Y .

In order for the approximations in equations (1.1) and (1.2) to work well, it is required that $L_X(X_s, X_t)$ and $L_Y(Y_s, Y_t)$ should decay relatively fast as $p \wedge q \rightarrow \infty$. The following assumption specifies the order of $L_X(X_s, X_t)$ and $L_Y(Y_s, Y_t)$.

Assumption 3. $D2$ $L_X(X, X') = O_p(a_p)$ and $L_Y(Y, Y') = O_p(b_q)$, where a_p, b_q are sequences of numbers such that

$$\begin{aligned} a_p &= o(1), b_q = o(1), \\ \tau_X^2 a_p^3 &= o(1), \tau_Y^2 b_q^3 = o(1), \tau a_p^2 b_q = o(1), \tau a_p b_q^2 = o(1). \end{aligned}$$

Remark 4. A sufficient condition for $L_X(X, X') = o_p(1)$ is that $E[L_X(X, X')^2] = o(1)$. Let $\Sigma_X = \text{cov}(X)$. By a straightforward calculation, we obtain $|X - X'|^2 = \sum_{j=1}^p (x_j - x'_j)^2$, $E|X - X'|^2 = 2 \sum_{j=1}^p \text{var}(x_j) = 2\text{tr}(\Sigma_X)$, and

$$E[L_X(X, X')^2] = \frac{\sum_{j,j'=1}^p [\text{cov}(x_j^2, x_{j'}^2) + 2\text{cov}^2(x_j, x_{j'})]}{2\text{tr}^2(\Sigma_X)}.$$

Therefore, $E[L_X(X, X')^2] = o(1)$ holds if the component-wise dependence within X is not too strong. To illustrate this point, we consider the general multivariate model,

$$X_{p \times 1} = \mathbf{A}_{p \times s_1} U_{s_1 \times 1} + \mu_{p \times 1},$$

where \mathbf{A} is a constant matrix with $s_1 \geq p$, μ is the mean vector for X , and $U = (u_1, \dots, u_{s_1})^T$ has i.i.d components with mean zero and variance one. Suppose

$$\frac{\text{tr}(\mathbf{A}\mathbf{A}^T \mathbf{A}\mathbf{A}^T)}{\text{tr}^2(\mathbf{A}\mathbf{A}^T)} = \frac{\text{tr}(\Sigma_X^2)}{\text{tr}^2(\Sigma_X)} = O(p^{-1})$$

and $\sup_s E[u_s^4] < \infty$. Then the multivariate model satisfies Assumption 3 with $a_p = 1/\sqrt{p}$, see Section 1.7.2 of the Appendix for more details.

Distance Covariance

Distance covariance was first introduced by Székely et al. (2007) to measure the dependence between two random vectors of arbitrary dimensions. For two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, the (squared) distance

covariance is defined as

$$dCov^2(X, Y) = \int_{\mathbb{R}^{p+q}} \frac{|\phi_{X,Y}(t, s) - \phi_X(t)\phi_Y(s)|^2}{c_p c_q |t|^{1+p} |s|^{1+q}} dt ds,$$

where $c_p = \pi^{(1+p)/2} / \Gamma((1+p)/2)$, $|\cdot|$ is the (complex) Euclidean norm defined as $|x| = \sqrt{\bar{x}^T x}$ for any vector x in the complex vector space (\bar{x} denotes the conjugate of x), ϕ_X and ϕ_Y are the characteristic functions of X and Y respectively, $\phi_{X,Y}$ is the joint characteristic function. According to Theorem 7 of Székely and Rizzo (2009), an alternative definition of distance covariance is given by

$$dCov^2(X, Y) = E|X - X'| |Y - Y'| + E|X - X'| E|Y - Y'| - 2E|X - X'| |Y - Y''|, \quad (1.4)$$

where (X', Y') and (X'', Y'') are independent copies of (X, Y) . It has been shown that $dCov^2(X, Y) = 0$ if and only if X and Y are independent. Therefore, it is able to measure any type of dependence including non-linear and non-monotonic dependence between X and Y , whereas the commonly used Pearson correlation can only measure the linear dependence and the rank correlation coefficients (Kendall's τ and Spearman's ρ) can only capture the monotonic dependence.

Notice that in the above setting, p, q are arbitrary positive integers. Therefore, distance covariance is applicable to the high dimensional setting, where we allow $p, q \rightarrow \infty$. However, it is unclear whether this metric can still retain the power to detect the nonlinear dependence or not when the dimension is high. Distance correlation ($dCor$) is the normalized version of distance covariance, which is defined as

$$dCor^2(X, Y) = \begin{cases} \frac{dCov^2(X, Y)}{\sqrt{dCov^2(X, X) dCov^2(Y, Y)}}, & dCov^2(X, X) dCov^2(Y, Y) > 0, \\ 0, & dCov^2(X, X) dCov^2(Y, Y) = 0. \end{cases}$$

Following Székely and Rizzo (2014), we introduce the \mathcal{U} -centering based unbiased sample distance covariance ($dCov_n^2$) as follows.

$$dCov_n^2(\mathbf{X}, \mathbf{Y}) = (\tilde{\mathbf{A}} \cdot \tilde{\mathbf{B}}),$$

where $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}$ are the \mathcal{U} -centered versions of $\mathbf{A} = (a_{st})_{s,t=1}^n, \mathbf{B} = (b_{st})_{s,t=1}^n$ respectively and $a_{st} = |X_s - X_t|, b_{st} = |Y_s - Y_t|$ for $s, t = 1, \dots, n$. Correspondingly, the sample distance correlation ($dCor_n^2$) is given as

$$dCor_n^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{dCov_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{dCov_n^2(\mathbf{X}, \mathbf{X}) dCov_n^2(\mathbf{Y}, \mathbf{Y})}}, & dCov_n^2(\mathbf{X}, \mathbf{X}) dCov_n^2(\mathbf{Y}, \mathbf{Y}) > 0, \\ 0, & dCov_n^2(\mathbf{X}, \mathbf{X}) dCov_n^2(\mathbf{Y}, \mathbf{Y}) = 0. \end{cases}$$

Here, for $s \neq t$, we can apply the approximation in Proposition 2, that is

$$\frac{a_{st}}{\tau_X} = 1 + \frac{1}{2} L_X(X_s, X_t) + R_X(X_s, X_t), \quad (1.5)$$

$$\frac{b_{st}}{\tau_Y} = 1 + \frac{1}{2} L_Y(Y_s, Y_t) + R_Y(Y_s, Y_t), \quad (1.6)$$

where

$$L_X(X_s, X_t) = \frac{|X_s - X_t|^2 - \tau_X^2}{\tau_X^2}, \quad L_Y(Y_s, Y_t) = \frac{|Y_s - Y_t|^2 - \tau_Y^2}{\tau_Y^2},$$

and R_X, R_Y are the remainder terms from the approximation. The approximation of the pair-wise L^2 distance in Equations (1.5) and (1.6) is our building block to decompose the unbiased sample (squared) distance covariance ($dCov_n^2$) into a leading term plus a negligible remainder term under the HDLSS setting. The following main theorem summarizes the decomposition properties of sample distance covariance ($dCov_n^2$).

Theorem 1. *Under Assumption 1, we can show that*

(i)

$$dCov_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) + \mathcal{R}_n. \quad (1.7)$$

Here

$$cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) = \frac{1}{\binom{n}{4}} \sum_{s < t < u < v} h(x_{si}, x_{ti}, x_{ui}, x_{vi}; y_{sj}, y_{tj}, y_{uj}, y_{vj}),$$

and the kernel h is defined as

$$h(x_{si}, x_{ti}, x_{ui}, x_{vi}; y_{sj}, y_{tj}, y_{uj}, y_{vj}) = \frac{1}{4!} \sum_{*}^{(s,t,u,v)} \frac{1}{4} (x_{si} - x_{ti})(y_{sj} - y_{tj})(x_{ui} - x_{vi})(y_{uj} - y_{vj}),$$

where the summation $\sum_{*}^{(s,t,u,v)}$ is over all permutations of the 4-tuples of indices (s, t, u, v) and \mathcal{R}_n is the remainder term. $cov_n^2(\mathcal{X}_i, \mathcal{Y}_j)$ is a fourth-order U-statistic and is an unbiased estimator for the squared covariance between x_i and y_j , i.e., $E[cov_n^2(\mathcal{X}_i, \mathcal{Y}_j)] = cov^2(x_i, y_j)$.

(ii) Further suppose Assumption 3 holds. Then

$$\begin{aligned} \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) &= O_p(\tau a_p b_q), \\ \mathcal{R}_n &= O_p(\tau a_p^2 b_q + \tau a_p b_q^2) = o_p(1), \end{aligned}$$

thus the remainder term is of smaller order compared to the leading term and therefore is asymptotically negligible.

Equation (1.7) in Theorem 1 shows that the leading term for sample distance covariance is the sum of all component-wise squared sample cross-covariances scaled by τ , which depends on the marginal variances of X and Y . This theorem suggests that in the HDLSS setting, the sample distance covariance can only measure the component-wise linear dependence between the two random vectors.

As argued previously, sample distance covariance ($dCov_n^2$) based tests suffer from power loss when X and Y are component-wisely non-linear dependent but uncorrelated. To remedy this drawback, we can consider the following aggregation of marginal sample distance covariances,

$$mdCov_n^2(\mathbf{X}, \mathbf{Y}) = \sqrt{\binom{n}{2}} \sum_{i=1}^p \sum_{j=1}^q dCov_n^2(\mathcal{X}_i, \mathcal{Y}_j),$$

where $dCov_n^2(\mathcal{X}_i, \mathcal{Y}_j) = (\tilde{\mathbf{A}}(i) \cdot \tilde{\mathbf{B}}(j))$, $\tilde{\mathbf{A}}(i)$ and $\tilde{\mathbf{B}}(j)$ are the \mathcal{U} -centered versions of $\mathbf{A}(i) = (a_{st}(i))_{s,t=1}^n$, $\mathbf{B}(j) = (b_{st}(j))_{s,t=1}^n$ respectively and $a_{st}(i) = |x_{si} - x_{ti}|$, $b_{st}(j) = |y_{sj} - y_{tj}|$.

Note that $mdCov_n^2$ captures the pairwise low dimensional nonlinear dependence, which can be viewed as the main effects of the dependence between two high dimensional random vectors. It is natural in many fields of statistics to test for main effects first before proceeding to high order interactions. See Chakraborty and Zhang (2018) for some discussions on main effects and high order effects in the context of joint dependence testing. In the testing of mutual independence of a high dimensional vector, Yao et al. (2018) also approached the problem by testing the pairwise independence using distance covariance and demonstrated that there may be intrinsic difficulty to capture the effects beyond main effects (pairwise dependence in the mutual independence testing problem), as the tests that target joint dependence do not perform well in the high dimensional setting.

Hilbert-Schmidt Covariance

A generalization of the Distance Covariance ($dCov$) is Hilbert-Schmidt Covariance ($hCov$), first proposed and aka Hilbert-Schmidt independence criterion ($HSIC$) by Gretton et al. (2008). In particular, the (squared) Hilbert-Schmidt Covariance ($hCov$) is obtained by kernelizing the Euclidean distance in equation (1.4), i.e.,

$$hCov^2(X, Y) = E[K(X, X')L(Y, Y')] + E[K(X, X')]E[L(Y, Y')] - 2E[K(X, X')L(Y, Y'')],$$

where $(X', Y'), (X'', Y'')$ are independent copies of (X, Y) and K, L are user specified kernels. Following the literature, we consider the following widely used kernels

$$\begin{aligned} \text{Gaussian kernel: } K(x, y) &= \exp\left(-\frac{|x-y|^2}{2\gamma^2}\right), \\ \text{Laplacian kernel: } K(x, y) &= \exp\left(-\frac{|x-y|}{\gamma}\right), \end{aligned}$$

where γ is a bandwidth parameter. For later convenience, we focus on the kernels that can be represented compactly as $K(x, y) = f(|x-y|/\gamma)$ for some continuously differentiable function f . For example, the Gaussian and Laplacian kernel can be defined by choosing different function f ,

$$\begin{aligned} \text{Gaussian kernel: } K(x, y) &= f\left(\frac{|x-y|}{\gamma}\right), f(a) = \exp\left(-\frac{a^2}{2}\right), \\ \text{Laplacian kernel: } K(x, y) &= f\left(\frac{|x-y|}{\gamma}\right), f(a) = \exp(-a). \end{aligned}$$

In practice, the bandwidth parameter is usually set as the median of pair-wise sample L^2 distance. Thus, a natural estimator for $hCov^2(X, Y)$ is defined as

$$hCov_n^2(\mathbf{X}, \mathbf{Y}) = (\tilde{\mathbf{R}} \cdot \tilde{\mathbf{H}}),$$

where $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{H}}$ are the \mathcal{U} -centered versions of $\mathbf{R} = (r_{st})_{s,t=1}^n, \mathbf{H} = (h_{st})_{s,t=1}^n$ respectively and $r_{st} = h_{st} = 0$ if $s = t$, otherwise

$$\begin{cases} r_{st} = K(X_s, X_t, \mathbf{X}) = f\left(\frac{|X_s - X_t|}{\gamma_{\mathbf{X}}}\right), \gamma_{\mathbf{X}} = \text{median}\{|X_s - X_t|, s \neq t\}, \\ h_{st} = L(Y_s, Y_t, \mathbf{Y}) = g\left(\frac{|Y_s - Y_t|}{\gamma_{\mathbf{Y}}}\right), \gamma_{\mathbf{Y}} = \text{median}\{|Y_s - Y_t|, s \neq t\}. \end{cases}$$

Similar to the definition of distance correlation, the Hilbert-Schmidt Correlation ($hCor$) is defined as

$$hCor^2(X, Y) = \begin{cases} \frac{hCov^2(X, Y)}{\sqrt{hCov^2(X, X)hCov^2(Y, Y)}}, & hCov^2(X, X)hCov^2(Y, Y) > 0, \\ 0, & hCov^2(X, X)hCov^2(Y, Y) = 0, \end{cases}$$

and the sample Hilbert-Schmidt Correlation ($hCor_n^2$) is defined in the same way by replacing $hCov^2$ with the corresponding sample version.

Next, we can extend the decomposition results for sample distance covariance ($dCov_n^2$) to sample Hilbert-Schmidt covariance ($hCov_n^2$) as shown in the following theorem. Throughout the paper, we use $f^{(1)}$ and $f^{(2)}$ to denote the first and second derivative of the function f .

Theorem 2. *Under Assumption 1, we have*

(i)

$$\tau \times hCov_n^2(\mathbf{X}, \mathbf{Y}) = f^{(1)}\left(\frac{\tau_X}{\gamma_{\mathbf{X}}}\right) g^{(1)}\left(\frac{\tau_Y}{\gamma_{\mathbf{Y}}}\right) \frac{\tau_X}{\gamma_{\mathbf{X}}} \frac{\tau_Y}{\gamma_{\mathbf{Y}}} \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) + \mathcal{R}_n, \quad (1.8)$$

where cov_n^2 is defined the same as in Theorem 1 and \mathcal{R}_n is the remainder term.

(ii) Further suppose Assumption 3 holds. Then

$$\begin{aligned} f^{(1)}\left(\frac{\tau_X}{\gamma_{\mathbf{X}}}\right) g^{(1)}\left(\frac{\tau_Y}{\gamma_{\mathbf{Y}}}\right) \frac{\tau_X}{\gamma_{\mathbf{X}}} \frac{\tau_Y}{\gamma_{\mathbf{Y}}} &\asymp_p 1, \\ \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) &= O_p(\tau a_p b_q), \\ \mathcal{R}_n &= O_p(\tau a_p^2 b_q + \tau a_p b_q^2) = o_p(1). \end{aligned}$$

Thus the remainder term is of smaller order compared to the leading term and is therefore asymptotically negligible.

Notice that different from the decomposition of $dCov_n^2(\mathbf{X}, \mathbf{Y})$ as in Theorem 1, here we decompose $hCov_n^2$ multiplied by $\tau = \tau_X \tau_Y$. This is expected, since in $hCov_n^2$, each pair-wise distance is normalized by $\gamma_{\mathbf{X}}$ or $\gamma_{\mathbf{Y}}$, which has asymptotically the same magnitude as τ_X , τ_Y respectively. In the high dimensional case, the expansion (1.8) suggests that $hCov$ -based tests also suffer from power loss when X and Y are component-wisely uncorrelated but nonlinearly dependent.

To analyze the asymptotic property of sample Hilbert-Schmidt covariance, most literature would assume the bandwidth parameters to be fixed constants, see e.g. Gretton et al. (2008). In contrast, our approach can handle the case where these bandwidth parameters are selected to be the median of pairwise sample distance, which is random and whose magnitude increases with dimension.

Similar to the marginal distance covariance introduced in Section 1.2.1, we can also aggregate the marginal Hilbert-Schmidt Covariance ($mhCov$), which is defined as

$$mhCov_n^2(\mathbf{X}, \mathbf{Y}) = \sqrt{\binom{n}{2}} \sum_{i=1}^p \sum_{j=1}^q hCov_n^2(\mathcal{X}_i, \mathcal{Y}_j)$$

where $hCov_n^2(\mathcal{X}_i, \mathcal{Y}_j) = (\tilde{\mathbf{R}}(i) \cdot \tilde{\mathbf{H}}(j))$, $\tilde{\mathbf{R}}(i)$ and $\tilde{\mathbf{H}}(j)$ are \mathcal{U} -centered version of $\mathbf{R}(i) = (r_{st}(i))_{s,t=1}^n$, $\mathbf{H}(j) = (h_{st}(j))_{s,t=1}^n$ respectively and $r_{st}(i) = h_{st}(j) = 0$ if $s = t$, otherwise

$$\begin{cases} r_{st}(i) = K(x_{si}, x_{ti}, \mathcal{X}_i) = f\left(\frac{|x_{si} - x_{ti}|}{\gamma_{\mathcal{X}_i}}\right), \gamma_{\mathcal{X}_i} = \text{median}\{|x_{si} - x_{ti}|, s \neq t\}, \\ h_{st}(j) = L(y_{sj}, y_{tj}, \mathcal{Y}_j) = g\left(\frac{|y_{sj} - y_{tj}|}{\gamma_{\mathcal{Y}_j}}\right), \gamma_{\mathcal{Y}_j} = \text{median}\{|y_{sj} - y_{tj}|, s \neq t\}. \end{cases}$$

Remark 5. By the multi-linearity of the operator (\cdot, \cdot) , it can be easily seen that $mdCov_n^2(\mathbf{X}, \mathbf{Y})$ is equal to (up to a constant) $hCov_n^2(\mathbf{X}, \mathbf{Y})$ equipped with L^1 -distance and $mhCov_n^2(\mathbf{X}, \mathbf{Y})$ is equal to (up to a constant) $hCov_n^2(\mathbf{X}, \mathbf{Y})$ equipped with kernels

$$K'(X_s, X_t, \mathbf{X}) = \sum_{i=1}^p K(x_{si}, x_{ti}, \mathcal{X}_i) \text{ and } L'(Y_s, Y_t, \mathbf{Y}) = \sum_{j=1}^q L(y_{sj}, y_{tj}, \mathcal{Y}_j).$$

1.2.2 Studentized Test Statistics

In this section, we provide studentized version of the statistics introduced in Section 1.2.1. It is worth mentioning that we provide a unified approach to the asymptotic analysis of studentized $dCov$, $mdCov$ and further extend them to the analysis of studentized $hCov$.

Unified Approach

Firstly, we will present results that will be useful for deriving the studentized version of the interested statistics, i.e. distance covariance ($dCov$), marginal distance covariance ($mdCov$), Hilbert-Schmidt Covariance ($hCov$), marginal Hilbert-Schmidt Covariance ($mhCov$). It can be shown later that many previously mentioned statistics are asymptotically equal to the unified quantity $uCov_n^2(\mathbf{X}, \mathbf{Y})$ multiplied by some normalizing factor. Here, $uCov_n^2(\mathbf{X}, \mathbf{Y})$ is defined as

$$uCov_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{\sqrt{pq}} \sum_{i=1}^p \sum_{j=1}^q (\tilde{\mathbf{K}}(i) \cdot \tilde{\mathbf{L}}(j)),$$

where $\tilde{\mathbf{K}}(i)$ and $\tilde{\mathbf{L}}(j)$ are the \mathcal{U} -centered versions of $\mathbf{K}(i) = (k_{st}(i))_{s,t=1}^n$, $\mathbf{L}(j) = (l_{st}(j))_{s,t=1}^n$ respectively and $k_{st}(i) = l_{st}(j) = 0$ if $s = t$, otherwise $k_{st}(i)$, $l_{st}(j)$ are the double centered kernel distances, i.e., for bivariate kernels k and l ,

$$\begin{aligned} k_{st}(i) &= k(x_{si}, x_{ti}) - E[k(x_{si}, x_{ti})|x_{si}] - E[k(x_{si}, x_{ti})|x_{ti}] + E[k(x_{si}, x_{ti})], \\ l_{st}(j) &= l(y_{sj}, y_{tj}) - E[l(y_{sj}, y_{tj})|y_{sj}] - E[l(y_{sj}, y_{tj})|y_{tj}] + E[l(y_{sj}, y_{tj})]. \end{aligned}$$

The advantage of using the double centering kernel distance is that we can have 0 covariance between $k_{st}(i)$ and $k_{uv}(j)$ ($l_{st}(i)$ and $l_{uv}(j)$) for $\{s, t\} \neq \{u, v\}$ as shown in the following proposition.

Proposition 6. For all $1 \leq i, i' \leq p, 1 \leq j, j' \leq q$, if $\{s, t\} \neq \{u, v\}$, then

$$E[k_{st}(i)k_{uv}(i')] = E[l_{st}(j)l_{uv}(j')] = E[k_{st}(i)l_{uv}(j)] = 0.$$

To derive the limiting distribution of the unified quantity, we need the following assumptions.

Assumption 7. *D3 For fixed n , as $p \wedge q \rightarrow \infty$,*

$$\begin{pmatrix} p^{-1/2} \sum_{i=1}^p k_{st}(i) \\ q^{-1/2} \sum_{j=1}^q l_{uv}(j) \end{pmatrix}_{s < t, u < v} \xrightarrow{d} \begin{pmatrix} c_{st} \\ d_{uv} \end{pmatrix}_{s < t, u < v},$$

where $\{c_{st}, d_{uv}\}_{s < t, u < v}$ are jointly Gaussian. Naturally, we further assume the existence of the following constants that show up in the covariance matrix of $\{c_{st}, d_{uv}\}$,

$$\begin{aligned} \text{var}[c_{st}] &:= \sigma_x^2 = \lim_p \frac{1}{p} \sum_{i,j=1}^p \text{cov}[k_{st}(i), k_{st}(j)] \\ &= \begin{cases} \lim_p \frac{\sum_{i,j=1}^p d\text{Cov}^2(x_i, x_j)}{p}, & \text{if } k(x, y) = l(x, y) = |x - y|, \\ \lim_p \frac{\sum_{i,j=1}^p 4\text{cov}^2(x_i, x_j)}{p}, & \text{if } k(x, y) = l(x, y) = |x - y|^2, \end{cases} \\ \text{var}[d_{st}] &:= \sigma_y^2 = \lim_q \frac{1}{q} \sum_{i,j=1}^q \text{cov}[l_{st}(i), l_{st}(j)] \\ &= \begin{cases} \lim_q \frac{\sum_{i,j=1}^q d\text{Cov}^2(y_i, y_j)}{q}, & \text{if } k(x, y) = l(x, y) = |x - y|, \\ \lim_q \frac{\sum_{i,j=1}^q 4\text{cov}^2(y_i, y_j)}{q}, & \text{if } k(x, y) = l(x, y) = |x - y|^2, \end{cases} \end{aligned}$$

$$\begin{aligned} \text{cov}[c_{st}, d_{st}] &:= \sigma_{xy}^2 = \lim_{p,q} \frac{1}{\sqrt{pq}} \sum_{i=1}^p \sum_{j=1}^q \text{cov}[k_{st}(i), l_{st}(j)] \\ &= \begin{cases} \lim_{p,q} \frac{\sum_{i=1}^p \sum_{j=1}^q d\text{Cov}^2(x_i, y_j)}{\sqrt{pq}}, & \text{if } k(x, y) = l(x, y) = |x - y|, \\ \lim_{p,q} \frac{\sum_{i=1}^p \sum_{j=1}^q 4\text{cov}^2(x_i, y_j)}{\sqrt{pq}}, & \text{if } k(x, y) = l(x, y) = |x - y|^2. \end{cases} \end{aligned}$$

Remark 8. Notice that when $\{s, t\} \neq \{u, v\}$, we do not assume the form of $\text{cov}[c_{st}, c_{uv}]$, $\text{cov}[d_{st}, d_{uv}]$, $\text{cov}[c_{st}, d_{uv}]$ in Assumption 7, since it follows easily from Proposition 6 that $\text{cov}[c_{st}, c_{uv}] = 0$, $\text{cov}[d_{st}, d_{uv}] = 0$ and $\text{cov}[c_{st}, d_{uv}] = 0$ if $\{s, t\} \neq \{u, v\}$.

Remark 9. The above Central Limit Theorem (CLT) result can be derived under suitable moment and weak dependence assumptions for the components of X and Y . We refer the reader to Doukhan and Neumann (2008) for a relatively recent survey of weak dependence notions and the CLT results under such weak dependence. It is worth noting that the commonly used weak dependence assumptions in time series analysis, such as α -mixing, β -mixing and variants [Bradley (2007)], near epoch dependence [Gallant and White (1988), Davidson (1994)] and physical dependence measure [Wu (2005a)], all require the components have a natural time ordering. In our setting, the components do not necessarily have a natural ordering but our results still hold as long as there exists a permutation of components that satisfy the weak dependence assumption. Furthermore we remark that the weak dependence assumption typically rules out long range dependence and local strong dependence, under which we might have non-Gaussian limit and different norming rate. We shall examine the validity of our tests in these two scenarios via simulations.

The following theorem is our main result, which shows that the unified quantity converges in distribution to a quadratic form of random variables.

Theorem 3. Fixing n and letting $p \wedge q \rightarrow \infty$, under Assumptions 1 and 7,

$$\begin{aligned} uCov_n^2(\mathbf{X}, \mathbf{Y}) &\xrightarrow{d} \frac{1}{v} \mathbf{c}^T \mathbf{M} \mathbf{d}, \\ uCov_n^2(\mathbf{X}, \mathbf{X}) &\xrightarrow{d} \frac{1}{v} \mathbf{c}^T \mathbf{M} \mathbf{c} \stackrel{d}{=} \frac{\sigma_x^2}{v} \chi_v^2, \\ uCov_n^2(\mathbf{Y}, \mathbf{Y}) &\xrightarrow{d} \frac{1}{v} \mathbf{d}^T \mathbf{M} \mathbf{d} \stackrel{d}{=} \frac{\sigma_y^2}{v} \chi_v^2, \end{aligned}$$

where $v := n(n-3)/2$, \mathbf{M} is a projection matrix of rank v and

$$\begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \stackrel{d}{=} N \left(\mathbf{0}, \begin{pmatrix} \sigma_x^2 \mathbf{I}_{n(n-1)/2} & \sigma_{xy}^2 \mathbf{I}_{n(n-1)/2} \\ \sigma_{xy}^2 \mathbf{I}_{n(n-1)/2} & \sigma_y^2 \mathbf{I}_{n(n-1)/2} \end{pmatrix} \right).$$

For the exact form of \mathbf{M} , see the proof of Theorem 3 in the Appendix. Next, we define the quantity T_u as

$$T_u = \sqrt{v-1} \frac{uCor_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{1 - (uCor_n^2(\mathbf{X}, \mathbf{Y}))^2}},$$

where

$$uCor_n^2(\mathbf{X}, \mathbf{Y}) = \frac{uCov_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{uCov_n^2(\mathbf{X}, \mathbf{X})uCov_n^2(\mathbf{Y}, \mathbf{Y})}}.$$

We then define the constants v and ϕ that appear in the limiting distribution of T_u . Set $v = n(n-3)/2$ and $\phi = \sigma_{xy}^2 / \sqrt{\sigma_x^2 \sigma_y^2}$ such that

$$\phi = \phi_1 \mathbb{I}_{\{k(x,y)=l(x,y)=|x-y|\}} + \phi_2 \mathbb{I}_{\{k(x,y)=l(x,y)=|x-y|^2\}},$$

where

$$\begin{aligned} \phi_1 &:= \lim_{p,q} \frac{\sum_{i=1}^p \sum_{j=1}^q dCov^2(x_i, y_j)}{\sqrt{\sum_{i,j=1}^p dCov^2(x_i, x_j) \sum_{i,j=1}^q dCov^2(y_i, y_j)}}, \\ \phi_2 &:= \lim_{p,q} \frac{\sum_{i=1}^p \sum_{j=1}^q cov^2(x_i, y_j)}{\sqrt{\sum_{i,j=1}^p cov^2(x_i, x_j) \sum_{i,j=1}^q cov^2(y_i, y_j)}}. \end{aligned}$$

The limiting distribution of T_u is derived under both null (H_0) and alternative (H_A) hypothesis, i.e.,

$$\begin{aligned} \text{null hypothesis : } H_0 &= \{(X, Y) \mid X \perp Y\}, \\ \text{alternative hypothesis : } H_A &= \{(X, Y) \mid X \not\perp Y\}. \end{aligned}$$

In addition, we also consider the local alternative hypothesis $H_{A_l} \subset H_A$, i.e.,

$$H_{A_l} = \left\{ (X, Y) \mid X \not\perp Y, \phi = \frac{\phi_0}{\sqrt{v}} \right\},$$

where $v = n(n-3)/2$, $\phi_0 = \phi_{0,1} \mathbb{I}_{\{k(x,y)=l(x,y)=|x-y|\}} + \phi_{0,2} \mathbb{I}_{\{k(x,y)=l(x,y)=|x-y|^2\}}$ and $0 < \phi_{0,1}, \phi_{0,2} < \infty$ are constants with respect to n . It is also interesting to compare the asymptotic power under the following class

of alternatives $H_{A_s} \subset H_A$, i.e.,

$$H_{A_s} = \{(X, Y) \mid x_i \not\perp y_j, \text{cov}(x_i, y_j) = 0 \text{ for all } 1 \leq i \leq p, 1 \leq j \leq q\}.$$

In summary, the following table illustrates the value of ϕ under different cases we are considering,

ϕ	H_0	H_A	H_{A_l}	H_{A_s}
$k(x, y) = l(x, y) = x - y $	0	ϕ_1	$\frac{\phi_{0,1}}{\sqrt{v}}$	ϕ_1
$k(x, y) = l(x, y) = x - y ^2$	0	ϕ_2	$\frac{\phi_{0,2}}{\sqrt{v}}$	0

Next, denote by t_a the student t -distribution with degrees of freedom a . Let $t_a^{(\alpha)}$ be the $(1 - \alpha)$ th percentile of t_a and $t_{a,b}$ be the non-central t -distribution with degrees of freedom a and non-central parameter b . The asymptotic distribution of T_u is stated in the following proposition.

Proposition 10. *Fix n and let $p \wedge q \rightarrow \infty$. If Assumptions 1 and 7 hold, then for any fixed $t \in \mathbb{R}$,*

$$\begin{aligned} P_{H_0}(T_u \leq t) &\rightarrow P(t_{v-1} \leq t), \\ P_{H_A}(T_u \leq t) &\rightarrow E[P(t_{v-1,W} \leq t)], \end{aligned}$$

where $W \sim \sqrt{\frac{\phi^2}{1-\phi^2}} \chi_v^2$ and χ_v^2 is the chi-square distribution with degrees of freedom v .

Remark 11. *For the explicit form of $E[P(t_{v-1,W} \leq t)]$, see Lemma 26 in the Appendix.*

Below we derive the large sample approximation of the limiting distribution $E[P(t_{v-1,W} \leq t)]$ under the local alternative hypothesis (H_{A_l}).

Proposition 12. *Under H_{A_l} , if we allow n to grow and t is bounded as $n \rightarrow \infty$, $E[P(t_{v-1,W} \leq t)]$ can be approximated as*

$$E_{H_{A_l}}[P(t_{v-1,W} \leq t)] = P(t_{v-1,\phi_0} \leq t) + O\left(\frac{1}{v}\right),$$

where $\phi_0 = \phi_{0,1}\mathbb{I}_{\{k(x,y)=l(x,y)=|x-y|\}} + \phi_{0,2}\mathbb{I}_{\{k(x,y)=l(x,y)=|x-y|^2\}}$. In particular, the result still holds if we replace t with $t_{v-1}^{(\alpha)}$.

Studentized Tests

For testing the null, permutation test can be used to determine the critical value of the distance covariance ($dCov$), Hilbert-Schmidt covariance ($hCov$), marginal distance covariance ($mdCov$) and marginal Hilbert-Schmidt covariance ($mhCov$) respectively. If $dCov_n^2$, $hCov_n^2$, $mdCov_n^2$ or $mhCov_n^2$ is larger than the corresponding critical value, which can be determined by the empirical permutation distribution function, we reject the null. Alternatively, similar to the construction of T_u , we transform each of $dCov_n^2$, $hCov_n^2$, $mdCov_n^2$ and $mhCov_n^2$ into a statistic that has asymptotic t -distribution under the null. Thus, instead of using permutation test, which can be quite computationally expensive, we can determine the critical value using this asymptotic t -distribution. For each $R \in \{dCov, hCov, mdCov, mhCov\}$, the studentized test statistic T_R is defined as

$$T_R = \sqrt{v-1} \frac{R^*(\mathbf{X}, \mathbf{Y})}{\sqrt{1 - (R^*(\mathbf{X}, \mathbf{Y}))^2}},$$

where

$$R^*(\mathbf{X}, \mathbf{Y}) = \frac{R_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{R_n^2(\mathbf{X}, \mathbf{X})R_n^2(\mathbf{Y}, \mathbf{Y})}}.$$

The way to derive the asymptotic distribution of T_R is to show that for each $R \in \{dCov, hCov, mdCov\}$, $R_n^2(\mathbf{X}, \mathbf{Y})$ and $uCov_n^2(\mathbf{X}, \mathbf{Y})$ are asymptotically equal up to an asymptotically constant factor, as shown below.

Proposition 13. *Under Assumption 1,*

(i) *When $k(x, y) = l(x, y) = |x - y|^2$,*

$$\begin{aligned} dCov_n^2(\mathbf{X}, \mathbf{Y}) &= \frac{1}{4} \frac{\sqrt{pq}}{\tau} uCov_n^2(\mathbf{X}, \mathbf{Y}) + \mathcal{R}'_n, \\ \tau \times hCov_n^2(\mathbf{X}, \mathbf{Y}) &= \frac{\sqrt{pq}}{4\gamma_{\mathbf{X}}\gamma_{\mathbf{Y}}} f^{(1)}\left(\frac{\tau_{\mathbf{X}}}{\gamma_{\mathbf{X}}}\right) g^{(1)}\left(\frac{\tau_{\mathbf{Y}}}{\gamma_{\mathbf{Y}}}\right) uCov_n^2(\mathbf{X}, \mathbf{Y}) + \mathcal{R}''_n, \end{aligned}$$

where $\mathcal{R}'_n, \mathcal{R}''_n$ are the remainder terms. Further suppose Assumption 3 holds. Then

$$\begin{aligned} uCov_n^2(\mathbf{X}, \mathbf{Y}) &= O_p(\tau a_p b_q), \\ \mathcal{R}'_n &= O_p(\tau a_p^2 b_q + \tau a_p b_q^2) = o_p(1), \\ \mathcal{R}''_n &= O_p(\tau a_p^2 b_q + \tau a_p b_q^2) = o_p(1). \end{aligned}$$

Thus the remainder term is of smaller order compared to the leading term and therefore is asymptotically negligible.

(ii) *When $k(x, y) = l(x, y) = |x - y|$,*

$$mdCov_n^2(\mathbf{X}, \mathbf{Y}) = \sqrt{pq} \sqrt{\binom{n}{2}} uCov_n^2(\mathbf{X}, \mathbf{Y}).$$

As shown in Proposition 13, $k(x, y) = l(x, y) = |x - y|$ would correspond to the $mdCov$ -based t -test and $k(x, y) = l(x, y) = |x - y|^2$ would correspond to the $\{dCov, hCov\}$ -based t -tests. Then, for each $R \in \{dCov, hCov, mdCov\}$ the asymptotic distribution of T_R is given in the following Corollary.

Corollary 1. *If Assumptions 1, 3 and 7 hold, for any fixed t and each $R \in \{dCov, hCov, mdCov\}$, we have*

$$\begin{aligned} P_{H_0}(T_R \leq t) &\rightarrow P(t_{v-1} \leq t), \\ P_{H_A}(T_R \leq t) &\rightarrow E[P(t_{v-1, W} \leq t)], \text{ where } W \sim \sqrt{\frac{\phi^2}{1 - \phi^2}} \chi_v^2. \end{aligned}$$

After knowing the asymptotic distribution of T_R under the null, i.e. t -distribution with degrees of freedom $v - 1$, we can set critical value as $t_{v-1}^{(\alpha)}$. Then, from Proposition 10, under the alternative, the asymptotic power of testing the null can be written as a function of ϕ , i.e.,

$$Power_n(\phi) := E[P(t_{v-1, W} > t_{v-1}^{(\alpha)})],$$

and under H_{A_l} , if we allow n to grow

$$Power_\infty(\phi_0) := \lim_{n \rightarrow \infty} Power_n \left(\frac{\phi_0}{\sqrt{v}} \right) = \lim_{n \rightarrow \infty} P \left(t_{v-1, \phi_0} > t_{v-1}^{(\alpha)} \right).$$

Next, we can actually bound the ratio of ϕ_1 and ϕ_2 for standard normal random variables.

Proposition 14. *Suppose that*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \stackrel{d}{=} N \left(\mathbf{0}, \begin{pmatrix} \mathbf{I}_p & \Sigma_{XY} \\ \Sigma_{XY}^T & \mathbf{I}_q \end{pmatrix} \right),$$

where $\Sigma_{XY} = cov(X, Y)$. We have

$$0.89^2 \phi_2 \leq \phi_1 \leq \phi_2.$$

It will be shown later that ϕ_1 corresponds to the $mdCov$ -based test, whereas ϕ_2 corresponds to the $dCov$ and $hCov$ -based tests. Thus considering models described in Proposition 14, we expect a power loss for the $mdCov$ -based test comparing to the $dCov$ and $hCov$ -based tests. On the other hand, since ϕ_1 is bounded below by $0.89^2 \phi_2$, the power loss is expected to be moderate.

Using Corollary 1, we can theoretically compare the power of these t -tests under different cases and the results are summarized in the following table

<i>Power</i>	T_{mdCov}	T_{dCov}, T_{hCov}
under H_A	$Power_n(\phi_1)$	$Power_n(\phi_2)$
under H_{A_l} , allow n growing to infinity	$Power_\infty(\phi_{0,1})$	$Power_\infty(\phi_{0,2})$
under H_{A_s}	$Power_n(\phi_1)$	α

For the studentized version of $mhCov$, if we consider the bandwidth parameters to be fixed constants, then we can use the unified approach to get the limiting t -distribution of the transformed $mhCov_n^2$. On the other hand, if $\gamma_{\mathcal{X}_i}$ and $\gamma_{\mathcal{Y}_j}$ are treated to be median of sample distance along each dimension and are thus random, we encounter technical difficulties to derive the limiting distribution, as in this case the kernelized pair-wise distance along each dimension are correlated with each other. This is due to the choice of the bandwidth parameter and the high dimensional approximation used for $hCov_n^2$ can not be directly applied, since $\gamma_{\mathcal{X}_i}$ and $\gamma_{\mathcal{Y}_j}$ are calculated component-wisely. Nevertheless, we shall examine the testing efficiency using t -distribution approximation when the bandwidth parameters are chosen to be the median of sample distance in simulation.

Remark 15. *An anonymous referee inquired about the applicability of our tests to the setting when the p -dimensional data vector $X = (x_1, \dots, x_p)^T$ is a growth curve and thus can be viewed as a stochastic process or random function evaluated at time points $\{t_i\}_{i=1}^p$, say $0 \leq t_1 < t_2 < \dots < t_p \leq 1$. Under suitable conditions, one can show that the Euclidean norm of X after proper scaling converges to the L_2 norm of the random function when the number of sampling points goes to infinity. It is known that the Hilbert space $L_2([0, 1])$ is of strong negative type [Lyons (2013)], and thus the HSIC or the distance covariance based on the L_2 norm completely characterizes dependence. Therefore, the Euclidean norm is a proper norm to use if X is considered to be an element in $L_2([0, 1])$ and we want to use the L_2 norm to construct our distance metrics. However, the setting we are considering in this paper assumed the components of X and Y have*

weak dependence and the above growth curve example falls into the very strong dependent case, and thus our theoretical phenomenon does not apply. In practice, both strongly componentwise correlated high dimensional data and weakly componentwise dependent high dimensional data can be collected depending on the nature of data generating process. We shall illustrate the usefulness of our theory and proposed tests using an earthquake dataset in Section 1.3.

Our theory demonstrates the limitation of $dCov$ and $hCov$ in the high dimensional environment, which is intimately related to the use of Euclidean norm in their definitions. Similar phenomenon has been discovered for energy distance [Székely and Rizzo (2004)] and maximum mean discrepancy [Gretton et al. (2012a)] in the two sample testing problem recently; see Zhu and Shao (2019) and Chakraborty and Zhang (2019a). It is natural to ask what norm would be desirable to use in the high dimensional setting and in what sense? We shall leave these questions for future study.

1.3 Numerical Results

Here, we consider some numerical examples to compare the “joint” tests, where the distance/Hilbert-Schmidt covariance is applied to whole components of data jointly, with the “marginal” tests, where distance/Hilbert-Schmidt covariance is applied to one dimensional components and then being aggregated. To this end, we consider the following statistics

$$\begin{aligned} \text{“Joint”} & \left\{ \begin{array}{l} dCov : \text{distance covariance (permutation)} \\ T_{dCov} : \text{studentized distance covariance} \\ hCov : \text{Hilbert-Schmidt covariance (permutation)} \\ T_{hCov} : \text{studentized Hilbert-Schmidt covariance} \end{array} \right. \\ \text{“Marginal”} & \left\{ \begin{array}{l} mdCov : \text{marginal distance covariance (permutation)} \\ T_{mdCov} : \text{studentized marginal distance covariance,} \\ mhCov : \text{marginal Hilbert-Schmidt covariance (permutation)} \\ T_{mhCov} : \text{studentized marginal Hilbert-Schmidt covariance} \end{array} \right. \end{aligned}$$

In the above display, $dCov_n^2$ and $hCov_n^2$ are the two “joint” test statistics to measure the overall dependence between X and Y , $mdCov_n^2$ and $mhCov_n^2$ are the “marginal” test statistics, and these four test statistics are implemented as permutation tests; T_{dCov} from Székely and Rizzo (2013) is the studentized version of $dCov$, our proposed t -tests $T_{hCov}, T_{mdCov}, T_{mhCov}$ are the studentized versions of $hCov, mdCov, mhCov$ respectively. All these four tests are implemented using the t -distribution based critical value. We examine both the Gaussian kernel and Laplacian kernel for the Hilbert-Schmidt covariance based tests.

For the permutation-based tests, we randomly shuffle the samples $\{X_1, \dots, X_n\}$ and get $(X_{\pi(1)}, \dots, X_{\pi(n)})$, where π is the permutation map from $\{1, \dots, n\}$ to $\{1, \dots, n\}$. Then we calculate the test statistic based on the permuted sample $\{(X_{\pi(1)}, \dots, X_{\pi(n)}), (Y_1, \dots, Y_n)\}$. The p -value for permutation-based test is defined as the proportion of times that the test statistic based on the permuted samples is greater than the one based on the original sample. All the numerical results from permutation-based tests are based on 200 permutations and the empirical rejection rate of the tests are based on 5000 Monte Carlo repetitions.

We first examine the size of the afore-mentioned tests.

Example 1. Generate i.i.d. samples from the following models for $i = 1, \dots, n$.

$$(i) \ X_i = (x_{i1}, \dots, x_{ip})^T \sim N(\mathbf{0}, \mathbf{I}_p), Y_i = (y_{i1}, \dots, y_{ip})^T \sim N(\mathbf{0}, \mathbf{I}_p).$$

(ii) Let $AR(1)$ denotes the Gaussian autoregressive model of order 1 with parameter $\phi, X_i \sim AR(1), \phi = 0.5, Y_i \sim AR(1), \phi = -0.5$.

(iii) Let $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ and $\sigma_{ij} = 0.7^{|i-j|}$, $X_i = (x_{i1}, \dots, x_{ip})^T \sim N(\mathbf{0}, \Sigma), Y_i = (y_{i1}, \dots, y_{ip})^T \sim N(\mathbf{0}, \Sigma)$.

From Table 1.1, we can see that all the tests have quite accurate size. Although the t -tests are derived under the high dimensional scenario, they still have pretty accurate size even for relatively low dimension (e.g., $p = 5$).

Table 1.1: Size comparison from Example 1

	n	p	α	$dCov$	$mdCov$	T_{dCov}	T_{mdCov}	Gaussian Kernel				Laplacian Kernel			
								$hCov$	$mhCov$	T_{hCov}	T_{mhCov}	$hCov$	$mhCov$	T_{hCov}	T_{mhCov}
(i)	10	5	0.010	0.017	0.014	0.020	0.014	0.016	0.015	0.020	0.014	0.014	0.014	0.017	0.013
	10	5	0.050	0.055	0.055	0.062	0.061	0.055	0.060	0.062	0.061	0.055	0.050	0.064	0.050
	10	5	0.100	0.105	0.107	0.110	0.110	0.103	0.106	0.109	0.109	0.102	0.099	0.105	0.101
	10	30	0.010	0.015	0.015	0.013	0.011	0.015	0.016	0.012	0.012	0.014	0.014	0.011	0.011
	10	30	0.050	0.054	0.053	0.050	0.053	0.054	0.050	0.052	0.052	0.052	0.059	0.050	0.054
	10	30	0.100	0.102	0.104	0.099	0.102	0.102	0.105	0.100	0.103	0.102	0.107	0.101	0.105
	30	5	0.010	0.014	0.016	0.019	0.018	0.016	0.016	0.020	0.017	0.016	0.015	0.019	0.015
	30	5	0.050	0.052	0.053	0.062	0.059	0.052	0.057	0.061	0.059	0.054	0.055	0.061	0.058
	30	5	0.100	0.105	0.104	0.105	0.107	0.103	0.107	0.106	0.106	0.105	0.104	0.109	0.104
	30	30	0.010	0.014	0.014	0.011	0.012	0.014	0.017	0.010	0.013	0.014	0.017	0.011	0.013
	30	30	0.050	0.051	0.053	0.052	0.051	0.051	0.056	0.052	0.053	0.051	0.058	0.051	0.052
	30	30	0.100	0.097	0.105	0.096	0.103	0.097	0.105	0.095	0.101	0.099	0.104	0.100	0.102
	60	5	0.010	0.013	0.015	0.018	0.016	0.014	0.013	0.019	0.016	0.014	0.015	0.017	0.015
	60	5	0.050	0.052	0.055	0.061	0.057	0.054	0.061	0.060	0.064	0.053	0.057	0.058	0.058
	60	5	0.100	0.103	0.104	0.109	0.104	0.107	0.108	0.110	0.110	0.102	0.101	0.103	0.102
	60	30	0.010	0.019	0.017	0.016	0.012	0.019	0.015	0.015	0.013	0.020	0.016	0.015	0.014
	60	30	0.050	0.060	0.063	0.057	0.058	0.060	0.058	0.057	0.058	0.061	0.058	0.058	0.055
	60	30	0.100	0.113	0.112	0.110	0.107	0.113	0.109	0.111	0.105	0.110	0.111	0.107	0.107
	10	5	0.010	0.015	0.015	0.023	0.023	0.014	0.016	0.023	0.019	0.015	0.017	0.022	0.021
	10	5	0.050	0.051	0.054	0.064	0.066	0.053	0.058	0.064	0.066	0.054	0.058	0.066	0.062
	10	5	0.100	0.101	0.105	0.107	0.111	0.100	0.109	0.105	0.113	0.102	0.110	0.106	0.109
	10	30	0.010	0.014	0.018	0.013	0.016	0.014	0.017	0.014	0.013	0.017	0.018	0.017	0.013
	10	30	0.050	0.060	0.061	0.061	0.061	0.061	0.056	0.062	0.056	0.059	0.060	0.059	0.056
	10	30	0.100	0.105	0.105	0.110	0.107	0.105	0.105	0.109	0.099	0.106	0.108	0.111	0.104
(ii)	30	5	0.010	0.012	0.011	0.022	0.023	0.012	0.014	0.021	0.020	0.013	0.013	0.019	0.016
	30	5	0.050	0.046	0.048	0.055	0.056	0.046	0.052	0.055	0.059	0.047	0.053	0.051	0.059
	30	5	0.100	0.094	0.096	0.094	0.096	0.096	0.100	0.097	0.100	0.093	0.107	0.097	0.104
	30	30	0.010	0.016	0.016	0.017	0.015	0.017	0.015	0.017	0.011	0.017	0.015	0.017	0.012
	30	30	0.050	0.061	0.058	0.060	0.059	0.061	0.055	0.060	0.054	0.058	0.052	0.060	0.051
	30	30	0.100	0.109	0.105	0.110	0.107	0.111	0.101	0.110	0.098	0.111	0.102	0.113	0.097
	60	5	0.010	0.015	0.013	0.026	0.022	0.016	0.014	0.024	0.020	0.013	0.015	0.020	0.018
	60	5	0.050	0.055	0.052	0.062	0.061	0.055	0.053	0.061	0.059	0.055	0.052	0.061	0.054
	60	5	0.100	0.101	0.100	0.103	0.100	0.102	0.100	0.104	0.099	0.101	0.097	0.103	0.099
	60	30	0.010	0.013	0.014	0.013	0.014	0.013	0.016	0.014	0.013	0.014	0.015	0.013	0.012
	60	30	0.050	0.055	0.051	0.058	0.051	0.054	0.054	0.057	0.053	0.058	0.053	0.053	0.052
	60	30	0.100	0.105	0.102	0.105	0.100	0.106	0.103	0.105	0.102	0.107	0.105	0.107	0.104
	10	5	0.010	0.012	0.013	0.025	0.024	0.012	0.014	0.024	0.022	0.016	0.013	0.025	0.019
	10	5	0.050	0.051	0.051	0.068	0.069	0.053	0.051	0.068	0.062	0.053	0.049	0.067	0.056
	10	5	0.100	0.100	0.099	0.107	0.103	0.100	0.098	0.105	0.102	0.100	0.098	0.104	0.101
	10	30	0.010	0.014	0.015	0.016	0.014	0.014	0.015	0.016	0.013	0.015	0.015	0.017	0.013
	10	30	0.050	0.055	0.057	0.061	0.058	0.053	0.056	0.061	0.056	0.057	0.057	0.064	0.059
	10	30	0.100	0.104	0.105	0.105	0.107	0.103	0.105	0.104	0.107	0.106	0.110	0.106	0.112
	30	5	0.010	0.015	0.014	0.028	0.029	0.015	0.014	0.025	0.024	0.014	0.014	0.024	0.019
	30	5	0.050	0.052	0.054	0.060	0.062	0.051	0.052	0.062	0.062	0.048	0.052	0.058	0.059
	30	5	0.100	0.103	0.103	0.098	0.099	0.101	0.101	0.101	0.098	0.099	0.099	0.097	0.098
	30	30	0.010	0.017	0.015	0.019	0.017	0.016	0.015	0.019	0.015	0.013	0.016	0.018	0.012
	30	30	0.050	0.054	0.055	0.058	0.058	0.055	0.055	0.059	0.057	0.056	0.057	0.063	0.056
	30	30	0.100	0.102	0.105	0.105	0.103	0.101	0.099	0.103	0.102	0.104	0.107	0.105	0.105
(iii)	60	5	0.010	0.012	0.012	0.029	0.027	0.014	0.012	0.028	0.024	0.016	0.011	0.023	0.021
	60	5	0.050	0.052	0.052	0.063	0.064	0.050	0.048	0.063	0.059	0.050	0.052	0.059	0.061
	60	5	0.100	0.100	0.101	0.098	0.095	0.098	0.099	0.097	0.099	0.099	0.098	0.100	0.094
	60	30	0.010	0.017	0.015	0.020	0.019	0.016	0.017	0.020	0.017	0.016	0.015	0.019	0.014
	60	30	0.050	0.052	0.053	0.058	0.060	0.055	0.057	0.061	0.059	0.057	0.056	0.062	0.059
	60	30	0.100	0.103	0.106	0.107	0.103	0.102	0.106	0.107	0.105	0.103	0.102	0.106	0.101

As demonstrated in Theorem 1 and 2, the leading term in (1.7) and (1.8) can only measure the linear dependence as $p \wedge q \rightarrow \infty$, therefore we expect the “joint” test based on $dCov_n^2(\mathbf{X}, \mathbf{Y})$ or $hCov_n^2(\mathbf{X}, \mathbf{Y})$ may

fail to capture the non-linear dependence in high dimension. On the other hand, we consider the “marginal” test where we take the sum of pairwise sample distance/Hilbert-Schmidt covariances to measure the low dimensional dependence for all the pairs as the test proposed in Sections 1.2.1 and 1.2.1. The “marginal” test statistic measures the dependence marginally in a low-dimensional fashion so that it can preserve the ability to capture component-wise non-linear dependence. In the following two examples, we demonstrate the superiority of “marginal” tests.

Example 2. Generate i.i.d. samples from the following models for $i = 1, \dots, n$.

- (i) $X_i = (x_{i1}, \dots, x_{ip})^T \sim N(\mathbf{0}, \mathbf{I}_p), Y_i = (y_{i1}, \dots, y_{ip})^T$, where $y_{ij} = x_{ij}^2$ for $j = 1, \dots, p$.
- (ii) Let $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ and $\sigma_{ij} = 0.7^{|i-j|}$, $X_i = (x_{i1}, \dots, x_{ip})^T \sim N(\mathbf{0}, \Sigma), Y_i = (y_{i1}, \dots, y_{ip})^T$, where $y_{ij} = x_{ij}^2$ for $j = 1, \dots, p$.
- (iii) $X_i = (x_{i1}, \dots, x_{ip})^T \sim N(\mathbf{0}, \mathbf{I}_p), Y_i = (y_{i1}, \dots, y_{ip})^T$, where $y_{ij} = \log |x_{ij}|$ for $j = 1, \dots, p$.

Example 3. Generate i.i.d. samples from the following models for $i = 1, \dots, n$.

- (i) Let \circ denotes the Hadamard product, $X_i = (x_{i1}, \dots, x_{ip})^T \stackrel{i.i.d.}{\sim} U(-1, 1), Y_i = X_i \circ X_i$.
- (ii) $X_i = (x_{i1}, \dots, x_{ip})^T \stackrel{i.i.d.}{\sim} U(0, 1), Y_i = 4X_i \circ X_i - 3.6X_i + 0.8$.
- (iii) $Z_i = (z_{i1}, \dots, z_{ip})^T \stackrel{i.i.d.}{\sim} U(0, 2\pi), X_i = \sin(Z_i), Y_i = \cos(Z_i)$.

Notice that in the above two examples, $\text{cov}^2(x_i, y_j) = 0$ but $d\text{Cov}^2(x_i, y_j) \neq 0$ for all (i, j) s, that is, $(X, Y) \in H_{A_s}$. From Table 1.2, we can observe that for Example 2, the “joint” tests suffer substantial power loss as dimension increases for fixed sample size. The power loss is less severe in case (ii) than the ones in cases (i) and (iii), due to the dependence between the components. On the other hand, the powers corresponding to the marginal test statistics consistently outperform their joint counterparts with very little to none power reduction as the dimension increases. Similar phenomenon can be observed for Example 3; see Table 1.3. In addition, for all the cases in both Example 2 and Example 3, the power loss corresponding to Laplacian kernel is consistently less than that for Gaussian kernel. In general, we observe that the tests based on distance covariance, Hilbert-Schmidt covariance with Gaussian kernel, and Hilbert-Schmidt covariance with Laplacian kernel, are all admissible, as none of them dominate the others in all situations. In the following example, we examine the afore-mentioned tests on a real data set.

Example 4. We consider the Earthquake data set, which is originally from the Northern California Earthquake Data Center and has classes of positive and negative major earthquake events. There are 368 negative and 93 positive cases and each data point is of length 512. This data set can be downloaded from UCR Time Series Classification Archive [Dau et al. (2018)]. Here we only consider the negative cases. Let $Z_i = (z_{i,1}, z_{i,2}, \dots, z_{i,512})^T$ denote the record of a negative event, then set $X_i = (z_{i,150-p+1}, \dots, z_{i,150})^T$ and $Y_i = (z_{i,151}, \dots, z_{i,150+p})^T$, $i = 1, \dots, 368$. We apply all tests to test the independence between X_i and Y_i , which are expected to be dependent due to the serial nature of Z_i . For each $p = 5, 30$ and $n = 10, 30, 60$, we randomly sample n rows from the full dataset $(X_i, Y_i)_{i=1}^{368}$ without replacement and apply the afore-mentioned tests based on the subsample. Next, the above procedure is repeated 5000 times to calculate the power.

The results are presented in Table 1.4. It can be seen that the powers of the marginal tests increase as the dimension grows, whereas the powers of all joint tests experience a decay as p grows and are nearly

Table 1.2: Power comparison under H_{A_s} from Example 2

	n	p	α	$dCov$	$mdCov$	T_{dCov}	T_{mdCov}	Gaussian Kernel				Laplacian Kernel			
								$hCov$	$mhCov$	T_{hCov}	T_{mhCov}	$hCov$	$mhCov$	T_{hCov}	T_{mhCov}
(i)	10	5	0.010	0.113	0.285	0.144	0.321	0.110	0.493	0.138	0.516	0.172	0.801	0.226	0.813
	10	5	0.050	0.231	0.495	0.254	0.519	0.236	0.724	0.256	0.736	0.356	0.927	0.398	0.938
	10	5	0.100	0.325	0.618	0.332	0.628	0.325	0.828	0.336	0.834	0.495	0.968	0.506	0.969
	10	30	0.010	0.032	0.286	0.028	0.267	0.032	0.543	0.030	0.513	0.044	0.848	0.041	0.838
	10	30	0.050	0.101	0.526	0.098	0.523	0.098	0.769	0.099	0.763	0.124	0.945	0.128	0.947
	10	30	0.100	0.158	0.669	0.162	0.666	0.160	0.858	0.160	0.858	0.203	0.978	0.205	0.977
	30	5	0.010	0.440	0.997	0.499	0.999	0.518	1	0.583	1	0.924	1	0.956	1
	30	5	0.050	0.651	1.000	0.679	1.000	0.741	1	0.768	1	0.987	1	0.988	1
	30	5	0.100	0.766	1.000	0.773	1	0.836	1	0.845	1	0.994	1	0.995	1
	30	30	0.010	0.084	1.000	0.082	1.000	0.085	1	0.082	1	0.194	1	0.192	1
	30	30	0.050	0.190	1	0.187	1	0.192	1	0.192	1	0.365	1	0.365	1
	30	30	0.100	0.275	1	0.272	1	0.280	1	0.276	1	0.476	1	0.478	1
	60	5	0.010	0.948	1	0.976	1	0.983	1	0.992	1	1	1	1	1
	60	5	0.050	0.994	1	0.996	1	0.998	1	0.999	1	1	1	1	1
	60	5	0.100	0.999	1	0.999	1	1.000	1	1.000	1	1	1	1	1
	60	30	0.010	0.185	1	0.173	1	0.194	1	0.183	1	0.587	1	0.587	1
	60	30	0.050	0.346	1	0.346	1	0.361	1	0.360	1	0.779	1	0.782	1
	60	30	0.100	0.462	1	0.459	1	0.475	1	0.473	1	0.861	1	0.864	1
	10	5	0.010	0.167	0.232	0.237	0.296	0.192	0.347	0.263	0.410	0.279	0.595	0.391	0.652
	10	5	0.050	0.306	0.386	0.341	0.421	0.356	0.570	0.401	0.606	0.525	0.806	0.584	0.832
	10	5	0.100	0.401	0.489	0.409	0.500	0.479	0.699	0.487	0.709	0.674	0.892	0.689	0.901
	10	30	0.010	0.080	0.202	0.091	0.210	0.082	0.376	0.091	0.366	0.099	0.646	0.123	0.634
	10	30	0.050	0.178	0.369	0.191	0.378	0.179	0.605	0.192	0.610	0.229	0.834	0.252	0.837
	10	30	0.100	0.257	0.492	0.259	0.492	0.264	0.728	0.265	0.730	0.342	0.906	0.351	0.909
	30	5	0.010	0.623	0.847	0.781	0.950	0.895	0.999	0.957	1	0.995	1	0.999	1
	30	5	0.050	0.872	0.984	0.902	0.990	0.982	1	0.990	1	1.000	1	1	1
	30	5	0.100	0.940	0.996	0.945	0.995	0.994	1	0.994	1	1	1	1	1
	30	30	0.010	0.251	0.929	0.277	0.944	0.307	1	0.336	1	0.629	1	0.686	1
	30	30	0.050	0.419	0.982	0.434	0.985	0.499	1	0.517	1	0.830	1	0.849	1
	30	30	0.100	0.532	0.995	0.532	0.995	0.613	1	0.622	1	0.905	1	0.909	1
(ii)	60	5	0.010	0.999	1	1	1	1	1	1	1	1	1	1	1
	60	5	0.050	1	1	1	1	1	1	1	1	1	1	1	1
	60	5	0.100	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.010	0.643	1	0.684	1	0.790	1	0.833	1	0.996	1	0.999	1
	60	30	0.050	0.824	1	0.836	1	0.918	1	0.930	1	1.000	1	1.000	1
	60	30	0.100	0.894	1	0.896	1	0.955	1	0.958	1	1	1	1	1
	10	5	0.010	0.043	0.233	0.060	0.257	0.042	0.434	0.053	0.447	0.076	0.768	0.098	0.785
	10	5	0.050	0.121	0.466	0.141	0.490	0.119	0.680	0.137	0.698	0.191	0.924	0.214	0.927
	10	5	0.100	0.201	0.616	0.212	0.624	0.203	0.808	0.210	0.810	0.291	0.963	0.298	0.964
	10	30	0.010	0.017	0.260	0.013	0.242	0.017	0.482	0.012	0.445	0.021	0.830	0.017	0.811
	10	30	0.050	0.062	0.488	0.062	0.487	0.063	0.729	0.062	0.727	0.071	0.941	0.070	0.940
	10	30	0.100	0.120	0.632	0.116	0.630	0.118	0.837	0.115	0.836	0.131	0.972	0.130	0.975
(iii)	30	5	0.010	0.146	0.999	0.191	1	0.153	1	0.187	1	0.464	1	0.529	1
	30	5	0.050	0.346	1	0.375	1	0.347	1	0.380	1	0.723	1	0.747	1
	30	5	0.100	0.484	1	0.497	1	0.496	1	0.501	1	0.835	1	0.840	1
	30	30	0.010	0.024	1.000	0.022	1.000	0.026	1	0.022	1	0.038	1	0.037	1
	30	30	0.050	0.088	1	0.085	1	0.086	1	0.085	1	0.117	1	0.115	1
	30	30	0.100	0.149	1	0.147	1	0.148	1	0.144	1	0.195	1	0.193	1
	60	5	0.010	0.547	1	0.630	1	0.566	1	0.642	1	0.978	1	0.988	1
	60	5	0.050	0.802	1	0.835	1	0.808	1	0.836	1	0.997	1	0.998	1
	60	5	0.100	0.907	1	0.911	1	0.905	1	0.913	1	0.999	1	0.999	1
	60	30	0.010	0.038	1	0.030	1	0.038	1	0.029	1	0.089	1	0.080	1
	60	30	0.050	0.122	1	0.117	1	0.119	1	0.119	1	0.217	1	0.214	1
	60	30	0.100	0.198	1	0.196	1	0.199	1	0.197	1	0.326	1	0.325	1

Table 1.3: Power comparison under H_{A_s} from Example 3

	n	p	α	$dCov$	$mdCov$	T_{dCov}	T_{mdCov}	Gaussian Kernel				Laplacian Kernel			
								$hCov$	$mhCov$	T_{hCov}	T_{mhCov}	$hCov$	$mhCov$	T_{hCov}	T_{mhCov}
(i)	10	5	0.010	0.044	0.196	0.055	0.218	0.042	0.348	0.052	0.367	0.074	0.672	0.098	0.685
	10	5	0.050	0.120	0.390	0.136	0.416	0.114	0.582	0.129	0.604	0.183	0.859	0.209	0.870
	10	5	0.100	0.201	0.542	0.209	0.546	0.191	0.722	0.197	0.731	0.292	0.927	0.304	0.931
	10	30	0.010	0.018	0.212	0.014	0.194	0.017	0.387	0.014	0.362	0.022	0.722	0.017	0.706
	10	30	0.050	0.066	0.434	0.064	0.428	0.066	0.627	0.064	0.625	0.075	0.892	0.077	0.891
	10	30	0.100	0.123	0.571	0.121	0.568	0.123	0.749	0.119	0.750	0.135	0.944	0.132	0.946
	30	5	0.010	0.158	0.988	0.197	0.996	0.136	1	0.163	1	0.486	1	0.555	1
	30	5	0.050	0.341	1.000	0.369	1	0.303	1	0.328	1	0.725	1	0.756	1
	30	5	0.100	0.483	1	0.488	1	0.433	1	0.444	1	0.838	1	0.846	1
	30	30	0.010	0.026	0.996	0.023	0.996	0.027	1.000	0.022	1.000	0.043	1	0.038	1
	30	30	0.050	0.089	1.000	0.084	0.999	0.088	1	0.083	1	0.123	1	0.125	1
	30	30	0.100	0.153	1.000	0.152	1.000	0.151	1	0.152	1	0.209	1	0.204	1
	60	5	0.010	0.559	1	0.637	1	0.461	1	0.539	1	0.989	1	0.996	1
	60	5	0.050	0.816	1	0.847	1	0.738	1	0.774	1	1.000	1	1	1
	60	5	0.100	0.916	1	0.925	1	0.861	1	0.870	1	1	1	1	1
	60	30	0.010	0.037	1	0.032	1	0.036	1	0.031	1	0.091	1	0.085	1
	60	30	0.050	0.125	1	0.119	1	0.122	1	0.115	1	0.231	1	0.228	1
	60	30	0.100	0.208	1	0.207	1	0.204	1	0.202	1	0.350	1	0.346	1
	10	5	0.010	0.044	0.217	0.059	0.242	0.040	0.393	0.055	0.413	0.077	0.713	0.106	0.732
	10	5	0.050	0.124	0.432	0.141	0.453	0.117	0.637	0.131	0.655	0.202	0.886	0.224	0.895
	10	5	0.100	0.210	0.577	0.213	0.583	0.196	0.771	0.204	0.775	0.304	0.942	0.318	0.942
	10	30	0.010	0.020	0.247	0.013	0.224	0.019	0.439	0.013	0.409	0.022	0.774	0.018	0.763
	10	30	0.050	0.064	0.474	0.064	0.474	0.063	0.677	0.063	0.676	0.075	0.913	0.076	0.913
	10	30	0.100	0.126	0.606	0.125	0.604	0.126	0.795	0.126	0.790	0.141	0.956	0.138	0.955
	30	5	0.010	0.178	0.995	0.221	0.999	0.148	1	0.186	1	0.544	1	0.608	1
	30	5	0.050	0.376	1	0.409	1	0.333	1	0.358	1	0.775	1	0.797	1
	30	5	0.100	0.518	1	0.526	1	0.468	1	0.478	1	0.871	1	0.880	1
	30	30	0.010	0.027	0.998	0.023	0.998	0.026	1.000	0.022	1	0.043	1	0.038	1
	30	30	0.050	0.088	1.000	0.087	1.000	0.088	1	0.086	1	0.128	1	0.128	1
	30	30	0.100	0.155	1.000	0.152	1.000	0.154	1	0.152	1	0.218	1	0.213	1
	60	5	0.010	0.632	1	0.709	1	0.526	1	0.609	1	0.995	1	0.999	1
	60	5	0.050	0.870	1	0.895	1	0.792	1	0.826	1	1	1	1	1
	60	5	0.100	0.946	1	0.952	1	0.904	1	0.911	1	1	1	1	1
	60	30	0.010	0.044	1	0.037	1	0.043	1	0.036	1	0.105	1	0.096	1
	60	30	0.050	0.126	1	0.125	1	0.123	1	0.121	1	0.251	1	0.244	1
	60	30	0.100	0.213	1	0.211	1	0.211	1	0.206	1	0.368	1	0.366	1
(ii)	10	5	0.010	0.019	0.024	0.023	0.028	0.017	0.033	0.022	0.040	0.023	0.090	0.029	0.095
	10	5	0.050	0.058	0.079	0.068	0.089	0.057	0.111	0.067	0.115	0.068	0.232	0.081	0.242
	10	5	0.100	0.113	0.148	0.117	0.151	0.114	0.194	0.118	0.196	0.124	0.351	0.129	0.355
	10	30	0.010	0.016	0.026	0.012	0.020	0.016	0.037	0.012	0.030	0.017	0.089	0.013	0.076
	10	30	0.050	0.059	0.086	0.057	0.083	0.060	0.112	0.058	0.105	0.061	0.233	0.060	0.225
	10	30	0.100	0.111	0.156	0.108	0.153	0.112	0.199	0.108	0.193	0.112	0.357	0.109	0.346
	30	5	0.010	0.019	0.051	0.021	0.068	0.017	0.141	0.021	0.170	0.026	0.673	0.032	0.724
	30	5	0.050	0.061	0.166	0.070	0.188	0.058	0.339	0.066	0.360	0.083	0.889	0.091	0.903
	30	5	0.100	0.117	0.283	0.117	0.288	0.117	0.488	0.116	0.497	0.153	0.953	0.153	0.955
	30	30	0.010	0.017	0.074	0.012	0.065	0.017	0.182	0.012	0.165	0.017	0.754	0.012	0.742
	30	30	0.050	0.061	0.202	0.058	0.198	0.061	0.378	0.059	0.373	0.063	0.913	0.061	0.913
	30	30	0.100	0.112	0.309	0.110	0.307	0.113	0.518	0.110	0.517	0.117	0.960	0.114	0.959
	60	5	0.010	0.019	0.174	0.024	0.219	0.017	0.580	0.022	0.666	0.034	1.000	0.041	1
	60	5	0.050	0.066	0.421	0.073	0.458	0.061	0.853	0.069	0.883	0.108	1	0.119	1
	60	5	0.100	0.123	0.600	0.128	0.612	0.119	0.941	0.122	0.949	0.179	1	0.183	1
	60	30	0.010	0.013	0.251	0.009	0.233	0.013	0.680	0.010	0.665	0.014	1.000	0.010	1
	60	30	0.050	0.053	0.485	0.051	0.484	0.052	0.869	0.050	0.871	0.056	1	0.055	1
	60	30	0.100	0.105	0.620	0.101	0.619	0.106	0.930	0.101	0.929	0.107	1	0.106	1
	10	5	0.010	0.019	0.024	0.023	0.028	0.017	0.033	0.022	0.040	0.023	0.090	0.029	0.095
	10	5	0.050	0.058	0.079	0.068	0.089	0.057	0.111	0.067	0.115	0.068	0.232	0.081	0.242
	10	5	0.100	0.113	0.148	0.117	0.151	0.114	0.194	0.118	0.196	0.124	0.351	0.129	0.355
	10	30	0.010	0.016	0.026	0.012	0.020	0.016	0.037	0.012	0.030	0.017	0.089	0.013	0.076
	10	30	0.050	0.059	0.086	0.057	0.083	0.060	0.112	0.058	0.105	0.061	0.233	0.060	0.225
	10	30	0.100	0.111	0.156	0.108	0.153	0.112	0.199	0.108	0.193	0.112	0.357	0.109	0.346
	30	5	0.010	0.019	0.051	0.021	0.068	0.017	0.141	0.021	0.170	0.026	0.673	0.032	0.724
	30	5	0.050	0.061	0.166	0.070	0.188	0.058	0.339	0.066	0.360	0.083	0.889	0.091	0.903
	30	5	0.100	0.117	0.283	0.117	0.288	0.117	0.488	0.116	0.497	0.153	0.953	0.153	0.955
	30	30	0.010	0.017	0.074	0.012	0.065	0.017	0.182	0.012	0.165	0.017	0.754	0.012	0.742
	30	30	0.050	0.061	0.202	0.058	0.198	0.061	0.378	0.059	0.373	0.063	0.913	0.061	0.913
	30	30	0.100	0.112	0.309	0.110	0.307	0.113	0.518	0.110	0.517	0.117	0.960	0.114	0.959
	60	5	0.010	0.019	0.174	0.024	0.219	0.017	0.580	0.022	0.666	0.034	1.000	0.041	1
	60	5	0.050	0.066	0.421	0.073	0.458	0.061	0.853	0.069	0.883	0.108	1	0.119	1
	60	5	0.100	0.123	0.600	0.128	0.612	0.119	0.941	0.122	0.949	0.179	1	0.183	1
	60	30	0.010	0.013	0.251	0.009	0.233	0.013	0.680	0.010	0.665	0.014	1.000	0.010	1
	60	30	0.050	0.053	0.485	0.051	0.484	0.052	0.869	0.050	0.871	0.056	1	0.055	1
	60	30	0.100	0.105	0.620	0.101	0.619	0.106	0.930	0.101	0.929	0.107	1	0.106	1

trivial when $p = 30$. This finding is consistent for all tests including hCov-based ones with Gaussian and Laplacian kernels. In addition, we also note that the marginal tests with Gaussian or Laplacian kernel have consistently higher power as compared to the Euclidean distance based tests.

Table 1.4: Power Comparison on Earthquake data

n	p	α	$dCov$	$mdCov$	T_{dCov}	T_{mdCov}	Gaussian Kernel				Laplacian Kernel			
							$hCov$	$mhCov$	T_{hCov}	T_{mhCov}	$hCov$	$mhCov$	T_{hCov}	T_{mhCov}
10	5	0.010	0.021	0.054	0.041	0.079	0.021	0.851	0.038	0.883	0.035	0.937	0.060	0.952
10	5	0.050	0.070	0.144	0.085	0.160	0.065	0.927	0.080	0.937	0.091	0.975	0.112	0.979
10	5	0.100	0.120	0.235	0.126	0.230	0.114	0.956	0.117	0.959	0.155	0.985	0.160	0.985
10	30	0.010	0.012	0.218	0.013	0.226	0.012	1.000	0.012	1.000	0.013	1	0.016	1
10	30	0.050	0.046	0.412	0.047	0.421	0.046	1.000	0.046	1.000	0.050	1	0.054	1
10	30	0.100	0.095	0.537	0.093	0.541	0.096	1	0.095	1	0.094	1	0.091	1
30	5	0.010	0.034	0.155	0.055	0.196	0.026	1	0.042	1	0.078	1	0.124	1
30	5	0.050	0.106	0.333	0.128	0.356	0.082	1	0.096	1	0.188	1	0.210	1
30	5	0.100	0.177	0.460	0.183	0.461	0.139	1	0.146	1	0.278	1	0.273	1
30	30	0.010	0.009	0.936	0.009	0.941	0.009	1	0.009	1	0.011	1	0.012	1
30	30	0.050	0.040	0.977	0.041	0.976	0.043	1	0.043	1	0.040	1	0.041	1
30	30	0.100	0.081	0.988	0.080	0.988	0.086	1	0.085	1	0.085	1	0.082	1
60	5	0.010	0.060	0.473	0.086	0.549	0.034	1	0.050	1	0.171	1	0.245	1
60	5	0.050	0.147	0.722	0.171	0.749	0.096	1	0.107	1	0.341	1	0.370	1
60	5	0.100	0.240	0.835	0.244	0.838	0.162	1	0.167	1	0.457	1	0.458	1
60	30	0.010	0.006	1	0.006	1	0.006	1	0.006	1	0.008	1	0.008	1
60	30	0.050	0.030	1	0.031	1	0.032	1	0.031	1	0.033	1	0.034	1
60	30	0.100	0.066	1	0.065	1	0.068	1	0.070	1	0.067	1	0.066	1

1.4 Conclusion

In this article, we investigate the behavior of the distance covariance and Hilbert-Schmidt covariance in the high dimensional setting. Somewhat shockingly, we discover that the distance covariance and Hilbert-Schmidt covariance, which are well-known to capture nonlinear dependence in low/fixed dimensional context, can only capture linear componentwise cross-dependence (to the first order) in the high-dimensional environment. We believe that this is a new finding that may have significant implications to the design of tests for independence for high dimensional data. On one hand, we reveal the limitation of distance covariance and variants in the high dimensional context, and suggest to use marginally aggregated (sample) distance covariance as a way out, where the latter targets the low dimensional nonlinear dependence. On the other hand, we speculate whether it is possible to capture all kinds of dependence between high dimensional vectors X and Y , in a limited sample size framework. If the sample size is fixed, we would conjecture that an omnibus test does not exist; If the sample size can grow faster than the dimension, it seems possible but unclear to us how to develop an omnibus test in an asymptotic sense. We hope the results presented in this paper shed some light on the challenges in the high dimensional dependence testing and will motivate more work in this area.

1.5 High Dimension Medium Sample Size

Another type of asymptotics closely related to HDLSS is the high dimension medium sample size (HDMSS) setting [Aoshima et al. (2018)], where $p \wedge q \rightarrow \infty$ and $n \rightarrow \infty$ at a slower rate comparing to p, q . The HDMSS setting has been studied by Fan and Lv (2008) and Yata and Aoshima (2010), among others.

From the previous sections, we know that the distance/Hilbert-Schmidt covariance can only detect linear dependencies between pair-wise components when $p \wedge q \rightarrow \infty$ and n is fixed. In this section, we show that this surprising phenomenon still holds under the high dimension medium sample size setting. Consequently,

a unified approach is used to show that T_R converges in distribution to standard normal under the null hypothesis, but the technical details of handling the leading term and controlling the remainder are totally different from the fixed n case.

1.5.1 Distance Covariance and Variants

We first state the following assumption which can be seen as an extension of Assumption 3.

Assumption 16. *Denote $E[L_X(X, X')^2] = \alpha_p^2$, $E[L_Y(Y, Y')^2] = \beta_q^2$, $E[L_X(X, X')^4] = \gamma_p^2$ and $E[L_Y(Y, Y')^4] = \lambda_q^2$, where $\alpha_p, \beta_q, \gamma_p, \lambda_q$ are sequences of numbers such that as $n \wedge p \wedge q \rightarrow \infty$*

$$n\alpha_p = o(1), \quad n\beta_q = o(1), \\ \tau_X^2(\alpha_p\gamma_p + \gamma_p^2) = o(1), \tau_Y^2(\beta_q\lambda_q + \lambda_q^2) = o(1), \tau(\alpha_p\lambda_q + \gamma_p\beta_q + \gamma_p\lambda_q) = o(1).$$

Remark 17. *For the m -dependence structure, i.e., $x_i \perp x_j$ if $|i - j| > m$ and $y_{i'} \perp y_{j'}$ if $|i' - j'| > m'$, where $\sup_i E(x_i^8) < \infty$ and $\sup_i E(y_i^8) < \infty$, we can show that $\alpha_p = O(\sqrt{m/p})$, $\beta_q = O(\sqrt{m'/q})$, $\gamma_p = O(m/p)$ and $\lambda_q = O(m'/q)$. Thus, Assumption 16 holds under the m -dependence model if n and m, m' satisfies*

$$n^2m = o(p), \quad n^2m' = o(q), \\ m^3 = o(p), \quad m'^3 = o(q), \quad m'm^2 = o(p), \quad mm'^2 = o(q).$$

The following theorem shows that the decomposition property (1.7) for distance covariance still holds under high dimension medium sample size setting.

Theorem 4. *Under Assumption 1, we can show that*

(i)

$$dCov_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) + \mathcal{R}_n. \quad (1.9)$$

Here cov_n^2 is defined the same as in Theorem 1 and \mathcal{R}_n is the remainder term.

(ii) Further suppose Assumption 16 holds. Then we have

$$\frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) = O_p(\tau\alpha_p\beta_q), \\ \mathcal{R}_n = O_p(\tau\alpha_p\lambda_q + \tau\gamma_p\beta_q + \tau\gamma_p\lambda_q) = o_p(1).$$

Similarly, as shown in the following, $hCov$ also has the decomposition property under HDMSS.

Theorem 5. *Under Assumption D1, we have*

(i)

$$\tau \times hCov_n^2(\mathbf{X}, \mathbf{Y}) = f^{(1)} \left(\frac{\tau_X}{\gamma_X} \right) g^{(1)} \left(\frac{\tau_Y}{\gamma_Y} \right) \frac{\tau_X}{\gamma_X} \frac{\tau_Y}{\gamma_Y} \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) + \mathcal{R}_n, \quad (1.10)$$

where cov_n^2 is defined the same as in Theorem 1 and \mathcal{R}_n is the remainder term.

(ii) Further suppose Assumption 16 holds. Then

$$\begin{aligned} f^{(1)}\left(\frac{\tau_X}{\gamma_X}\right) g^{(1)}\left(\frac{\tau_Y}{\gamma_Y}\right) \frac{\tau_X}{\gamma_X} \frac{\tau_Y}{\gamma_Y} &\asymp_p 1, \\ \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q \text{cov}_n^2(\mathcal{X}_i, \mathcal{Y}_j) &= O_p(\tau \alpha_p \beta_q), \\ \mathcal{R}_n &= O_p(\tau \alpha_p \lambda_q + \tau \gamma_p \beta_q + \tau \gamma_p \lambda_q) = o_p(1). \end{aligned}$$

From Equations (1.9) and (1.10), we can see that under the HDMSS setting, it is still true that distance/Hilbert-Schmidt covariance can only detect the linear dependence between the components of X and Y .

1.5.2 Studentized Test Statistics

Similar to Section 1.2.2, we provide a unified approach to analyze the studentized $dCov, hCov, mdCov$. Since now the sample size is growing, the element-wise argument used to prove the results in Section 1.2.2 will no longer work. Inspired by Zhang et al. (2018) and Yao et al. (2018), we derive the asymptotic distribution by constructing a martingale sequence and using martingale CLT.

Unified Approach

For notational convenience, we first define the following metrics,

$$U(X_s, X_t) := \frac{1}{\sqrt{p}} \sum_{i=1}^p k_{st}(i), \quad V(Y_s, Y_t) := \frac{1}{\sqrt{q}} \sum_{i=1}^q l_{st}(i),$$

where $k_{st}(i)$ and $l_{st}(i)$ are defined in Section 1.2.2. To show that the studentized test statistic converges to standard normal, we essentially use the martingale CLT [Hall and Heyde (2014)] and the following assumptions are used to guarantee the conditions in martingale CLT.

Assumption 18. $D5$

$$\frac{E[U(X, X')^4]}{\sqrt{n}(E[U(X, X')^2])^2} \rightarrow 0, \quad (1.11)$$

$$\frac{E[U(X, X')U(X', X'')U(X'', X''')U(X''', X)]}{(E[U(X, X')^2])^2} \rightarrow 0, \quad (1.12)$$

and similar assumptions hold for Y .

Remark 19. When $k(x, y) = l(x, y) = |x - y|$, Assumption 18 has been studied in Propositions 2.1 and 2.2 of Zhang et al. (2018).

Remark 20. When $k(x, y) = l(x, y) = |x - y|^2$, Equations (1.11) and (1.12) can be simplified to

$$\begin{aligned} \frac{\sum_{i,j,r,w=1}^p E^2[(x_i - E[x_i])(x_j - E[x_j])(x_r - E[x_r])(x_w - E[x_w])]}{\sqrt{n} \text{Tr}^2(\Sigma_X^2)} &\rightarrow 0, \\ \frac{\text{Tr}(\Sigma_X^4)}{\text{Tr}^2(\Sigma_X^2)} &\rightarrow 0, \quad \text{where } \Sigma_X = \text{cov}(X, X). \end{aligned}$$

Notice that $\text{Tr}(\Sigma_X^2) = \sum_{i=1}^p \sum_{j=1}^p \text{cov}^2(x_i, x_j)$. Consider the m -dependence model in Remark 17. Assuming $\sup_i E(x_i^4) < \infty$, we have $\text{Tr}(\Sigma_X^4) = O(m^3 p)$ and

$$\sum_{i,j,r,w=1}^p E^2 [(x_i - E[x_i])(x_j - E[x_j])(x_r - E[x_r])(x_w - E[x_w])] = O(m^2 p^2).$$

Consequently, it can be seen that the m -dependence model in Remark 17 also satisfies Equations (1.11) and (1.12) by controlling the orders of n, m, m' .

Then, we can show that the normalized $uCov_n^2(\mathbf{X}, \mathbf{Y})$ converges to standard normal distribution under the high dimension medium sample size regime.

Theorem 6. *Let $n \wedge p \wedge q \rightarrow \infty$. Under H_0 and Assumption 18, we have*

$$\sqrt{\binom{n}{2}} \frac{uCov_n^2(\mathbf{X}, \mathbf{Y})}{\mathcal{S}} \xrightarrow{d} N(0, 1), \text{ where } \mathcal{S}^2 = E[U(X, X')^2]E[V(Y, Y')^2].$$

Consequently, we have the following result.

Proposition 21. *Let $n \wedge p \wedge q \rightarrow \infty$. Under H_0 and Assumption 18, we have*

$$T_u \xrightarrow{d} N(0, 1).$$

Studentized Tests

The following result shows that as $n \wedge p \wedge q \rightarrow \infty$, scaled $dCov$, $hCov$ and $mdCov$ are all equal to $uCov$ up to an asymptotically constant factor.

Proposition 22. *Under Assumption 1,*

(i) *When $k(x, y) = l(x, y) = |x - y|^2$,*

$$\begin{aligned} dCov_n^2(\mathbf{X}, \mathbf{Y}) &= \frac{1}{4} \frac{\sqrt{pq}}{\tau} uCov_n^2(\mathbf{X}, \mathbf{Y}) + \mathcal{R}'_n, \\ \tau \times hCov_n^2(\mathbf{X}, \mathbf{Y}) &= \frac{\sqrt{pq}}{4\gamma_{\mathbf{X}}\gamma_{\mathbf{Y}}} f^{(1)}\left(\frac{\tau_{\mathbf{X}}}{\gamma_{\mathbf{X}}}\right) g^{(1)}\left(\frac{\tau_{\mathbf{Y}}}{\gamma_{\mathbf{Y}}}\right) uCov_n^2(\mathbf{X}, \mathbf{Y}) + \mathcal{R}''_n, \end{aligned}$$

where $\mathcal{R}'_n, \mathcal{R}''_n$ are the remainder terms. Further suppose Assumption 16 holds. Then

$$\begin{aligned} uCov_n^2(\mathbf{X}, \mathbf{Y}) &= O_p(\tau \alpha_p \beta_q), \\ \mathcal{R}'_n &= O_p(\tau \alpha_p \lambda_q + \tau \gamma_p \beta_q + \tau \gamma_p \lambda_q) = o_p(1), \\ \mathcal{R}''_n &= O_p(\tau \alpha_p \lambda_q + \tau \gamma_p \beta_q + \tau \gamma_p \lambda_q) = o_p(1). \end{aligned}$$

(ii) *When $k(x, y) = l(x, y) = |x - y|$,*

$$mdCov_n^2(\mathbf{X}, \mathbf{Y}) = \sqrt{pq} \sqrt{\binom{n}{2}} uCov_n^2(\mathbf{X}, \mathbf{Y}).$$

Finally, by adopting a unified approach, we have the following Corollary.

Corollary 2. Let $n \wedge p \wedge q \rightarrow \infty$. Under H_0 and Assumption 18, we have

(i)

$$T_{mdCov} \xrightarrow{d} N(0, 1).$$

(ii) Further suppose Assumption 16 and

$$\frac{n}{\sqrt{\frac{1}{p} \text{Tr}(\Sigma_X^2) \frac{1}{q} \text{Tr}(\Sigma_Y^2)}} \tau(\alpha_p \lambda_q + \gamma_p \beta_q + \gamma_p \lambda_q) = o(1). \quad (1.13)$$

Then, for each $R \in \{dCov, hCov\}$, we have

$$T_R \xrightarrow{d} N(0, 1).$$

Remark 23. The m -dependence model in Remark 17 can also satisfies Equation (1.13) by controlling the orders of n, m, m' based on the magnitude of $\text{Tr}(\Sigma_X^2)/p$ and $\text{Tr}(\Sigma_Y^2)/q$.

1.6 Additional Simulation Examples and Comparisons

The asymptotic validity of our t -tests depend on the weak dependence assumption among the components of X and Y . In the following, we conduct some sensitivity analysis and examine a few cases where the weak dependence assumption is violated.

Example 5. Generate i.i.d. samples from the following models for $i = 1, \dots, n$.

(i) Let $ARFIMA(\phi, d, 0)$ denotes the autoregressive fractionally integrated moving average (ARFIMA) model with autoregressive order 1 parameter ϕ , moving average parameter 0 and fractional differencing parameter d , $X_i = (x_{i1}, \dots, x_{ip})^T \sim ARFIMA(0.5, d, 0)$ and $Y_i = (y_{i1}, \dots, y_{ip})^T \sim ARFIMA(0.5, d, 0)$.

(ii) We consider the compound symmetric covariance structure, i.e., let

$$\Sigma_d = (0.5 + 0.5\mathbb{I}_{\{i=j\}})_{i,j=1}^d \quad \text{and} \quad \Sigma = \begin{pmatrix} \mathbf{I}_{\lfloor p/d \rfloor} \otimes \Sigma_d & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p - \lfloor p/d \rfloor \times d} \end{pmatrix},$$

where \otimes is the Kronecker product. Then, $X_i = (x_{i1}, \dots, x_{ip})^T \sim N(\mathbf{0}, \Sigma)$ and $Y_i = (y_{i1}, \dots, y_{ip})^T \sim N(\mathbf{0}, \Sigma)$.

(iii) Let $X_i = (x_{i1}, \dots, x_{ip})^T \sim N(\mathbf{0}, \Sigma)$ and $Y_i = (y_{i1}, \dots, y_{ip})^T \sim N(\mathbf{0}, \Sigma)$, where

$$\Sigma_d = (0.5 + 0.5\mathbb{I}_{\{i=j\}})_{i,j=1}^d \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_d & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-d} \end{pmatrix}.$$

In Example 5 (i), as d increases from 0 to 0.45, the dependence among components become stronger. The weak dependence assumptions, i.e. Assumptions 3 and 7, are not expected to hold, thus the limiting null distribution of our test statistics may not be t -distribution. The results in Table 1.5 show there is a mild size distortion and the approximation by t -distribution might still work in this case. For Example 5 (ii) and

(iii), the larger d is, the more dependence there are in the components. For relatively large d , i.e., $d = 30$ in Example 5 (ii) & (iii), some of the t -tests have size over 0.07, showing the impact of strong componentwise dependence. In general, the magnitude of distortion is moderate under all examples in 5.

Table 1.5: Size comparison from Example 5

	d	n	$dCov$	$mdCov$	T_{dCov}	T_{mdCov}	Gaussian Kernel				Laplacian Kernel			
							$hCov$	$mhCov$	T_{hCov}	T_{mhCov}	$hCov$	$mhCov$	T_{hCov}	T_{mhCov}
(i)	0	10	0.057	0.059	0.059	0.058	0.058	0.053	0.059	0.051	0.056	0.055	0.058	0.051
	0.050	10	0.057	0.058	0.058	0.059	0.056	0.053	0.059	0.054	0.055	0.054	0.057	0.051
	0.100	10	0.055	0.058	0.058	0.060	0.056	0.053	0.058	0.055	0.055	0.052	0.057	0.049
	0.150	10	0.056	0.057	0.058	0.059	0.055	0.054	0.057	0.051	0.053	0.051	0.056	0.050
	0.200	10	0.054	0.055	0.058	0.062	0.055	0.052	0.057	0.055	0.053	0.052	0.057	0.054
	0.250	10	0.053	0.055	0.059	0.059	0.053	0.052	0.057	0.054	0.054	0.053	0.059	0.058
	0.350	10	0.054	0.055	0.063	0.061	0.053	0.056	0.063	0.057	0.053	0.058	0.062	0.061
	0.450	10	0.058	0.054	0.066	0.060	0.056	0.053	0.066	0.059	0.056	0.052	0.061	0.058
	0	30	0.055	0.052	0.058	0.052	0.055	0.053	0.059	0.054	0.058	0.049	0.061	0.049
	0.050	30	0.054	0.051	0.059	0.051	0.054	0.052	0.058	0.052	0.058	0.052	0.059	0.049
	0.100	30	0.053	0.049	0.060	0.052	0.055	0.052	0.059	0.051	0.056	0.049	0.059	0.047
	0.150	30	0.055	0.049	0.059	0.055	0.055	0.050	0.059	0.054	0.055	0.048	0.059	0.049
	0.200	30	0.055	0.050	0.058	0.057	0.055	0.048	0.058	0.054	0.054	0.051	0.058	0.048
	0.250	30	0.054	0.051	0.060	0.059	0.053	0.052	0.058	0.054	0.054	0.056	0.059	0.057
	0.350	30	0.052	0.052	0.057	0.058	0.053	0.051	0.059	0.056	0.054	0.057	0.062	0.058
	0.450	30	0.053	0.052	0.062	0.061	0.052	0.055	0.061	0.059	0.054	0.057	0.060	0.059
(ii)	2	10	0.056	0.056	0.054	0.056	0.055	0.058	0.053	0.056	0.057	0.056	0.053	0.051
	3	10	0.055	0.056	0.056	0.054	0.055	0.052	0.057	0.051	0.056	0.055	0.056	0.054
	5	10	0.060	0.058	0.063	0.059	0.060	0.052	0.064	0.052	0.059	0.054	0.062	0.051
	6	10	0.058	0.052	0.062	0.057	0.055	0.059	0.061	0.059	0.056	0.059	0.062	0.058
	10	10	0.055	0.055	0.064	0.064	0.057	0.059	0.065	0.061	0.058	0.060	0.064	0.060
	15	10	0.054	0.051	0.065	0.062	0.054	0.050	0.065	0.059	0.055	0.050	0.065	0.057
	30	10	0.056	0.053	0.075	0.072	0.056	0.054	0.074	0.069	0.053	0.055	0.073	0.068
	2	30	0.056	0.054	0.055	0.054	0.056	0.059	0.055	0.057	0.056	0.057	0.055	0.055
	3	30	0.056	0.057	0.057	0.058	0.056	0.053	0.057	0.053	0.056	0.055	0.059	0.052
	5	30	0.055	0.056	0.059	0.055	0.057	0.054	0.059	0.055	0.059	0.050	0.061	0.050
	6	30	0.051	0.053	0.056	0.054	0.052	0.051	0.055	0.051	0.053	0.051	0.056	0.051
	10	30	0.057	0.052	0.066	0.060	0.057	0.051	0.065	0.060	0.056	0.053	0.064	0.054
	15	30	0.050	0.051	0.064	0.064	0.050	0.051	0.063	0.059	0.054	0.050	0.064	0.059
	30	30	0.053	0.054	0.067	0.069	0.054	0.053	0.068	0.065	0.052	0.052	0.065	0.062
(iii)	2	10	0.057	0.056	0.056	0.055	0.057	0.058	0.055	0.056	0.055	0.055	0.055	0.051
	6	10	0.058	0.055	0.057	0.054	0.057	0.054	0.057	0.052	0.057	0.052	0.058	0.049
	10	10	0.057	0.060	0.062	0.061	0.057	0.058	0.062	0.059	0.057	0.057	0.061	0.059
	14	10	0.060	0.056	0.063	0.061	0.060	0.058	0.062	0.061	0.056	0.059	0.062	0.056
	18	10	0.055	0.053	0.067	0.064	0.056	0.053	0.067	0.060	0.059	0.052	0.068	0.057
	24	10	0.051	0.052	0.067	0.065	0.051	0.053	0.065	0.063	0.054	0.053	0.068	0.062
	30	10	0.056	0.053	0.075	0.072	0.056	0.054	0.074	0.069	0.053	0.055	0.073	0.068
	2	30	0.049	0.054	0.048	0.052	0.050	0.056	0.048	0.053	0.050	0.056	0.049	0.053
	6	30	0.049	0.050	0.048	0.050	0.048	0.055	0.047	0.055	0.048	0.057	0.049	0.055
	10	30	0.050	0.053	0.052	0.057	0.049	0.052	0.051	0.054	0.047	0.053	0.050	0.051
	14	30	0.052	0.054	0.058	0.059	0.049	0.050	0.056	0.059	0.049	0.053	0.056	0.054
	18	30	0.053	0.055	0.062	0.061	0.054	0.053	0.062	0.058	0.054	0.051	0.059	0.054
	24	30	0.044	0.046	0.055	0.060	0.046	0.048	0.055	0.060	0.046	0.047	0.056	0.055
	30	30	0.053	0.054	0.067	0.069	0.054	0.053	0.068	0.065	0.052	0.052	0.065	0.062

Notice that under the high dimensional case, the “joint” tests can be seen as the aggregation of component-wise sample squared covariances. On the other hand, the “marginal” tests are the accumulation of component-wise sample distance/Hilbert-Schmidt covariances. When (X, Y) are generated from the model in Proposition 14, it is expected that there is power loss for $mdCov$ and $mhCov$ based permutation test comparing to $dCov$ and $hCov$ based permutation tests and similar phenomenon is expected for $mdCov$ and $mhCov$ based t -tests comparing to $dCov$ and $hCov$ based t -tests. The following example demonstrates this phenomenon.

Example 6. Generate i.i.d. samples from the following models for $i = 1, \dots, n$.

(i) Let $\rho = 0.5$,

$$\begin{aligned} Z_i &= (z_{i1}, \dots, z_{ip}) \sim N(\mathbf{0}, \mathbf{I}_p), \\ X_i &= (x_{i1}, \dots, x_{ip}) \sim N(\mathbf{0}, \mathbf{I}_p), \\ Y_i &= \frac{\rho X_i + (1-\rho)Z_i}{\sqrt{\rho^2 + (1-\rho)^2}}. \end{aligned}$$

(ii) Let $\rho = 0.7$ and (X_i, Y_i, Z_i) be defined in the same way as in (i).

(iii) Let $\rho = 0.5$ and \otimes denote the Kronecker product. Define

$$\begin{aligned} Z_i &= (z_{i1}, \dots, z_{ip}) \sim N(\mathbf{0}, \mathbf{I}_p), \\ X_i &= (x_{i1}, \dots, x_{ip}) \sim N(\mathbf{0}, \mathbf{I}_p), \\ Y_i &= \frac{\rho \Sigma X_i + (1-\rho)Z_i}{\sqrt{\rho^2 + (1-\rho)^2}}, \end{aligned}$$

where $\Sigma = \mathbf{I} \otimes \mathbf{A}$ and \mathbf{A} is an orthogonal matrix defined as

$$\mathbf{A} = \begin{pmatrix} 0 & \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{5}} & -\sqrt{\frac{1}{4}} & -\sqrt{\frac{3}{10}} \\ \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{5}} & \sqrt{\frac{1}{4}} & \sqrt{\frac{2}{15}} \\ -\sqrt{\frac{2}{3}} & 0 & \sqrt{\frac{1}{5}} & 0 & \sqrt{\frac{2}{15}} \\ \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{4}} & \sqrt{\frac{1}{5}} & -\sqrt{\frac{1}{4}} & \sqrt{\frac{2}{15}} \\ 0 & -\sqrt{\frac{1}{4}} & \sqrt{\frac{1}{5}} & \sqrt{\frac{1}{4}} & -\sqrt{\frac{3}{10}} \end{pmatrix}.$$

From Table 2.5, we can see that there is indeed a power loss for the “marginal” tests compared to the “joint” tests, but the loss of power appears fairly moderate, which is consistent with our theory. For Example 6, it can also be observed that the power decrease for the Hilbert-Schmidt covariance based tests is a bit more than the power decrease of distance covariance based tests. Moreover, the power drop is slightly smaller for Gaussian kernel comparing with Laplacian kernel.

We also compare the marginal distance/Hilbert-Schmidt covariance based statistics with five recently proposed nonparametric tests, namely Heller-Heller-Gorfine test (HHG) [Heller et al. (2012)], Projection Correlation (PCOR) [Zhu et al. (2017)], Multiscale Graph Correlation (MGC) [Shen et al. (2018)], Kendall’s tau (R_τ) [Han et al. (2017)] and Spearman’s rho (R_ρ) [Han et al. (2017)]. We want to remark that Han et al. (2017) aim to test the mutual independence among the components of X based on a random sample $\{X_i\}_{i=1}^n$, whereas we test the independence of X and Y based on paired high dimensional samples $\{(X_i, Y_i)\}_{i=1}^n$. Thus, the asymptotic theory in Han et al. (2017) is not directly applicable to our setting. To circumvent the difficulty, we shall apply a simple permutation procedure to the following test statistics R_τ and R_ρ , where

$$R_\rho = \max_{j=1}^p \max_{k=1}^q \rho_{j,k}^2 \text{ and } R_\tau = \max_{j=1}^p \max_{k=1}^q \tau_{j,k}^2,$$

where $\rho_{j,k}$ is the Spearman’s ρ and $\tau_{j,k}$ is the Kendall’s tau between component samples \mathcal{X}_j and \mathcal{Y}_k respectively. Let Q_{ni}^j and Q_{ni}^k be the ranks of $x_{i,j}$ and $x_{i,k}$ among $\{x_{1,j}, \dots, x_{n,j}\}$ and $\{x_{1,k}, \dots, x_{n,k}\}$ respectively,

Table 1.6: Power comparison from Example 6

	n	p	α	$dCov$	$mdCov$	T_{dCov}	T_{mdCov}	Gaussian Kernel				Laplacian Kernel			
								$hCov$	$mhCov$	T_{hCov}	T_{mhCov}	$hCov$	$mhCov$	T_{hCov}	T_{mhCov}
(i)	10	5	0.010	0.635	0.560	0.691	0.597	0.629	0.371	0.685	0.392	0.516	0.237	0.585	0.246
	10	5	0.050	0.833	0.774	0.855	0.792	0.825	0.598	0.849	0.610	0.741	0.450	0.772	0.458
	10	5	0.100	0.910	0.861	0.914	0.867	0.906	0.717	0.912	0.721	0.839	0.581	0.851	0.586
	10	30	0.010	0.795	0.654	0.788	0.634	0.796	0.410	0.787	0.379	0.769	0.247	0.762	0.219
	10	30	0.050	0.936	0.849	0.937	0.851	0.935	0.648	0.937	0.644	0.921	0.468	0.924	0.460
	10	30	0.100	0.970	0.914	0.970	0.916	0.970	0.767	0.970	0.768	0.963	0.604	0.964	0.603
	30	5	0.010	1	1	1	1	1	0.999	1	0.998	1	0.980	1	0.982
	30	5	0.050	1	1	1	1	1	1.000	1	1.000	1	0.996	1	0.996
	30	5	0.100	1	1	1	1	1	1.000	1	1.000	1	0.998	1	0.998
	30	30	0.010	1	1	1	1	1	1	1	1	1	0.996	1	0.996
	30	30	0.050	1	1	1	1	1	1	1	1	1	0.999	1	0.999
	30	30	0.100	1	1	1	1	1	1	1	1	1	1.000	1	1.000
	60	5	0.010	1	1	1	1	1	1	1	1	1	1	1	1
	60	5	0.050	1	1	1	1	1	1	1	1	1	1	1	1
	60	5	0.100	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.010	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.050	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.100	1	1	1	1	1	1	1	1	1	1	1	1
	10	5	0.010	1.000	0.999	1.000	0.999	1.000	0.986	1.000	0.989	0.997	0.935	0.999	0.942
	10	5	0.050	1.000	1.000	1.000	1.000	1.000	0.997	1.000	0.997	0.999	0.983	1.000	0.983
	10	5	0.100	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.992	1.000	0.993
	10	30	0.010	1	1	1	1.000	1	0.998	1	0.998	1	0.973	1	0.970
	10	30	0.050	1	1	1	1	1	1.000	1	1.000	1	0.995	1	0.995
	10	30	0.100	1	1	1	1	1	1.000	1	1.000	1	0.997	1	0.997
(ii)	30	5	0.010	1	1	1	1	1	1	1	1	1	1	1	1
	30	5	0.050	1	1	1	1	1	1	1	1	1	1	1	1
	30	5	0.100	1	1	1	1	1	1	1	1	1	1	1	1
	30	30	0.010	1	1	1	1	1	1	1	1	1	1	1	1
	30	30	0.050	1	1	1	1	1	1	1	1	1	1	1	1
	30	30	0.100	1	1	1	1	1	1	1	1	1	1	1	1
	60	5	0.010	1	1	1	1	1	1	1	1	1	1	1	1
	60	5	0.050	1	1	1	1	1	1	1	1	1	1	1	1
	60	5	0.100	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.010	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.050	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.100	1	1	1	1	1	1	1	1	1	1	1	1
	10	5	0.010	0.635	0.497	0.685	0.537	0.633	0.238	0.681	0.260	0.525	0.138	0.584	0.135
	10	5	0.050	0.831	0.728	0.848	0.748	0.824	0.460	0.844	0.477	0.740	0.311	0.768	0.323
	10	5	0.100	0.903	0.830	0.911	0.835	0.899	0.597	0.905	0.604	0.835	0.440	0.844	0.446
	10	30	0.010	0.790	0.583	0.784	0.555	0.789	0.273	0.785	0.247	0.763	0.147	0.761	0.122
	10	30	0.050	0.928	0.800	0.930	0.797	0.928	0.490	0.930	0.486	0.915	0.331	0.919	0.324
	10	30	0.100	0.966	0.888	0.964	0.889	0.965	0.628	0.964	0.626	0.960	0.460	0.957	0.453
	30	5	0.010	1	1.000	1	1	1	0.985	1	0.989	1.000	0.890	1.000	0.898
	30	5	0.050	1	1	1	1	1	0.996	1	0.997	1	0.971	1	0.971
	30	5	0.100	1	1	1	1	1	0.999	1	0.999	1	0.984	1	0.984
	30	30	0.010	1	1	1	1	1	0.998	1	0.999	1	0.950	1	0.948
	30	30	0.050	1	1	1	1	1	1.000	1	1.000	1	0.990	1	0.990
	30	30	0.100	1	1	1	1	1	1.000	1	1.000	1	0.997	1	0.997
(iii)	60	5	0.010	1	1	1	1	1	1	1	1	1	1.000	1	1
	60	5	0.050	1	1	1	1	1	1	1	1	1	1	1	1
	60	5	0.100	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.010	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.050	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.100	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.010	1	1	1	1	1	1	1	1	1	1	1	1
	60	30	0.100	1	1	1	1	1	1	1	1	1	1	1	1

$\rho_{j,k}$ and $\tau_{j,k}$ are formally defined as

$$\rho_{j,k} = \frac{\sum_{i=1}^n (Q_{ni}^j - \bar{Q}_n^j)(Q_{ni}^k - \bar{Q}_n^k)}{\{\sum_{i=1}^n (Q_{ni}^j - \bar{Q}_n^j)^2 \sum_{i=1}^n (Q_{ni}^k - \bar{Q}_n^k)^2\}^{1/2}},$$

$$\tau_{j,k} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \text{sign}(x_{i',j} - x_{i,j}) \text{sign}(y_{i',k} - y_{i,k}).$$

From Table 1.7, the sizes for all three tests appear very accurate due to the use of permutation-based critical values except for the two rank-correlation based tests at small sample size $n = 10$. Comparing Table 2.5 to Table 1.8, we can see that the performance of dCov/hCov based tests (except for Laplacian kernel) are comparable to PCOR and MGC, which outperform all other tests (HHG, R_τ , R_ρ). The power of R_τ and R_ρ are almost trivial in the case of Example 6 (iii).

A comparison of Table 1.9 with Table 1.2 (Table 1.10 with Table 1.3) shows that the *mdCov*/*mhCov* based statistics have superior power compared with HHG, PCOR, MGC, R_ρ and R_τ for Example 2 (iii) and Example 3 (i) & (ii). In addition, HHG, PCOR and MGC all experience a power drop as the dimension grows. For Example 2 (i) & (ii), MGC and PCOR have a noticeable power drop for higher dimensional case, while HHG performs comparably with marginal distance/Hilbert-Schmidt covariance based statistics. Next, for Example 3 (iii), *mhCov* based statistics have better power than HHG, PCOR and MGC, while the performance of *mdCov* based statistics and HHG are similar. In addition, for Example 2 (i) & (ii) and Example 3 (iii), R_ρ and R_τ experience low power.

Finally, the experimental results of these five tests on the Earthquake data are shown in Table 1.11, from which we can observe that HHG, PCOR and MGC all suffer substantial power drop as the dimension increases, while the tests R_ρ and R_τ exhibit higher power with increasing dimension. Overall, HHG, PCOR and MGC exhibit similar phenomenon as what we observed for distance/Hilbert-Schmidt covariance applied to the whole components jointly and their ability of detecting nonlinear dependence gets compromised as the dimension grows with a fixed sample size. As for R_ρ and R_τ , although they do not seem to suffer from high dimensionality, these two tests have little power for detecting some none-monotone dependence due to their rank-based nature. The class of kernels we consider, however, does not include any of these five tests, so a rigorous theoretical justification for these five tests would be interesting and merits future investigation.

Table 1.7: Size comparison from Example 1

n	p	α	(i)					(ii)					(iii)				
			HHG	PCOR	MGC	R_ρ	R_τ	HHG	PCOR	MGC	R_ρ	R_τ	HHG	PCOR	MGC	R_ρ	R_τ
10	5	0.010	0.008	0.012	0.008	0.014	0.011	0.008	0.010	0.009	0.012	0.010	0.008	0.009	0.008	0.014	0.011
10	5	0.050	0.048	0.052	0.054	0.051	0.039	0.052	0.048	0.053	0.045	0.04	0.044	0.050	0.051	0.049	0.041
10	5	0.100	0.095	0.104	0.099	0.092	0.074	0.104	0.100	0.097	0.089	0.072	0.095	0.106	0.099	0.096	0.079
10	30	0.010	0.011	0.011	0.010	0.007	0.005	0.009	0.011	0.010	0.006	0.005	0.010	0.008	0.007	0.009	0.007
10	30	0.050	0.047	0.049	0.053	0.036	0.030	0.045	0.056	0.058	0.039	0.033	0.050	0.046	0.049	0.037	0.034
10	30	0.100	0.098	0.099	0.096	0.075	0.047	0.096	0.109	0.105	0.079	0.051	0.100	0.102	0.094	0.078	0.053
30	5	0.010	0.007	0.010	0.010	0.013	0.012	0.008	0.005	0.008	0.015	0.016	0.008	0.009	0.009	0.016	0.013
30	5	0.050	0.044	0.048	0.055	0.054	0.048	0.048	0.041	0.048	0.055	0.052	0.044	0.045	0.051	0.055	0.052
30	5	0.100	0.098	0.104	0.099	0.105	0.102	0.095	0.099	0.090	0.102	0.097	0.095	0.100	0.095	0.105	0.107
30	30	0.010	0.009	0.008	0.007	0.013	0.010	0.008	0.011	0.011	0.016	0.016	0.011	0.010	0.011	0.016	0.013
30	30	0.050	0.047	0.048	0.054	0.050	0.050	0.042	0.057	0.061	0.056	0.053	0.049	0.051	0.058	0.053	0.051
30	30	0.100	0.096	0.099	0.091	0.097	0.097	0.095	0.110	0.104	0.101	0.101	0.095	0.103	0.099	0.104	0.098
60	5	0.010	0.008	0.011	0.010	0.016	0.016	0.011	0.009	0.009	0.017	0.016	0.011	0.008	0.009	0.016	0.015
60	5	0.050	0.050	0.047	0.051	0.056	0.056	0.054	0.048	0.052	0.056	0.056	0.046	0.047	0.051	0.054	0.055
60	5	0.100	0.101	0.105	0.100	0.107	0.102	0.112	0.101	0.098	0.103	0.101	0.095	0.098	0.093	0.101	0.102
60	30	0.010	0.011	0.013	0.007	0.016	0.016	0.009	0.008	0.012	0.014	0.015	0.008	0.012	0.010	0.014	0.014
60	30	0.050	0.048	0.055	0.057	0.054	0.056	0.047	0.049	0.056	0.055	0.052	0.047	0.054	0.056	0.057	0.055
60	30	0.100	0.092	0.112	0.104	0.106	0.105	0.100	0.108	0.106	0.105	0.102	0.096	0.101	0.097	0.107	0.11

Table 1.8: Power comparison from Example 6

n	p	α	(i)					(ii)					(iii)				
			HHG	PCOR	MGC	R_ρ	R_τ	HHG	PCOR	MGC	R_ρ	R_τ	HHG	PCOR	MGC	R_ρ	R_τ
10	5	0.010	0.350	0.614	0.266	0.311	0.27	0.994	1.000	0.711	0.945	0.928	0.358	0.613	0.277	0.062	0.049
10	5	0.050	0.580	0.844	0.847	0.622	0.572	0.999	1.000	0.987	0.998	0.995	0.573	0.840	0.842	0.199	0.161
10	5	0.100	0.681	0.917	0.929	0.787	0.742	1.000	1	0.998	1	1	0.674	0.916	0.922	0.334	0.288
10	30	0.010	0.395	0.749	0.341	0.114	0.098	0.998	1	0.674	0.850	0.808	0.367	0.751	0.348	0.017	0.013
10	30	0.050	0.622	0.934	0.917	0.336	0.302	0.999	1	0.968	0.994	0.984	0.615	0.925	0.916	0.068	0.060
10	30	0.100	0.724	0.973	0.974	0.516	0.406	1	1	0.991	0.999	0.998	0.717	0.967	0.967	0.138	0.094
30	5	0.010	0.999	1	0.992	1	1	1	1	1	1	1	0.998	1	0.994	0.438	0.429
30	5	0.050	1	1	1	1	1	1	1	1	1	1	1.000	1	1	0.771	0.756
30	5	0.100	1	1	1	1	1	1	1	1	1	1	1	1	1	0.902	0.893
30	30	0.010	1	1	0.966	1	1	1	1	0.997	1	1	1	1	0.967	0.268	0.257
30	30	0.050	1	1	1	1	1	1	1	1.000	1	1	1	1	1	0.563	0.553
30	30	0.100	1	1	1	1	1	1	1	1	1	1	1	1	1	0.748	0.734
60	5	0.010	1	1	1	1	1	1	1	1	1	1	1	1	1	0.927	0.924
60	5	0.050	1	1	1	1	1	1	1	1	1	1	1	1	1	0.995	0.996
60	5	0.100	1	1	1	1	1	1	1	1	1	1	1	1	1	0.999	1
60	30	0.010	1	1	1	1	1	1	1	1	1	1	1	1	1	0.893	0.899
60	30	0.050	1	1	1	1	1	1	1	1	1	1	1	1	1	0.994	0.992
60	30	0.100	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.999

Table 1.9: Power comparison from Example 2

n	p	α	(i)					(ii)					(iii)				
			HHG	PCOR	MGC	R_ρ	R_τ	HHG	PCOR	MGC	R_ρ	R_τ	HHG	PCOR	MGC	R_ρ	R_τ
10	5	0.010	0.403	0.075	0.200	0.076	0.092	0.591	0.107	0.256	0.068	0.080	0.032	0.027	0.052	0.076	0.092
10	5	0.050	0.643	0.187	0.430	0.132	0.169	0.814	0.240	0.525	0.128	0.153	0.113	0.108	0.141	0.132	0.169
10	5	0.100	0.742	0.280	0.516	0.200	0.238	0.879	0.346	0.606	0.190	0.211	0.191	0.188	0.208	0.200	0.238
10	30	0.010	0.185	0.021	0.072	0.124	0.135	0.377	0.044	0.140	0.112	0.117	0.011	0.009	0.013	0.124	0.135
10	30	0.050	0.423	0.082	0.265	0.218	0.236	0.627	0.139	0.391	0.186	0.196	0.052	0.055	0.067	0.218	0.236
10	30	0.100	0.534	0.146	0.354	0.275	0.292	0.730	0.216	0.491	0.235	0.239	0.106	0.115	0.118	0.275	0.292
30	5	0.010	1	0.266	0.961	0.056	0.133	1	0.433	0.992	0.047	0.107	0.282	0.080	0.613	0.056	0.133
30	5	0.050	1	0.510	0.999	0.127	0.230	1	0.774	1	0.110	0.185	0.587	0.266	0.791	0.127	0.230
30	5	0.100	1	0.643	0.999	0.196	0.308	1	0.901	1	0.170	0.249	0.720	0.419	0.806	0.196	0.308
30	30	0.010	0.984	0.048	0.323	0.055	0.190	0.999	0.129	0.658	0.053	0.157	0.010	0.013	0.051	0.055	0.190
30	30	0.050	0.997	0.135	0.767	0.135	0.332	1	0.278	0.974	0.122	0.259	0.069	0.072	0.134	0.135	0.332
30	30	0.100	0.999	0.222	0.792	0.207	0.425	1	0.392	0.979	0.192	0.340	0.130	0.133	0.186	0.207	0.425
60	5	0.010	1	0.781	1	0.056	0.140	1	0.993	1	0.043	0.106	0.862	0.338	0.999	0.056	0.140
60	5	0.050	1	0.955	1	0.123	0.248	1	1	1	0.104	0.196	0.985	0.676	1	0.123	0.248
60	5	0.100	1	0.988	1	0.198	0.330	1	1	1	0.161	0.258	0.995	0.828	1	0.198	0.330
60	30	0.010	1	0.081	0.826	0.043	0.203	1	0.316	0.991	0.043	0.161	0.018	0.019	0.171	0.043	0.203
60	30	0.050	1	0.212	0.999	0.109	0.333	1	0.562	1	0.114	0.263	0.095	0.093	0.409	0.109	0.333
60	30	0.100	1	0.324	1.000	0.176	0.427	1	0.689	1	0.178	0.350	0.188	0.165	0.434	0.176	0.427

Table 1.10: Power comparison from Example 3

n	p	α	(i)						(ii)						(iii)					
			HHG	PCOR	MGC	R_ρ	R_τ	HHG	PCOR	MGC	R_ρ	R_τ	HHG	PCOR	MGC	R_ρ	R_τ	HHG	PCOR	MGC
10	5	0.010	0.077	0.028	0.067	0.076	0.093	0.080	0.030	0.073	0.047	0.053	0.010	0.010	0.015	0.011	0.008	0.010	0.010	0.010
10	5	0.050	0.199	0.107	0.163	0.142	0.181	0.202	0.109	0.168	0.096	0.11	0.054	0.048	0.056	0.038	0.034	0.054	0.054	0.054
10	5	0.100	0.292	0.194	0.236	0.210	0.250	0.294	0.196	0.238	0.154	0.171	0.117	0.107	0.096	0.078	0.064	0.107	0.107	0.107
10	30	0.010	0.020	0.013	0.022	0.134	0.151	0.021	0.011	0.023	0.065	0.071	0.012	0.011	0.011	0.012	0.010	0.012	0.012	0.010
10	30	0.050	0.080	0.059	0.085	0.222	0.241	0.079	0.056	0.089	0.129	0.132	0.062	0.055	0.052	0.041	0.037	0.062	0.062	0.062
10	30	0.100	0.143	0.118	0.142	0.271	0.293	0.138	0.123	0.146	0.178	0.173	0.125	0.111	0.093	0.079	0.056	0.125	0.125	0.125
30	5	0.010	0.728	0.101	0.737	0.055	0.128	0.737	0.112	0.768	0.027	0.052	0.034	0.008	0.034	0.009	0.011	0.034	0.034	0.034
30	5	0.050	0.893	0.270	0.825	0.125	0.234	0.891	0.301	0.854	0.078	0.114	0.172	0.053	0.070	0.04	0.037	0.172	0.172	0.172
30	5	0.100	0.935	0.413	0.838	0.193	0.314	0.938	0.445	0.862	0.135	0.180	0.305	0.107	0.109	0.078	0.074	0.305	0.305	0.305
30	30	0.010	0.198	0.016	0.083	0.054	0.191	0.186	0.016	0.093	0.024	0.066	0.093	0.011	0.011	0.015	0.015	0.093	0.093	0.093
30	30	0.050	0.371	0.072	0.165	0.122	0.323	0.364	0.074	0.179	0.076	0.139	0.232	0.053	0.052	0.052	0.052	0.232	0.232	0.232
30	30	0.100	0.482	0.141	0.216	0.188	0.411	0.478	0.140	0.226	0.135	0.202	0.337	0.112	0.094	0.104	0.096	0.337	0.337	0.337
60	5	0.010	0.999	0.349	1	0.049	0.143	0.999	0.401	1	0.024	0.041	0.263	0.010	0.073	0.013	0.011	0.263	0.263	0.263
60	5	0.050	1	0.680	1	0.124	0.25	1.000	0.734	1	0.075	0.102	0.633	0.053	0.109	0.041	0.045	0.633	0.633	0.633
60	5	0.100	1	0.836	1	0.198	0.332	1	0.871	1	0.13	0.155	0.793	0.117	0.146	0.079	0.078	0.793	0.793	0.793
60	30	0.010	0.758	0.019	0.338	0.049	0.217	0.730	0.022	0.366	0.024	0.045	0.487	0.009	0.012	0.015	0.014	0.487	0.487	0.487
60	30	0.050	0.882	0.095	0.471	0.116	0.354	0.873	0.098	0.511	0.074	0.108	0.706	0.047	0.046	0.052	0.050	0.706	0.706	0.706
60	30	0.100	0.925	0.173	0.494	0.19	0.442	0.915	0.179	0.530	0.123	0.173	0.801	0.105	0.088	0.099	0.097	0.801	0.801	0.801

Table 1.11: Power Comparison on Earthquake data

n	p	α	HHG	PCOR	MGC	R_ρ	R_τ
10	5	0.010	0.067	0.085	0.032	0.791	0.874
10	5	0.050	0.169	0.179	0.102	0.873	0.958
10	5	0.100	0.236	0.256	0.156	0.914	0.983
10	30	0.010	0.008	0.009	0.007	1	1
10	30	0.050	0.050	0.043	0.050	1	1
10	30	0.100	0.097	0.094	0.090	1	1
30	5	0.010	0.333	0.413	0.136	0.922	1.000
30	5	0.050	0.548	0.570	0.313	0.983	1
30	5	0.100	0.638	0.657	0.350	0.994	1
30	30	0.010	0.009	0.007	0.015	1	1
30	30	0.050	0.048	0.038	0.050	1	1
30	30	0.100	0.100	0.088	0.088	1	1
60	5	0.010	0.725	0.812	0.427	0.997	1
60	5	0.050	0.906	0.903	0.750	1	1
60	5	0.100	0.945	0.936	0.779	1	1
60	30	0.010	0.005	0.003	0.017	1	1
60	30	0.050	0.037	0.027	0.056	1	1
60	30	0.100	0.077	0.069	0.085	1	1

1.7 Technical Details

1.7.1 Proof of Proposition 2

Proof. Denote $f^{(2)}(t) = -\frac{1}{4}(1+t)^{-\frac{3}{2}}$. The remainder term can be written as

$$R_X(X_s, X_t) = \int_0^1 \int_0^1 v f^{(2)}(uv L_X(X_s, X_t)) dudv \times (L_X(X_s, X_t))^2.$$

Set $\varphi(x) = \int_0^1 \int_0^1 v f^{(2)}(uvx) dudv$. Then $\varphi(x)$ is continuous at 0. Next, by the continuous mapping theorem, we have

$$\int_0^1 \int_0^1 v f^{(2)}(uv L_X(X_s, X_t)) dudv \xrightarrow{p} \int_0^1 \int_0^1 v f^{(2)}(0) dudv.$$

So, $R_X(X_s, X_t) \asymp_p (L_X(X_s, X_t))^2$. Similar arguments hold for $R_Y(Y_s, Y_t)$. \square

1.7.2 Proof of Remark 4

Proof. Denote $\mathbf{C} = (c_{ij}) = \mathbf{A}^T \mathbf{A} \in \mathbb{R}^{s_1 \times s_1}$. We obtain that

$$\begin{aligned} \text{var}[L(X, X')] &= \frac{\text{var}[(U - U')^T \mathbf{A}^T \mathbf{A} (U - U')]}{4\text{tr}^2(\mathbf{A}\mathbf{A}^T)} \\ &= \frac{\text{var}\left[\sum_{i=1}^{s_1} \sum_{j=1}^{s_1} c_{ij} (u_i - u'_i)(u_j - u'_j)\right]}{4\text{tr}^2(\mathbf{A}\mathbf{A}^T)} \\ &\leq \frac{\sum_{i=1}^{s_1} \sum_{j=1}^{s_1} c_{ij}^2 \text{var}[(u_i - u'_i)(u_j - u'_j)]}{\text{tr}^2(\mathbf{A}\mathbf{A}^T)} \\ &\leq \frac{C \sum_{i=1}^{s_1} \sum_{j=1}^{s_1} c_{ij}^2}{\text{tr}^2(\mathbf{A}\mathbf{A}^T)} \\ &= \frac{C \text{tr}(\mathbf{A}^T \mathbf{A} \mathbf{A}^T \mathbf{A})}{\text{tr}^2(\mathbf{A}\mathbf{A}^T)} = O(p^{-1}) \end{aligned}$$

for some constant $C > 0$, where we have used the fact that the fourth moment of u_i is uniformly bounded. It follows that $a_p = 1/\sqrt{p}$. □

1.7.3 Proof of Theorem 1

Proof. (i) Recall that $d\text{Cov}_n^2(\mathbf{X}, \mathbf{Y}) = (\tilde{\mathbf{A}} \cdot \tilde{\mathbf{B}})$. For any matrix \mathbf{A} , the matrix \mathbf{A}_{-D} is just \mathbf{A} with its diagonal elements setting to be 0. Using the expansion for a_{st} in Proposition 2, we have

$$\frac{1}{\tau_X} \tilde{\mathbf{A}} = (\widetilde{\mathbf{1}_{n \times n}})_{-D} + \frac{1}{2} \tilde{\mathbf{L}}_X + \tilde{\mathbf{R}}_X = \frac{1}{2} \tilde{\mathbf{L}}_X + \tilde{\mathbf{R}}_X,$$

where $\mathbf{L}_X = ((L_X(X_s, X_t))_{s,t=1}^n)_{-D}$ and $\mathbf{R}_X = ((R_X(X_s, X_t))_{s,t=1}^n)_{-D}$. Similarly, $\frac{1}{\tau_Y} \tilde{\mathbf{B}} = \frac{1}{2} \tilde{\mathbf{L}}_Y + \tilde{\mathbf{R}}_Y$. Then, we have

$$\begin{aligned} \frac{d\text{Cov}_n^2(\mathbf{X}, \mathbf{Y})}{\tau} &= ((\frac{1}{2} \tilde{\mathbf{L}}_X + \tilde{\mathbf{R}}_X) \cdot (\frac{1}{2} \tilde{\mathbf{L}}_Y + \tilde{\mathbf{R}}_Y)) \\ &= \frac{1}{4} (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{L}}_Y) + \frac{1}{2} (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{R}}_Y) + \frac{1}{2} (\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{L}}_Y) + (\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{R}}_Y). \end{aligned}$$

Let $R_n = \frac{1}{2} (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{R}}_Y) + \frac{1}{2} (\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{L}}_Y) + (\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{R}}_Y)$. We show that $\frac{1}{4} (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{L}}_Y)$ can be written as sum of sample component-wise cross-covariances up to a constant factor in the following Lemma.

Lemma 24.

$$\frac{1}{4} (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{L}}_Y) = \frac{1}{\tau^2} \sum_{i=1}^p \sum_{j=1}^q \text{cov}_n^2(x_i, y_j).$$

Proof. By Lemma A.1. of Park et al. (2015), since all diagonal entries of distance matrices \mathbf{A} and \mathbf{B} are equal to 0, we have $(\tilde{\mathbf{A}} \cdot \tilde{\mathbf{B}}) = (\mathbf{A} \cdot \mathbf{B})$. Then, it can be directly verified that for any $1 \leq s, t \leq n$,

$\sum_{u=1}^n \tilde{b}_{ut} = \sum_{v=1}^n \tilde{b}_{sv} = 0$ and it further implies that $\tilde{\tilde{\mathbf{B}}} = \tilde{\mathbf{B}}$. Direct calculation shows that

$$(\tilde{\mathbf{A}} \cdot \tilde{\mathbf{B}}) = (\mathbf{A} \cdot \tilde{\mathbf{B}}) = \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} a_{st} b_{st} + \frac{1}{\binom{n}{4}} \frac{1}{4!} \sum_{(s,t,u,v) \in \mathbf{i}_4^n} a_{st} b_{uv} - \frac{2}{\binom{n}{3}} \frac{1}{3!} \sum_{(s,t,u) \in \mathbf{i}_3^n} a_{st} b_{su}, \quad (1.14)$$

where \mathbf{i}_m^n denotes the set of all m -tuples drawn without replacement from $\{1, 2, \dots, n\}$. Equation (1.14) can be used as equivalent definition of the sample distance covariance. Notice that

$$\tilde{\mathbf{L}}_X = \tilde{\mathbf{D}}_X - (\mathbf{1}_{n \times n})_{-D} = \tilde{\mathbf{D}}_X,$$

where $\mathbf{D}_X = \frac{1}{\tau_X^2} ((|X_s - X_t|^2)_{s,t=1}^n)_{-D}$. Similarly, $\tilde{\mathbf{L}}_Y = \tilde{\mathbf{D}}_Y$. Next, it can be verified directly that for any vector $\mathbf{a} \in \mathbf{R}^n$,

$$(\mathbf{a} \mathbf{1}_n^T)_{-D} + (\mathbf{1}_n \mathbf{a})_{-D} = 0.$$

Using this fact, we then can further decompose $\tilde{\mathbf{D}}_X$ as follows,

$$\tilde{\mathbf{D}}_X = \tilde{\mathbf{D}}_{X,1} + \tilde{\mathbf{D}}_{X,2} + \tilde{\mathbf{D}}_{X,3} = \tilde{\mathbf{D}}_{X,2},$$

where $\mathbf{D}_{X,1} = \frac{1}{\tau_X^2} ((X_s^T X_s)_{s,t=1}^n)_{-D}$, $\mathbf{D}_{X,2} = -2 \frac{1}{\tau_X^2} ((X_s^T X_t)_{s,t=1}^n)_{-D}$ and $\mathbf{D}_{X,3} = \frac{1}{\tau_X^2} ((X_t^T X_t)_{s,t=1}^n)_{-D}$. Similarly, $\tilde{\mathbf{D}}_Y = \tilde{\mathbf{D}}_{Y,2}$. Next, using Equation (1.14), we have

$$\begin{aligned} & \tau^2 \times (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{L}}_Y) \\ &= \tau^2 \times (\tilde{\mathbf{D}}_{X,2} \cdot \tilde{\mathbf{D}}_{Y,2}) \\ &= 4 \left\{ \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} X_s^T X_t Y_s^T Y_t + \right. \\ & \quad \left. \frac{1}{\binom{n}{4}} \frac{1}{4!} \sum_{(s,t,u,v) \in \mathbf{i}_4^n} X_s^T X_t Y_u^T Y_v - \frac{2}{\binom{n}{3}} \frac{1}{3!} \sum_{(s,t,u) \in \mathbf{i}_3^n} X_s^T X_t Y_s^T Y_u \right\} \\ &= 4 \sum_{i=1}^p \sum_{j=1}^q \left\{ \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} x_{si} x_{ti} y_{sj} y_{tj} + \right. \\ & \quad \left. \frac{1}{\binom{n}{4}} \frac{1}{4!} \sum_{(s,t,u,v) \in \mathbf{i}_4^n} x_{si} x_{ti} y_{uj} y_{vj} - \frac{2}{\binom{n}{3}} \frac{1}{3!} \sum_{(s,t,u) \in \mathbf{i}_3^n} x_{si} x_{ti} y_{sj} y_{uj} \right\} \\ &= 4 \sum_{i=1}^p \sum_{j=1}^q \left\{ \frac{1}{\binom{n}{4}} \sum_{k < l < s < t} \frac{1}{4!} \sum_{*}^{(k,l,s,t)} \frac{(x_{ki} - x_{li})(y_{kj} - y_{lj})(x_{si} - x_{ti})(y_{sj} - y_{tj})}{4} \right\} \\ &= 4 \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j). \end{aligned}$$

□

Therefore, by Lemma 24, we have the following decomposition,

$$dCov_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) + \mathcal{R}_n,$$

where $\mathcal{R}_n = \tau R_n$.

(ii) Note $L_X(X_s, X_t) = O_p(a_p) = o_p(1)$ and $L_Y(Y_s, Y_t) = O_p(b_q) = o_p(1)$ for $s \neq t \in \{1, \dots, n\}$. We can then apply Proposition 2, obtain that $R_X(X_s, X_t) = O_p(L_X(X_s, X_t)^2)$ and $R_Y(Y_s, Y_t) = O_p(L_Y(Y_s, Y_t)^2)$. For the leading term $\tau(\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{L}}_Y)$, it can be easily seen from Equation (1.14) that $(\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{L}}_Y) = O_p(a_p b_q)$. Similarly, for the remainder terms, $(\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{R}}_Y) = O_p(a_p b_q^2)$, $(\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{L}}_Y) = O_p(a_p^2 b_q)$ and $(\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{R}}_Y) = O_p(a_p^2 b_q^2)$. Thus, we have $R_n = O_p(a_p^2 b_q + a_p b_q^2)$ and $\mathcal{R}_n = \tau R_n = O_p(\tau a_p^2 b_q + \tau a_p b_q^2) = o_p(1)$. Therefore the remainder terms are negligible comparing to the leading term. \square

1.7.4 Proof of Theorem 2

Proof. (i) We first show that $\gamma_{\mathbf{X}}$ is asymptotically equal to τ_X (similar result applies to $\gamma_{\mathbf{Y}}$ and τ_Y). Recall that for all $s \neq t$,

$$L_X(X_s, X_t) = \frac{|X_s - X_t|^2 - \tau_X^2}{\tau_X^2}.$$

Since $L_X(X_s, X_t) = O_p(a_p) = o_p(1)$, we have $\frac{|X_s - X_t|^2}{\tau_X^2} \xrightarrow{p} 1$. Then

$$\frac{\text{median}\{|X_s - X_t|^2\}}{\tau_X^2} \xrightarrow{p} 1$$

and thus

$$\frac{\tau_X}{\gamma_{\mathbf{X}}} = \sqrt{\frac{\tau_X^2}{\text{median}\{|X_i - X_j|^2\}}} \xrightarrow{p} 1.$$

Similar arguments can also be used to show that $\frac{\tau_Y}{\gamma_{\mathbf{Y}}} \xrightarrow{p} 1$. Next, under Proposition 2, we can deduce that

$$\begin{aligned} & f\left(\frac{|X_s - X_t|}{\gamma_{\mathbf{X}}}\right) \\ &= f\left(\frac{|X_s - X_t|}{\tau_X} \frac{\tau_X}{\gamma_{\mathbf{X}}}\right) \\ &= f\left(\left\{1 + \frac{L_X(X_s, X_t)}{2} + R_X(X_s, X_t)\right\} \frac{\tau_X}{\gamma_{\mathbf{X}}}\right) \\ &= f\left(\frac{\tau_X}{\gamma_{\mathbf{X}}}\right) + f^{(1)}\left(\frac{\tau_X}{\gamma_{\mathbf{X}}}\right) \left\{\frac{L_X(X_s, X_t)}{2} + R_X(X_s, X_t)\right\} \frac{\tau_X}{\gamma_{\mathbf{X}}} + R_f(X_s, X_t), \end{aligned}$$

where $R_f(X_s, X_t)$ is the remainder term. Similarly,

$$g\left(\frac{|Y_s - Y_t|}{\gamma_{\mathbf{Y}}}\right) = g\left(\frac{\tau_Y}{\gamma_{\mathbf{Y}}}\right) + g^{(1)}\left(\frac{\tau_Y}{\gamma_{\mathbf{Y}}}\right) \left\{\frac{L_Y(Y_s, Y_t)}{2} + R_Y(Y_s, Y_t)\right\} \frac{\tau_Y}{\gamma_{\mathbf{Y}}} + R_g(Y_s, Y_t).$$

Similar to the proof of Theorem 1,

$$\begin{aligned} hCov_n^2(\mathbf{X}, \mathbf{Y}) &= (\tilde{\mathbf{R}} \cdot \tilde{\mathbf{H}}) \\ &= \frac{1}{4} f^{(1)} \left(\frac{\tau_X}{\gamma_X} \right) g^{(1)} \left(\frac{\tau_Y}{\gamma_Y} \right) \frac{\tau_X}{\gamma_X} \frac{\tau_Y}{\gamma_Y} (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{L}}_Y) + \frac{1}{2} f^{(1)} \left(\frac{\tau_X}{\gamma_X} \right) \frac{\tau_X}{\gamma_X} (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{R}}_Y) \\ &\quad + \frac{1}{2} g^{(1)} \left(\frac{\tau_Y}{\gamma_Y} \right) \frac{\tau_Y}{\gamma_Y} (\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{L}}_Y) + (\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{R}}_Y), \end{aligned}$$

where $\mathbf{L}_X = ((L_X(X_s, X_t))_{s,t=1}^n)_{-D}$, $\mathbf{L}_Y = ((L_Y(Y_s, Y_t))_{s,t=1}^n)_{-D}$ and

$$\begin{aligned} \mathbf{R}_X &= \left(\left(f^{(1)} \left(\frac{\tau_X}{\gamma_X} \right) \frac{\tau_X}{\gamma_X} R_X(X_s, X_t) + R_f(X_s, X_t) \right)_{s,t=1}^n \right)_{-D}, \\ \mathbf{R}_Y &= \left(\left(g^{(1)} \left(\frac{\tau_Y}{\gamma_Y} \right) \frac{\tau_Y}{\gamma_Y} R_Y(Y_s, Y_t) + R_g(Y_s, Y_t) \right)_{s,t=1}^n \right)_{-D}. \end{aligned}$$

Denote $R_n = \frac{1}{2} f^{(1)} \left(\frac{\tau_X}{\gamma_X} \right) \frac{\tau_X}{\gamma_X} (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{R}}_Y) + \frac{1}{2} g^{(1)} \left(\frac{\tau_Y}{\gamma_Y} \right) \frac{\tau_Y}{\gamma_Y} (\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{L}}_Y) + (\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{R}}_Y)$ and $\mathcal{R}_n = \tau R_n$. By Lemma 24, we have

$$\tau \times hCov_n^2(\mathbf{X}, \mathbf{Y}) = f^{(1)} \left(\frac{\tau_X}{\gamma_X} \right) g^{(1)} \left(\frac{\tau_Y}{\gamma_Y} \right) \frac{\tau_X}{\gamma_X} \frac{\tau_Y}{\gamma_Y} \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) + \mathcal{R}_n.$$

(ii) We present the following lemma which would be useful in subsequent arguments.

Lemma 25. *Suppose $f^{(2)}$ and $g^{(2)}$ are continuous on some open interval containing 1. Then under the assumptions of Theorem 2,*

$$R_f(X_s, X_t) = O_p(L_X(X_s, X_t)^2), \quad R_g(Y_s, Y_t) = O_p(L_Y(Y_s, Y_t)^2).$$

Proof. The remainder term can be written as

$$\begin{aligned} R_f(X_s, X_t) &= \int_0^1 \int_0^1 v f^{(2)} \left(\frac{\tau_X}{\gamma_X} + uv \left\{ \frac{L_X(X_s, X_t)}{2} + R_X(X_s, X_t) \right\} \frac{\tau_X}{\gamma_X} \right) dudv \\ &\quad \times \left(\frac{\tau_X}{\gamma_X} \right)^2 \left(\frac{L_X(X_s, X_t)}{2} + R_X(X_s, X_t) \right)^2. \end{aligned} \quad (1.15)$$

Set $\varphi(x, y) = \int_0^1 \int_0^1 v f^{(2)}(x + uv y) dudv$. Then $\varphi(x, y)$ is continuous at $(1, 0)$. By the continuous mapping theorem, we have

$$\int_0^1 \int_0^1 v f^{(2)} \left(\frac{\tau_X}{\gamma_X} + uv \left\{ \frac{L_X(X_s, X_t)}{2} + R_X(X_s, X_t) \right\} \frac{\tau_X}{\gamma_X} \right) dudv \xrightarrow{p} \int_0^1 \int_0^1 v f^{(2)}(1) dudv.$$

So $R_f(X_s, X_t) = O_p(1) \left(\frac{L_X(X_s, X_t)}{2} + R_X(X_s, X_t) \right)^2 = O_p(L_X(X_s, X_t)^2)$. Similar argument holds for $R_g(Y_s, Y_t)$. \square

Both the Gaussian and Laplacian kernel have continuous second order derivatives. From Lemma 25, we

know

$$\begin{aligned} f^{(1)} \left(\frac{\tau_X}{\gamma_{\mathbf{X}}} \right) \frac{\tau_X}{\gamma_{\mathbf{X}}} R_X(X_s, X_t) + R_f(X_s, X_t) &= O_p(L_X(X_s, X_t)^2), \\ g^{(1)} \left(\frac{\tau_Y}{\gamma_{\mathbf{Y}}} \right) \frac{\tau_Y}{\gamma_{\mathbf{Y}}} R_Y(Y_s, Y_t) + R_g(Y_s, Y_t) &= O_p(L_Y(Y_s, Y_t)^2). \end{aligned}$$

Thus, similar arguments in Theorem 1 can be used to show that $\mathcal{R}_n = O_p(\tau a_p^2 b_q + \tau a_p b_q^2) = o_p(1)$. \square

1.7.5 Proof of Proposition 6

Proof. Clearly, $E[k_{st}(i)l_{uv}(j)] = 0$ when $\{s, t\} \cap \{u, v\} = \emptyset$. For any $1 \leq i, i' \leq p, 1 \leq j, j' \leq q$,

$$\begin{aligned} &E[k_{st}(i)l_{su}(j)] \\ &= E[E[k_{st}(i)l_{su}(j)|x_{si}, y_{sj}]] \\ &= E[E[k_{st}(i)|x_{si}, y_{sj}]E[l_{su}(j)|x_{si}, y_{sj}]]. \end{aligned}$$

Notice that

$$\begin{aligned} &E[k_{st}(i)|x_{si}, y_{sj}] \\ &= E\{k(x_{si}, x_{ti}) - E[k(x_{si}, x_{ti})|x_{si}] - E[k(x_{si}, x_{ti})|x_{ti}] + E[k(x_{si}, x_{ti})]|x_{si}, y_{sj}\} \\ &= E[k(x_{si}, x_{ti})|x_{si}, y_{sj}] - E[k(x_{si}, x_{ti})|x_{si}] - E[k(x_{si}, x_{ti})] + E[k(x_{si}, x_{ti})] \\ &= 0. \end{aligned}$$

Thus $E[k_{st}(i)l_{su}(j)] = 0$. Similarly, $E[k_{st}(i)k_{su}(i')] = E[l_{st}(j)l_{su}(j')] = 0$. \square

1.7.6 Proof of Theorem 3

Proof. Let $\tilde{\mathbf{K}} = ((\tilde{k}_{st})_{s,t=1}^n)_{-D}$ and $\tilde{\mathbf{L}} = ((\tilde{l}_{st})_{s,t=1}^n)_{-D}$. Notice that

$$\begin{aligned} uCov_n^2(\mathbf{X}, \mathbf{Y}) &= (pq)^{-1/2} \sum_{i=1}^p \sum_{j=1}^q \frac{1}{n(n-3)} \sum_{s \neq t} \tilde{k}_{st}(i) \tilde{l}_{st}(j) \\ &= \frac{1}{n(n-3)} \sum_{s \neq t} \left(p^{-1/2} \sum_{i=1}^p \tilde{k}_{st}(i) \right) \left(q^{-1/2} \sum_{j=1}^q \tilde{l}_{st}(j) \right). \end{aligned}$$

Under Assumption 7, we have

$$\begin{aligned}
& p^{-1/2} \sum_{i=1}^p \tilde{k}_{st}(i) \\
&= p^{-1/2} \sum_{i=1}^p k_{st}(i) - \frac{1}{n-2} \sum_{u \neq t} p^{-1/2} \sum_{i=1}^p k_{ut}(i) \\
&\quad - \frac{1}{n-2} \sum_{v \neq s} p^{-1/2} \sum_{i=1}^p k_{sv}(i) + \frac{1}{(n-1)(n-2)} \sum_{u \neq v} p^{-1/2} \sum_{i=1}^p k_{uv}(i) \\
&\xrightarrow{d} c_{st} - \frac{1}{n-2} \sum_{u \neq t} c_{ut} - \frac{1}{n-2} \sum_{v \neq s} c_{vs} + \frac{1}{(n-1)(n-2)} \sum_{u \neq v} c_{uv}.
\end{aligned}$$

Then we get

$$\begin{aligned}
& n(n-3) \times uCov_n^2(\mathbf{X}, \mathbf{Y}) \xrightarrow{d} \\
& \sum_{s \neq t} \left(c_{st} - \frac{1}{n-2} \sum_{u \neq t} c_{ut} - \frac{1}{n-2} \sum_{v \neq s} c_{sv} + \frac{1}{(n-1)(n-2)} \sum_{u \neq v} c_{uv} \right) \\
& \quad \times \left(d_{st} - \frac{1}{n-2} \sum_{u \neq t} d_{ut} - \frac{1}{n-2} \sum_{v \neq s} d_{sv} + \frac{1}{(n-1)(n-2)} \sum_{u \neq v} d_{uv} \right).
\end{aligned}$$

Set

$$\begin{aligned}
\mathbf{c} &= (c_{12}, c_{13}, \dots, c_{1n}, c_{23}, \dots, c_{2n}, c_{34}, \dots, c_{n(n-1)})^T, \\
\mathbf{d} &= (d_{12}, d_{13}, \dots, d_{1n}, d_{23}, \dots, d_{2n}, d_{34}, \dots, d_{n(n-1)})^T.
\end{aligned}$$

Under Assumption 7 and by Proposition 6, we know that

$$\begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \sim N \left(\mathbf{0}, \begin{pmatrix} \sigma_x^2 \mathbf{I}_{n(n-1)/2} & \sigma_{xy}^2 \mathbf{I}_{n(n-1)/2} \\ \sigma_{xy}^2 \mathbf{I}_{n(n-1)/2} & \sigma_y^2 \mathbf{I}_{n(n-1)/2} \end{pmatrix} \right).$$

Define $\mathbf{C} = (c_{st})_{s,t=1}^n$ such that $c_{st} = c_{ts}$ and $\tilde{\mathbf{C}} = (\tilde{c}_{st})_{s,t=1}^n$. Here we assume that $c_{ss} = 0$. Following Park et al. (2015), for any matrix \mathbf{A} , define $\dot{\mathbf{A}}$ as

$$\dot{\mathbf{A}} = \mathbf{A} - \frac{1}{n-2} \mathbf{A} \mathbf{J} - \frac{1}{n-2} \mathbf{J} \mathbf{A} + \frac{1}{(n-1)(n-2)} \mathbf{J} \mathbf{A} \mathbf{J},$$

where $\mathbf{J} = \mathbf{1}_{n \times n}$ is the $n \times n$ matrix of ones. Let $\text{vec}(\mathbf{A})$ is the usual vectorization of matrix \mathbf{A} , then it is straightforward to see that

$$\text{vec}(\dot{\mathbf{A}}) = \mathbf{S} \text{vec}(\mathbf{A}),$$

where

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

Also notice that $\tilde{\mathbf{A}}$ is just \mathbf{A} with diagonal replaced by 0. Putting things together, we have

$$\text{vec}(\tilde{\mathbf{C}}) = \mathbf{F}\text{vec}(\mathbf{C}) = \mathbf{F}\mathbf{S}\mathbf{F}\text{vec}(\mathbf{C}),$$

where \mathbf{F} is the matrix of the linear operator that sets the diagonal of a matrix to be 0, i.e., $\text{vec}(\mathbf{B}_{-D}) = \mathbf{F}\text{vec}(\mathbf{B})$, \mathbf{B}_{-D} is \mathbf{B} with its diagonal set to be 0.

Next, to simplify the following proof, we will use a different vectorization operator, which will align the upper triangular elements first, then the lower triangular elements and lastly the diagonal elements, i.e., define

$$\begin{aligned}\widetilde{\text{vec}}(\mathbf{C}) &= (\mathbf{c}_u^T, \mathbf{c}_l^T, \mathbf{c}_d^T)^T, \\ \mathbf{c}_u^T &= (c_{12}, c_{13}, \dots, c_{1n}, c_{23}, \dots, c_{2n}, c_{34}, \dots, c_{(n-1)n}), \\ \mathbf{c}_l^T &= (c_{21}, c_{31}, \dots, c_{n1}, c_{32}, \dots, c_{n2}, c_{43}, \dots, c_{n(n-1)}), \\ \mathbf{c}_d^T &= (c_{11}, c_{22}, \dots, c_{nn}).\end{aligned}$$

Notice that there is a permutation matrix \mathbf{P}_1 such that $\widetilde{\text{vec}}(\mathbf{C}) = \mathbf{P}_1\text{vec}(\mathbf{C})$. Then

$$\widetilde{\text{vec}}(\tilde{\mathbf{C}}) = \mathbf{P}_1\mathbf{F}\mathbf{S}\mathbf{F}\mathbf{P}_1^T\widetilde{\text{vec}}(\mathbf{C}).$$

Observe that for any symmetric matrix \mathbf{C} with 0 diagonal, both the column sum and row sum of $\tilde{\mathbf{C}}$ are 0. We can verify that $\tilde{\mathbf{C}} = \tilde{\mathbf{C}}^T$. Set $\mathbf{U} = \mathbf{P}_1\mathbf{F}\mathbf{S}\mathbf{F}\mathbf{P}_1^T$. It follows that $\mathbf{U}^2\widetilde{\text{vec}}(\mathbf{C}) = \mathbf{U}\widetilde{\text{vec}}(\mathbf{C})$ and thus

$$(\mathbf{U}^2 - \mathbf{U})\widetilde{\text{vec}}(\mathbf{C}) = 0. \quad (1.16)$$

We partition \mathbf{U} into three blocks corresponding to the upper triangular, lower triangular and diagonal elements respective, i.e., we write

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{U}_2 & \mathbf{0} \\ \mathbf{U}_2 & \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where we have used the symmetry for \mathbf{U} . Equation (1.16) is then equivalent to $(\mathbf{U}_1^2 + \mathbf{U}_2^2 - \mathbf{U}_1)\mathbf{c}_u + (\mathbf{U}_1\mathbf{U}_2 + \mathbf{U}_2\mathbf{U}_1 - \mathbf{U}_2)\mathbf{c}_l = 0$. Using the facts that \mathbf{c}_u is arbitrary and $\mathbf{c}_u = \mathbf{c}_l$, we can conclude that

$$\mathbf{U}_1^2 + \mathbf{U}_2^2 + \mathbf{U}_1\mathbf{U}_2 + \mathbf{U}_2\mathbf{U}_1 = \mathbf{U}_1 + \mathbf{U}_2.$$

Next, let \mathbf{C}^u (\mathbf{C}^l) be the matrix by setting the lower (upper) triangular and diagonal elements in \mathbf{C} to be zero. Denote

$$\mathbf{P}_2 = \begin{pmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

Then, we see that $\widetilde{\text{vec}}(\mathbf{C}^l) = \mathbf{P}_2 \widetilde{\text{vec}}(\mathbf{C}^u)$ and

$$\mathbf{U} \widetilde{\text{vec}}(\mathbf{C}) = \mathbf{U} \widetilde{\text{vec}}(\mathbf{C}^u) + \mathbf{U} \mathbf{P}_2 \widetilde{\text{vec}}(\mathbf{C}^u) = \mathbf{U}(\mathbf{I} + \mathbf{P}_2) \widetilde{\text{vec}}(\mathbf{C}^u) = \mathbf{U}(\mathbf{I} + \mathbf{P}_2) \mathbf{D} \mathbf{c}.$$

We note that

$$\mathbf{W} := \mathbf{D}^T (\mathbf{I} + \mathbf{P}_2) \mathbf{U} \mathbf{U} (\mathbf{I} + \mathbf{P}_2) \mathbf{D} = \mathbf{D}^T (\mathbf{U} + \mathbf{U} \mathbf{P}_2 + \mathbf{P}_2 \mathbf{U} + \mathbf{P}_2 \mathbf{U} \mathbf{P}_2) \mathbf{D}.$$

Then we have $\mathbf{W} = 2(\mathbf{U}_1 + \mathbf{U}_2)$. Also notice that

$$\mathbf{W}^2 = 4(\mathbf{U}_1 + \mathbf{U}_2)^2 = 4(\mathbf{U}_1^2 + \mathbf{U}_2^2 + \mathbf{U}_1 \mathbf{U}_2 + \mathbf{U}_2 \mathbf{U}_1) = 4(\mathbf{U}_1 + \mathbf{U}_2) = 2\mathbf{W},$$

which indicates that \mathbf{W} has eigenvalues which are either equal to two or zero. It remains to show that the rank of \mathbf{W} is $n(n-3)/2$ or equivalently, the trace of $\mathbf{W}/2 = \mathbf{U}_1 + \mathbf{U}_2$ is $n(n-3)/2$. Note that

$$\text{Tr}(\mathbf{W}/2) = \text{Tr}(\mathbf{U}_1 + \mathbf{U}_2) = \sum_{i=1}^{n(n-1)/2} \frac{\mathbf{r}_i^T \mathbf{U} \mathbf{r}_i}{2} = \frac{n(n-1)}{4} \widetilde{\text{vec}}(\tilde{\mathbf{E}}_1)^T \widetilde{\text{vec}}(\tilde{\mathbf{E}}_1),$$

where $\mathbf{r}_i = (\mathbf{e}_i^T, \mathbf{e}_i^T, \mathbf{0}^T)^T$ and \mathbf{e}_i is a $n(n-1)/2$ -dimensional vector with 1 on the i th position and zero otherwise; $\tilde{\mathbf{E}}_i$ denotes the \mathcal{U} -centering version of the matrix \mathbf{E}_i such that $\widetilde{\text{vec}}(\mathbf{E}_i) = \mathbf{r}_i$. Direct calculation shows that

$$\begin{aligned} \text{vec}(\tilde{\mathbf{E}}_1)^T \text{vec}(\tilde{\mathbf{E}}_1) &= \frac{2(n-3)^2}{(n-1)^2} + 4(n-2) \frac{(n-3)^2}{(n-1)^2(n-2)^2} \\ &\quad + (n-2)(n-3) \frac{4}{(n-1)^2(n-2)^2} = \frac{2(n-3)}{n-1}, \end{aligned}$$

which implies that $4^{-1}n(n-1)\widetilde{\text{vec}}(\tilde{\mathbf{E}}_1)^T \widetilde{\text{vec}}(\tilde{\mathbf{E}}_1) = n(n-3)/2$. Using the above results and setting $\mathbf{M} = \mathbf{W}/2$, we have

$$\text{vec}(\tilde{\mathbf{C}})^T \text{vec}(\tilde{\mathbf{C}}) = \widetilde{\text{vec}}(\tilde{\mathbf{C}})^T \widetilde{\text{vec}}(\tilde{\mathbf{C}}) = \widetilde{\text{vec}}(\mathbf{C})^T \mathbf{U}^2 \widetilde{\text{vec}}(\mathbf{C}) = 2\mathbf{c}^T \mathbf{M} \mathbf{c} \sim 2\sigma_x^2 \chi_{n(n-3)/2}^2.$$

Thus,

$$uCov_n^2(\mathbf{X}, \mathbf{X}) \xrightarrow{d} \frac{2}{n(n-3)} \mathbf{c}^T \mathbf{M} \mathbf{c} \stackrel{d}{=} \frac{2}{n(n-3)} \sigma_x^2 \chi_{n(n-3)/2}^2.$$

Similarly,

$$\begin{aligned} uCov_n^2(\mathbf{X}, \mathbf{Y}) &\xrightarrow{d} \frac{2}{n(n-3)} \mathbf{c}^T \mathbf{M} \mathbf{d}, \\ uCov_n^2(\mathbf{Y}, \mathbf{Y}) &\xrightarrow{d} \frac{2}{n(n-3)} \mathbf{d}^T \mathbf{M} \mathbf{d} \stackrel{d}{=} \frac{2}{n(n-3)} \sigma_y^2 \chi_{n(n-3)/2}^2. \end{aligned}$$

□

1.7.7 Proof of Proposition 10

Proof. Since

$$\begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \stackrel{d}{=} N \left(\mathbf{0}, \begin{pmatrix} \sigma_x^2 \mathbf{I}_{n(n-1)/2} & \sigma_{xy}^2 \mathbf{I}_{n(n-1)/2} \\ \sigma_{xy}^2 \mathbf{I}_{n(n-1)/2} & \sigma_y^2 \mathbf{I}_{n(n-1)/2} \end{pmatrix} \right),$$

we have

$$\mathbf{c}|\mathbf{d} \stackrel{d}{=} N(\mu\mathbf{d}, \sigma^2 \mathbf{I}_{n(n-1)/2}),$$

where $\mu = \sigma_{xy}^2/\sigma_y^2$, $\sigma^2 = (\sigma_x^2\sigma_y^2 - \sigma_{xy}^4)/\sigma_y^2$. Set

$$\mathbf{z} = \frac{\mathbf{M}\mathbf{d}}{\sqrt{(\mathbf{d}^T \mathbf{M} \mathbf{d})}}.$$

It can be easily seen that conditional on \mathbf{d} ,

$$\mathbf{c}^T \mathbf{z} / \sigma \sim N(\mu \mathbf{z}^T \mathbf{d} / \sigma, 1),$$

which implies that $(\mathbf{c}^T \mathbf{z})^2 / \sigma^2 | \mathbf{d} \sim \chi_1^2(W^2)$, where $\chi_1^2(W^2)$ is the non-central chi-squared distribution and $W^2 = \frac{\mu^2}{\sigma^2} \mathbf{d}^T \mathbf{M} \mathbf{d}$ is the non-centrality parameter. Note that conditioned on \mathbf{d} ,

$$\mathbf{M}(\mathbf{I} - \mathbf{z}\mathbf{z}^T)\mathbf{c} / \sigma \sim N(\mathbf{0}, \mathbf{M}(\mathbf{I} - \mathbf{z}\mathbf{z}^T)\mathbf{M}),$$

where we have used the fact that $\mathbf{M}(\mathbf{I} - \mathbf{z}\mathbf{z}^T)\mathbf{d} = \mathbf{0}$. As $\mathbf{M}(\mathbf{I} - \mathbf{z}\mathbf{z}^T)\mathbf{M} = \mathbf{M} - \frac{\mathbf{M}\mathbf{d}\mathbf{d}^T\mathbf{M}}{\mathbf{d}^T\mathbf{M}\mathbf{d}}$ is a projection matrix with rank $v - 1$, it is easy to see that conditioned on \mathbf{d} ,

$$\mathbf{c}^T (\mathbf{I} - \mathbf{z}\mathbf{z}^T) \mathbf{M} (\mathbf{I} - \mathbf{z}\mathbf{z}^T) \mathbf{c} / \sigma^2 \sim \chi_{v-1}^2.$$

Next, conditioned on \mathbf{d} , as $\mathbf{z}^T \mathbf{c}$ and $(\mathbf{I} - \mathbf{z}\mathbf{z}^T)\mathbf{c}$ are independent, we have $(\mathbf{c}^T \mathbf{z})^2/\sigma^2$ and $\mathbf{c}^T(\mathbf{I} - \mathbf{z}\mathbf{z}^T)\mathbf{M}(\mathbf{I} - \mathbf{z}\mathbf{z}^T)\mathbf{c}$ are independent. Then,

$$\begin{aligned}
P_{H_A}(T_u < t) &\rightarrow P \left(\sqrt{v-1} \frac{\frac{\mathbf{c}^T \mathbf{z}}{\sqrt{(\mathbf{c}^T \mathbf{M} \mathbf{c})}}}{\sqrt{1 - \left(\frac{\mathbf{c}^T \mathbf{z}}{\sqrt{(\mathbf{c}^T \mathbf{M} \mathbf{c})}} \right)^2}} < t \right) \\
&= E \left[P \left(\sqrt{v-1} \frac{\frac{\mathbf{c}^T \mathbf{z}}{\sqrt{(\mathbf{c}^T \mathbf{M} \mathbf{c})}}}{\sqrt{1 - \left(\frac{\mathbf{c}^T \mathbf{z}}{\sqrt{(\mathbf{c}^T \mathbf{M} \mathbf{c})}} \right)^2}} < t \middle| \mathbf{d} \right) \right] \\
&= E \left[P \left(\sqrt{v-1} \frac{\mathbf{c}^T \mathbf{z}}{\sqrt{\mathbf{c}^T \mathbf{M} \mathbf{c} - (\mathbf{c}^T \mathbf{z})^2}} < t \middle| \mathbf{d} \right) \right] \\
&= E \left[P \left(\frac{\mathbf{c}^T \mathbf{z}}{\sqrt{\frac{1}{v-1} \mathbf{c}^T (\mathbf{I} - \mathbf{z}\mathbf{z}^T) \mathbf{M} (\mathbf{I} - \mathbf{z}\mathbf{z}^T) \mathbf{c}}} < t \middle| \mathbf{d} \right) \right] \\
&= E [P(t_{v-1, W} < t)]
\end{aligned}$$

where $t_{v-1, W}$ is a noncentral t -distribution with $v-1$ degrees of freedom and noncentrality parameter $W = \frac{\mu}{\sigma} \sqrt{\mathbf{d}^T \mathbf{M} \mathbf{d}} \stackrel{d}{=} c\chi_v$ for $c = \frac{\sigma_{xy}^2}{\sqrt{\sigma_x^2 \sigma_y^2 - \sigma_{xy}^4}}$. By setting $c = 0$, we get $P_{H_0}(T_u < t) \rightarrow P(t_{v-1} < t)$. \square

1.7.8 Proof of Proposition 12

Proof. Notice that

$$\phi = \frac{\phi_0}{\sqrt{v}} \Rightarrow c = \frac{\phi_0}{\sqrt{v - \phi_0^2}} = \frac{\phi_0}{\sqrt{v}} \left(1 + O\left(\frac{1}{v}\right) \right).$$

Next, by the definition of non-central t -distribution,

$$\begin{aligned}
P(t_{v-1, u} < t) &= P \left(\frac{Z + u}{\sqrt{\chi_{v-1}^2/(v-1)}} < t \right) \\
&= P \left(Z < t\sqrt{\chi_{v-1}^2/(v-1)} - u \right) \\
&= E \left[P \left(Z < t\sqrt{\chi_{v-1}^2/(v-1)} - u \middle| \chi_{v-1}^2 \right) \right] \\
&= E \left[\Phi \left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}} - u \right) \right],
\end{aligned}$$

where Φ is the cdf of standard normal. For notational convenience, set

$$g(u) = E \left[\Phi \left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}} - u \right) \right].$$

Notice that $P(t_{v-1, W} < t) = g(W)$. By the following asymptotic series [see Laforgia and Natalini (2012);

Tricomi and Erdélyi (1951)],

$$\begin{aligned}\frac{\Gamma(J+1/2)}{\Gamma(J)} &= \sqrt{J} \left(1 - \frac{1}{8J} + \frac{1}{128J^2} + \frac{5}{1024J^3} - \frac{21}{32768J^4} + \cdots \right) \\ &= \sqrt{J} \left(1 + O\left(\frac{1}{J}\right) \right),\end{aligned}$$

we can get,

$$\begin{aligned}E[(W - \phi_0)] &= \frac{\phi_0}{\sqrt{v}} \left(1 + O\left(\frac{1}{v}\right) \right) \sqrt{2} \frac{\Gamma((v+1)/2)}{\Gamma(v/2)} - \phi_0 \\ &= \phi_0 \left(1 + O\left(\frac{1}{v}\right) \right) - \phi_0 \\ &= O\left(\frac{1}{v}\right),\end{aligned}$$

as well as

$$\begin{aligned}E[(W - \phi_0)^2] &= \phi_0^2 E \left[\left(\frac{\chi_v}{\sqrt{v}} \left(1 + O\left(\frac{1}{v}\right) \right) - 1 \right)^2 \right] \\ &= \phi_0^2 E \left[\frac{\chi_v^2}{v} \left(1 + O\left(\frac{1}{v}\right) \right) - 2 \frac{\chi_v}{\sqrt{v}} \left(1 + O\left(\frac{1}{v}\right) \right) + 1 \right] \\ &= \phi_0^2 \left\{ \left(1 + O\left(\frac{1}{v}\right) \right) - 2 \left(1 + O\left(\frac{1}{v}\right) \right) + 1 \right\} \\ &= O\left(\frac{1}{v}\right),\end{aligned}$$

and

$$\begin{aligned}E[W(W - \phi_0)^2] &= \phi_0^3 E \left[\frac{\chi_v}{\sqrt{v}} \left(1 + O\left(\frac{1}{v}\right) \right) \left(\frac{\chi_v}{\sqrt{v}} \left(1 + O\left(\frac{1}{v}\right) \right) - 1 \right)^2 \right] \\ &= \phi_0^3 E \left[\frac{\chi_v^3}{v^{3/2}} - 2 \frac{\chi_v^2}{v} + \frac{\chi_v}{\sqrt{v}} \right] \left(1 + O\left(\frac{1}{v}\right) \right) \\ &= \phi_0^3 \left\{ \frac{(v+1)}{v^{3/2}} \sqrt{v} \left(1 + O\left(\frac{1}{v}\right) \right) - 2 + 1 + O\left(\frac{1}{v}\right) \right\} \left(1 + O\left(\frac{1}{v}\right) \right) \\ &= O\left(\frac{1}{v}\right).\end{aligned}$$

We note that

$$\begin{aligned}\frac{\partial}{\partial u}\Phi\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right) &= -\phi\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right) \\ \frac{\partial^2}{\partial u^2}\Phi\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right) &= -\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right)\phi\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right).\end{aligned}$$

Thus,

$$\begin{aligned}|g^{(2)}(u)| &= \left|\frac{\partial^2}{\partial u^2}E\left[\Phi\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right)\right]\right| \\ &= \left|E\left[\frac{\partial^2}{\partial u^2}\Phi\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right)\right]\right| \\ &= \left|E\left[-\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right)\phi\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right)\right]\right| \\ &\leq E\left[\left|-\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right)\right|\phi\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right)\right] \\ &\leq E\left[\left(|t|\sqrt{\frac{\chi_{v-1}^2}{v-1}}+u\right)\phi\left(t\sqrt{\frac{\chi_{v-1}^2}{v-1}}-u\right)\right] \\ &< E\left[\left(|t|\sqrt{\frac{\chi_{v-1}^2}{v-1}}+|u|\right)\right] \\ &\leq \left(|t|E\sqrt{\frac{\chi_{v-1}^2}{v-1}}+|u|\right) \\ &\leq \sqrt{2}|t|+|u|.\end{aligned}$$

Next, we can bound the following integral,

$$\begin{aligned}&\left|\int_0^1\int_0^1 ag^{(2)}(\phi_0+ab(W-\phi_0))dbda\right| \\ &\leq \int_0^1\int_0^1 \left|ag^{(2)}(\phi_0+ab(W-\phi_0))\right|dbda \\ &\leq \int_0^1\int_0^1 \sqrt{2}|t|+|\phi_0+ab(W-\phi_0)|dbda \\ &\leq \int_0^1\int_0^1 \sqrt{2}|t|+\phi_0+|W|dbda \\ &=\sqrt{2}|t|+\phi_0+W.\end{aligned}$$

To calculate $E[P(t_{v-1,W} < t)] = E[g(W)]$, taking the Taylor expansion of $g(W)$ around ϕ_0 , the asymptotic

mean of W , we get

$$\begin{aligned}
&= E[g(W)] \\
&= g(\phi_0) + g^{(1)}(\phi_0)E[(W - \phi_0)] \\
&\quad + E\left[\int_0^1 \int_0^1 ag^{(2)}(\phi_0 + ab(W - \phi_0))dbda (W - \phi_0)^2\right] \\
&= P(t_{v-1, \phi_0} < t) + O\left(\frac{1}{v}\right) \\
&\quad + E\left[\int_0^1 \int_0^1 ag^{(2)}(\phi_0 + ab(W - \phi_0))dbda (W - \phi_0)^2\right].
\end{aligned}$$

Notice that,

$$\begin{aligned}
&\left| E\left[\int_0^1 \int_0^1 ag^{(2)}(\phi_0 + ab(W - \phi_0))dbda (W - \phi_0)^2\right] \right| \\
&\leq E\left[\left|\int_0^1 \int_0^1 ag^{(2)}(\phi_0 + ab(W - \phi_0))dbda (W - \phi_0)^2\right|\right] \\
&\leq E\left[(\sqrt{2}|t| + \phi_0 + W)(W - \phi_0)^2\right] \\
&\leq (\sqrt{2}|t| + \phi_0)E[(W - \phi_0)^2] + E[W(W - \phi_0)^2] \\
&= O\left(\frac{1}{v}\right).
\end{aligned}$$

In conclusion, we have $E[P(t_{v-1, W} < t)] = P(t_{v-1, \phi_0} < t) + O\left(\frac{1}{v}\right)$. Since $t_{v-1}^{(\alpha)} \rightarrow Z^{(\alpha)}$ as $n \rightarrow \infty$, where $Z^{(\alpha)}$ is the $(1 - \alpha)$ th percentile of standard normal, $t_{v-1}^{(\alpha)}$ is bounded. Then, all the above analysis still holds if we replace t with t_{v-1}^α . \square

Let $B(\cdot, \cdot)$ denote the beta function and $I_y(\cdot, \cdot)$ denote the regularized incomplete beta function. In the following, we express $E[P(t_{v-1, W} \leq t)]$ as a sum of infinite series.

Lemma 26. $E[P(t_{v-1, W} \leq t)]$ can be calculated exactly as

$$\begin{aligned}
E[P(t_{v-1, W} < t)] &= \left(\frac{1}{c^2 + 1}\right)^{v/2} \left\{ P(t_{v-1} \leq t) + \right. \\
&\quad \left. \sum_{j=1}^{\infty} \left(\frac{c^2}{c^2 + 1}\right)^{j/2} \frac{1}{jB(j/2, v/2)} \left((-1)^j + I_{\frac{t^2}{t^2 + v - 1}}\left(\frac{j+1}{2}, \frac{v-1}{2}\right)\right) \right\}.
\end{aligned}$$

Proof. Notice that from Walck (1996), the CDF of non-central t -distribution for $t \geq 0$ can be written as

$$P(t_{v-1, W} < t) = \frac{1}{2\sqrt{\pi}} \times \sum_{j=0}^{\infty} \frac{2^{\frac{j}{2}}}{j!} W^j \exp\left\{-\frac{W^2}{2}\right\} \Gamma\left(\frac{j+1}{2}\right) \left((-1)^j + I_z\left(\frac{j+1}{2}, \frac{v-1}{2}\right)\right),$$

where

$$z = \frac{t^2}{t^2 + v - 1}, \quad v = \frac{n(n-3)}{2},$$

$I_y(\cdot, \cdot)$ is the regularized incomplete beta function,

$$W = \frac{\mu}{\sigma} \sqrt{\mathbf{d}^T \mathbf{M} \mathbf{d}} \stackrel{d}{=} c \chi_v, \quad c = \frac{\sigma_{xy}^2}{\sqrt{\sigma_x^2 \sigma_y^2 - \sigma_{xy}^4}}.$$

Next, we calculate the expectation by constructing a generalized gamma distribution,

$$\begin{aligned} & E \left[W^j \exp \left\{ -\frac{W^2}{2} \right\} \right] \\ &= \int_0^\infty w^j \exp \left\{ -\frac{w^2}{2} \right\} \frac{1}{c} \frac{1}{2^{v/2-1} \Gamma(v/2)} \left(\frac{w}{c} \right)^{v-1} \exp \left\{ -\frac{w^2}{2c^2} \right\} dw \\ &= \frac{1}{c^v} \frac{1}{2^{v/2-1} \Gamma(v/2)} \int_0^\infty \exp \left\{ -\left(\frac{w}{\sqrt{2c^2/(c^2+1)}} \right)^2 \right\} w^{j+v-1} dw \\ &= \frac{1}{c^v} \frac{1}{2^{v/2-1} \Gamma(v/2)} \frac{\Gamma(j/2 + v/2) (\sqrt{2c^2/(c^2+1)})^{j+v}}{2} \\ &= \frac{(\sqrt{2c^2/(c^2+1)})^{j+v}}{c^v} \frac{1}{2^{v/2}} \frac{\Gamma(j/2 + v/2)}{\Gamma(v/2)}. \end{aligned}$$

Then,

$$\begin{aligned} E[P(t_{v-1, W} < t)] &= \frac{1}{2\sqrt{\pi}} \left(\sqrt{\frac{1}{c^2+1}} \right)^v \times \\ &\quad \sum_{j=0}^\infty \left(\frac{4c^2}{c^2+1} \right)^{\frac{j}{2}} \frac{\Gamma((j+1)/2) \Gamma(j/2 + v/2)}{j! \Gamma(v/2)} \left((-1)^j + I_z \left(\frac{j+1}{2}, \frac{v-1}{2} \right) \right). \end{aligned}$$

According to the gamma duplicate formula,

$$\Gamma \left(\frac{j+1}{2} \right) = \frac{\sqrt{\pi}}{2^j} \frac{\Gamma(j+1)}{\Gamma(j/2+1)},$$

which further implies that

$$\begin{aligned} \frac{\Gamma((j+1)/2) \Gamma(j/2 + v/2)}{j! \Gamma(v/2)} &= \frac{\sqrt{\pi}}{2^j} \frac{\Gamma(j+1)}{\Gamma(j/2+1)} \frac{\Gamma(j/2 + v/2)}{j! \Gamma(v/2)} \\ &= \begin{cases} \sqrt{\pi}, & j = 0 \\ \frac{\sqrt{\pi}}{2^{j-1}} \frac{1}{j \Gamma(j/2)} \frac{\Gamma(j/2 + v/2)}{\Gamma(v/2)}, & j \geq 1 \end{cases} \\ &= \begin{cases} \sqrt{\pi}, & j = 0 \\ \frac{\sqrt{\pi}}{j 2^{j-1}} \frac{1}{B(j/2, v/2)}, & j \geq 1 \end{cases} \end{aligned}$$

where $B(\cdot, \cdot)$ is the beta function. Then, the expectation can be further simplified as

$$E[P(t_{v-1, W} < t)] = \frac{1}{2} \left(\frac{1}{c^2 + 1} \right)^{v/2} \left(1 + I_z \left(\frac{1}{2}, \frac{v-1}{2} \right) \right) + \left(\frac{1}{c^2 + 1} \right)^{v/2} \sum_{j=1}^{\infty} \left(\frac{c^2}{c^2 + 1} \right)^{\frac{j}{2}} \frac{1}{jB(j/2, v/2)} \left((-1)^j + I_z \left(\frac{j+1}{2}, \frac{v-1}{2} \right) \right).$$

Notice that

$$\frac{1}{2} \left(1 + I_z \left(\frac{1}{2}, \frac{v-1}{2} \right) \right) = P(t_{v-1} \leq t).$$

Thus,

$$E[P(t_{v-1, W} < t)] = \left(\frac{1}{c^2 + 1} \right)^{v/2} \left\{ P(t_{v-1} \leq t) + \sum_{j=1}^{\infty} \left(\frac{c^2}{c^2 + 1} \right)^{j/2} \frac{1}{jB(j/2, v/2)} \left((-1)^j + I_z \left(\frac{j+1}{2}, \frac{v-1}{2} \right) \right) \right\}.$$

□

1.7.9 Proof of Proposition 14

Proof. Since we have

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left(\mathbf{0}, \begin{pmatrix} \mathbf{I}_p & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{XY}^T & \mathbf{I}_q \end{pmatrix} \right),$$

from Theorem 7 in Székely et al. (2007), by setting $c = \frac{1}{4(\pi/3 - \sqrt{3} + 1)}$, we obtain

$$c \leq \frac{dCor^2(x_i, y_j)}{cor^2(x_i, y_j)} \leq 1,$$

$cov^2(x_i, y_j) = cor^2(x_i, y_j)$ and $dCor^2(x_i, y_j) = dCov^2(x_i, y_j)\pi/c$. Combine these results, we have

$$c \leq \frac{dCov^2(x_i, y_j)\pi/c}{cov^2(x_i, y_j)} \leq 1.$$

Notice also that $dCov^2(x_i, x_i) = dCov^2(y_j, y_j) = c/\pi$ and $cov^2(x_i, x_i) = cov^2(y_j, y_j) = 1$. We finally get $0.89^2\phi_2 \leq \phi_1 \leq \phi_2$. □

1.7.10 Proof of Proposition 13

Proof. (i) When $k(x, y) = l(x, y) = |x - y|^2$,

$$k_{st}(i) = -2(x_{si} - E(x_{si}))(x_{ti} - E(x_{ti})),$$

$$l_{st}(j) = -2(y_{sj} - E(y_{sj}))(y_{tj} - E(y_{tj})).$$

Thus, letting $\mathbf{D}_X(i) = ((x_{si}x_{ti})_{s,t=1}^n)_{-D}$ and $\mathbf{D}_Y(j) = ((y_{sj}y_{tj})_{s,t=1}^n)_{-D}$, we have

$$\begin{aligned}
uCov_n^2(\mathbf{X}, \mathbf{Y}) &= \frac{1}{\sqrt{pq}} \sum_{i=1}^p \sum_{j=1}^q (\tilde{\mathbf{K}}(i) \cdot \tilde{\mathbf{L}}(j)) \\
&= \frac{1}{\sqrt{pq}} \sum_{i=1}^p \sum_{j=1}^q 4(\tilde{\mathbf{D}}_X(i) \cdot \tilde{\mathbf{D}}_Y(j)) \\
&= 4 \frac{1}{\sqrt{pq}} \sum_{i=1}^p \sum_{j=1}^q \left\{ \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in i_2^n} x_{si}x_{ti}y_{sj}y_{tj} + \right. \\
&\quad \left. \frac{1}{\binom{n}{4}} \frac{1}{4!} \sum_{(s,t,u,v) \in i_4^n} x_{si}x_{ti}y_{uj}y_{vj} - \frac{2}{\binom{n}{3}} \frac{1}{3!} \sum_{(s,t,u) \in i_3^n} x_{si}x_{ti}y_{sj}y_{uj} \right\} \\
&= 4 \frac{1}{\sqrt{pq}} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j).
\end{aligned}$$

Thus,

$$dCov_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) + \mathcal{R}'_n = \frac{1}{4} \frac{\sqrt{pq}}{\tau} uCov_n^2(\mathbf{X}, \mathbf{Y}) + \mathcal{R}'_n$$

and

$$\begin{aligned}
&\tau \times hCov_n^2(\mathbf{X}, \mathbf{Y}) \\
&= f^{(1)} \left(\frac{\tau_X}{\gamma_X} \right) g^{(1)} \left(\frac{\tau_Y}{\gamma_Y} \right) \frac{\tau_X}{\gamma_X} \frac{\tau_Y}{\gamma_Y} \frac{1}{\tau} \sum_{i=1}^p \sum_{j=1}^q cov_n^2(\mathcal{X}_i, \mathcal{Y}_j) + \mathcal{R}''_n \\
&= f^{(1)} \left(\frac{\tau_X}{\gamma_X} \right) g^{(1)} \left(\frac{\tau_Y}{\gamma_Y} \right) \frac{\tau_X}{\gamma_X} \frac{\tau_Y}{\gamma_Y} \frac{1}{4} \frac{\sqrt{pq}}{\tau} uCov_n^2(\mathbf{X}, \mathbf{Y}) + \mathcal{R}''_n.
\end{aligned}$$

(ii) When $k(x, y) = l(x, y) = |x - y|$, we have

$$\tilde{\mathbf{K}}(i) = \tilde{\mathbf{K}}_1(i) - \tilde{\mathbf{K}}_2(i) - \tilde{\mathbf{K}}_3(i) + \tilde{\mathbf{K}}_4(i) = \tilde{\mathbf{K}}_1(i),$$

where

$$\begin{aligned}
\mathbf{K}_1(i) &= ((k(x_{si}, x_{ti}))_{s,t=1}^n)_{-D}, \mathbf{K}_2(i) = ((E[k(x_{si}, x_{ti})|x_{si}])_{s,t=1}^n)_{-D}, \\
\mathbf{K}_3(i) &= ((E[k(x_{si}, x_{ti})|x_{ti}])_{s,t=1}^n)_{-D}, \mathbf{K}_4(i) = (E[k(x_{si}, x_{ti})])_{s,t=1}^n)_{-D}.
\end{aligned}$$

Similarly, $\tilde{\mathbf{L}}(j) = \tilde{\mathbf{L}}_1(j)$ with $\mathbf{L}_1(j) = (l(y_{sj}, l_{tj}))_{s,t=1}^n$. Then, we have

$$\begin{aligned} uCov_n^2(\mathbf{X}, \mathbf{Y}) &= \frac{1}{\sqrt{pq}} \sum_{i=1}^p \sum_{j=1}^q (\tilde{\mathbf{K}}_1(i) \cdot \tilde{\mathbf{L}}_1(j)) \\ &= \frac{1}{\sqrt{pq}} \sum_{i=1}^p \sum_{j=1}^q dCov_n^2(\mathcal{X}_i, \mathcal{Y}_j) \\ &= \frac{1}{\sqrt{pq}} \frac{1}{\sqrt{\binom{n}{2}}} mdCov_n^2(\mathbf{X}, \mathbf{Y}). \end{aligned}$$

□

1.7.11 Proof of Corollary 1

Proof. For any fixed t and each $R \in \{dCov, hCov, mdCov\}$, Proposition 13 and Theorem 3 imply that

$$T_R \xrightarrow{d} \sqrt{v-1} \frac{\Upsilon}{\sqrt{1-(\Upsilon)^2}}, \text{ where } \Upsilon = \frac{\mathbf{c}^T \mathbf{M} \mathbf{d}}{\sqrt{(\mathbf{c}^T \mathbf{M} \mathbf{c})(\mathbf{d}^T \mathbf{M} \mathbf{d})}}.$$

Then the results follow similarly from the proof of Proposition 10. □

1.7.12 Proof of Remark 17

Proof. For notational convenience, set $z_i = (x_i - x'_i)^2 - E[(x_i - x'_i)^2]$. Since $\sup_i E(x_i^8) < \infty$, we get $\sup_i E(z_i^4) < \infty$. Then, we have

$$\begin{aligned} \alpha_p^2 &\asymp \frac{E\left[(\sum_{i=1}^p z_i)^2\right]}{p^2} \\ &= \frac{E\left[\sum_{s=1}^p \sum_{t \in [s-m, s+m]} z_s z_t\right]}{p^2} \\ &\leq \frac{(2m+1)p}{p^2} \sup_i E(z_i^2) \\ &= O\left(\frac{m}{p}\right) \end{aligned}$$

and

$$\begin{aligned} \gamma_p^2 &\asymp \frac{E\left[(\sum_{i=1}^p z_i)^4\right]}{p^4} \\ &\asymp \frac{m^3 p + m^2 p^2}{p^4} \sup_i E(z_i^4) \\ &= O\left(\frac{m^2}{p^2}\right). \end{aligned}$$

Similarly, we can show that

$$\beta_q^2 = O\left(\frac{m'}{q}\right) \text{ and } \lambda_q^2 = O\left(\frac{m'^2}{q^2}\right).$$

Next, it follows that

$$\tau\alpha_p\lambda_q = O\left(\frac{m'\sqrt{m}}{\sqrt{q}}\right) = o(1).$$

The other results can be proved in a similar fashion. \square

1.7.13 Proof of Theorem 4

Proof. (i)&(ii) Following the proof of Theorem 1, we only need to check that $\mathcal{R}_n = o_p(1)$ still holds as $n \wedge p \wedge q \rightarrow \infty$. Recall that the leading term is $\tau \times (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{L}}_Y)$ and the remainder term is given as

$$\mathcal{R}_n = \frac{1}{2}\tau(\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{R}}_Y) + \frac{1}{2}\tau(\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{L}}_Y) + \tau(\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{R}}_Y).$$

Then, using Equation (1.14), we have

$$\begin{aligned} (\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{R}}_Y) &= \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} L_X(X_s, X_t) R_Y(Y_s, Y_t) \\ &\quad + \frac{1}{\binom{n}{4}} \frac{1}{4!} \sum_{(s,t,u,v) \in \mathbf{i}_4^n} L_X(X_s, X_t) R_Y(Y_u, Y_v) \\ &\quad - \frac{2}{\binom{n}{3}} \frac{1}{3!} \sum_{(s,t,u) \in \mathbf{i}_3^n} L_X(X_s, X_t) R_Y(Y_s, Y_u), \end{aligned}$$

and

$$\begin{aligned} (\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{R}}_Y) &= \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} R_X(X_s, X_t) R_Y(Y_s, Y_t) \\ &\quad + \frac{1}{\binom{n}{4}} \frac{1}{4!} \sum_{(s,t,u,v) \in \mathbf{i}_4^n} R_X(X_s, X_t) R_Y(Y_u, Y_v) \\ &\quad - \frac{2}{\binom{n}{3}} \frac{1}{3!} \sum_{(s,t,u) \in \mathbf{i}_3^n} R_X(X_s, X_t) R_Y(Y_s, Y_u). \end{aligned}$$

To show that \mathcal{R}_n is asymptotically negligible, we consider the events $B_{\mathbf{X}}, B_{\mathbf{Y}}$ and their complements $B_{\mathbf{X}}^c, B_{\mathbf{Y}}^c$, where

$$B_{\mathbf{Y}} = \left\{ \min_{1 \leq s < t \leq n} \frac{|Y_s - Y_t|^2}{\tau_X^2} \leq \frac{1}{2} \text{ or } \max_{1 \leq s < t \leq n} \frac{|Y_s - Y_t|^2}{\tau_X^2} \geq \frac{3}{2} \right\}.$$

Then, under Assumption 16, as $n \wedge p \wedge q \rightarrow \infty$

$$\begin{aligned}
P(B_{\mathbf{Y}}) &= P\left(\min_{1 \leq s < t \leq n} L_Y(Y_s, Y_t) \leq -\frac{1}{2} \text{ or } \max_{1 \leq s < t \leq n} L_Y(Y_s, Y_t) \geq \frac{1}{2}\right) \\
&= P\left(\bigcup_{1 \leq s < t \leq n} \left\{L_Y(Y_s, Y_t) \leq -\frac{1}{2} \text{ or } L_Y(Y_s, Y_t) \geq \frac{1}{2}\right\}\right) \\
&\leq \sum_{1 \leq s < t \leq n} P\left(|L_Y(Y_s, Y_t)| \geq \frac{1}{2}\right) \\
&< n^2 P\left(|L_Y(Y_1, Y_2)| \geq \frac{1}{2}\right) \\
&\leq 4n^2 E[L_Y(Y_1, Y_2)^2] \\
&= o(1).
\end{aligned}$$

Also notice that $P(B_{\mathbf{Y}} B_{\mathbf{X}}^c) \leq P(B_{\mathbf{Y}}) = o(1)$. Similarly, we have $P(B_{\mathbf{X}}) = o(1)$, $P(B_{\mathbf{X}} B_{\mathbf{Y}}^c) = o(1)$ and $P(B_{\mathbf{Y}} B_{\mathbf{X}}) = o(1)$. By the proof of Proposition 2, the remainder term can be written as

$$R_X(X_s, X_t) = \int_0^1 \int_0^1 v f^{(2)}(uv L_X(X_s, X_t)) dudv \times (L_X(X_s, X_t))^2,$$

where $f^{(2)}(t) = -\frac{1}{4}(1+t)^{-\frac{3}{2}}$ and similar formula holds for Y . Conditioned on the event $B_{\mathbf{X}}^c B_{\mathbf{Y}}^c$, we can easily show that

$$|R_X(X_s, X_t)| \leq \frac{\sqrt{2}}{4} (L_X(X_s, X_t))^2, |R_Y(Y_s, Y_t)| \leq \frac{\sqrt{2}}{4} (L_Y(Y_s, Y_t))^2. \quad (1.17)$$

Notice that

$$\begin{aligned}
&\frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} R_X(X_s, X_t) R_Y(Y_s, Y_t) \\
&= \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} R_X(X_s, X_t) R_Y(Y_s, Y_t) \mathbb{I}_{\{B_{\mathbf{X}}^c B_{\mathbf{Y}}^c\}} \\
&\quad + \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} R_X(X_s, X_t) R_Y(Y_s, Y_t) \mathbb{I}_{\{B_{\mathbf{X}} B_{\mathbf{Y}}^c\}} \\
&\quad + \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} R_X(X_s, X_t) R_Y(Y_s, Y_t) \mathbb{I}_{\{B_{\mathbf{X}}^c B_{\mathbf{Y}}\}} \\
&\quad + \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} R_X(X_s, X_t) R_Y(Y_s, Y_t) \mathbb{I}_{\{B_{\mathbf{X}} B_{\mathbf{Y}}\}} \\
&= i + ii + iii + iv.
\end{aligned}$$

For any $\epsilon > 0$, $P(|\tau \times ii| > \epsilon) \leq P(B_{\mathbf{X}} B_{\mathbf{Y}}^c) = o(1)$, which implies that $\tau \times ii = o_p(1)$. Similarly, $\tau \times iii = o_p(1)$ and $\tau \times iv = o_p(1)$. For term i , by Equation (1.17), we have

$$\begin{aligned} |i| &\leq \left\{ \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} |R_X(X_s, X_t) R_Y(Y_s, Y_t)| \right\} B_{\mathbf{X}}^c B_{\mathbf{Y}}^c \\ &\leq \frac{1}{8} \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} L_X(X_s, X_t)^2 L_Y(Y_s, Y_t)^2 \\ &\leq \frac{1}{8} \left\{ \left(\frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} L_X(X_s, X_t)^4 \right) \left(\frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} L_Y(Y_s, Y_t)^4 \right) \right\}^{\frac{1}{2}}. \end{aligned}$$

Next, by the Markov's inequality

$$\begin{aligned} P \left(\frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} L_X(X_s, X_t)^4 > \epsilon \right) &\leq \frac{E \left[\frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} L_X(X_s, X_t)^4 \right]}{\epsilon} \\ &= \frac{1}{\epsilon} E [L_X(X_1, X_2)^4] \\ &= \frac{1}{\epsilon} \gamma_p^2. \end{aligned}$$

Thus, we have $\frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} L_X(X_s, X_t)^4 = O_p(\gamma_p^2)$ and similar proof shows that $\frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} L_Y(Y_s, Y_t)^4 = O_p(\lambda_q^2)$. So, we have $\tau i = O_p(\tau \gamma_p \lambda_q)$ and

$$\tau \frac{1}{\binom{n}{2}} \frac{1}{2!} \sum_{(s,t) \in \mathbf{i}_2^n} R_X(X_s, X_t) R_Y(Y_s, Y_t) = O_p(\tau \gamma_p \lambda_q).$$

Similarly, it can be shown that

$$\begin{aligned} \tau \frac{1}{\binom{n}{4}} \frac{1}{4!} \sum_{(s,t,u,v) \in \mathbf{i}_4^n} R_X(X_s, X_t) R_Y(Y_u, Y_v) &= O_p(\tau \gamma_p \lambda_q), \\ \tau \frac{2}{\binom{n}{3}} \frac{1}{3!} \sum_{(s,t,u) \in \mathbf{i}_3^n} R_X(X_s, X_t) R_Y(Y_s, Y_u) &= O_p(\tau \gamma_p \lambda_q). \end{aligned}$$

In conclusion, we have $\tau(\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{R}}_Y) = O_p(\tau \gamma_p \lambda_q)$. Similarly, it can be shown that $\tau(\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{L}}_Y) = O_p(\tau \alpha_p \beta_q)$, $\tau(\tilde{\mathbf{L}}_X \cdot \tilde{\mathbf{R}}_Y) = O_p(\tau \alpha_p \lambda_q)$ and $\tau(\tilde{\mathbf{R}}_X \cdot \tilde{\mathbf{L}}_Y) = O_p(\tau \gamma_p \beta_q)$. \square

1.7.14 Proof of Theorem 5

Proof. (i)&(ii) Continuing with the proof of Theorem 2, we need to show that $\mathcal{R}_n = o_p(1)$ and $\gamma_{\mathbf{X}}$ is asymptotically equal to τ_X as $n \wedge p \wedge q \rightarrow \infty$ (similar result applies to $\gamma_{\mathbf{Y}}$ and τ_Y). Recall that for all $s \neq t$,

$$L_X(X_s, X_t) = \frac{|X_s - X_t|^2 - \tau_X^2}{\tau_X^2}.$$

Since for any $\epsilon > 0$, under Assumption 16,

$$\begin{aligned}
& P \left(\left| \frac{\text{median}\{|X_s - X_t|^2\}}{\tau_X^2} - 1 \right| > \epsilon \right) \\
& \leq P \left(\min_{1 \leq s < t \leq n} L_X(X_s, X_t) \leq -\epsilon \text{ or } \max_{1 \leq s < t \leq n} L_X(X_s, X_t) \geq \epsilon \right) \\
& = P \left(\bigcup_{1 \leq s < t \leq n} \{L_X(X_s, X_t) \leq -\epsilon \text{ or } L_X(X_s, X_t) \geq \epsilon\} \right) \\
& \leq \sum_{1 \leq s < t \leq n} P(|L_X(X_s, X_t)| \geq \epsilon) \\
& < n^2 P(|L_X(X_1, X_2)| \geq \epsilon) \\
& \leq \frac{1}{\epsilon^2} n^2 E[L_X(X_1, X_2)^2] \\
& = o(1).
\end{aligned}$$

Thus, we have $\frac{\text{median}\{|X_s - X_t|^2\}}{\tau_X^2} \xrightarrow{p} 1$ and $\frac{\tau_X}{\gamma_X} = \sqrt{\frac{\tau_X^2}{\text{median}\{|X_i - X_j|^2\}}} \xrightarrow{p} 1$. Similar arguments can also be used to show that $\frac{\tau_Y}{\gamma_Y} \xrightarrow{p} 1$.

Notice that conditioned on $B_{\mathbf{X}}^c B_{\mathbf{Y}}^c$, for all $1 \leq s < t \leq n$, we have

$$|L_X(X_s, X_t)| < 1/2 \text{ and } \frac{1}{2} < \frac{|X_s - X_t|^2}{\tau_X^2} < \frac{3}{2}. \quad (1.18)$$

Next, Inequalities (1.17) and (1.18) together imply that

$$\left| \frac{\tau_X}{\gamma_X} + uv \left\{ \frac{L_X(X_s, X_t)}{2} + R_X(X_s, X_t) \right\} \frac{\tau_X}{\gamma_X} \right| \leq c,$$

where c is some constant. Since we choose kernels k and l to be the Gaussian or Laplacian kernel, it can be shown that

$$\left| \int_0^1 \int_0^1 v f^{(2)} \left(\frac{\tau_X}{\gamma_X} + uv \left\{ \frac{L_X(X_s, X_t)}{2} + R_X(X_s, X_t) \right\} \frac{\tau_X}{\gamma_X} \right) dudv \right| \leq c',$$

where c' is some constant. Then, we can easily see from Equation (1.15) that $|R_f(X_s, X_t)| \leq c' L_X(X_s, X_t)^2$. Similar result holds for Y . Finally, Theorem 5 can be shown using similar arguments as in the proof of Theorem 4. \square

1.7.15 Proof of Remark 20

Proof. When $k(x, y) = l(x, y) = |x - y|^2$,

$$\begin{aligned}
k_{st}(i) &= -2(x_{si} - E(x_{si}))(x_{ti} - E(x_{ti})), \\
l_{st}(j) &= -2(y_{sj} - E(y_{sj}))(y_{tj} - E(y_{tj})).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
& E[U(X_s, X_t)^2] \\
&= E \left[\frac{1}{p} \sum_{i=1}^p \sum_{j=1}^p k_{st}(i) k_{st}(j) \right] \\
&= \frac{4}{p} \sum_{i=1}^p \sum_{j=1}^p E[(x_{si} - E[x_{si}])(x_{ti} - E[x_{ti}])(x_{sj} - E[x_{sj}])(x_{tj} - E[x_{tj}])] \\
&= \frac{4}{p} \sum_{i=1}^p \sum_{j=1}^p cov^2(x_i, x_j) \\
&= \frac{4}{p} Tr(\Sigma_X^2),
\end{aligned}$$

and

$$\begin{aligned}
& E[U(X_s, X_t)^4] \\
&= E \left[\frac{1}{p^2} \sum_{i,j,r,w=1}^p k_{st}(i) k_{st}(j) k_{st}(r) k_{st}(w) \right] \\
&= \frac{16}{p^2} \sum_{i,j,r,w=1}^p E \left[(x_{si} - E[x_{si}])(x_{ti} - E[x_{ti}])(x_{sj} - E[x_{sj}])(x_{tj} - E[x_{tj}]) \right. \\
&\quad \left. (x_{sr} - E[x_{sr}])(x_{tr} - E[x_{tr}])(x_{sw} - E[x_{sw}])(x_{tw} - E[x_{tw}]) \right] \\
&= \frac{16}{p^2} \sum_{i,j,r,w=1}^p E^2[(x_i - E[x_i])(x_j - E[x_j])(x_r - E[x_r])(x_w - E[x_w])] \\
&\asymp \frac{m^3 p + m^2 p^2}{p^2} \sup_i E^2(x_i^4) \\
&= O(m^2).
\end{aligned}$$

Also,

$$\begin{aligned}
& E[U(X_s, X_t)U(X_t, X_u)U(X_u, X_v)U(X_v, X_s)] \\
&= E \left[\frac{1}{p^2} \sum_{i,j,r,w=1}^p k_{st}(i)k_{tu}(j)k_{uv}(r)k_{vs}(w) \right] \\
&= \frac{16}{p^2} \sum_{i,j,r,w=1}^p E \left[(x_{si} - E[x_{si}]) (x_{ti} - E[x_{ti}]) (x_{tj} - E[x_{tj}]) (x_{uj} - E[x_{uj}]) \right. \\
&\quad \left. (x_{ur} - E[x_{ur}]) (x_{vr} - E[x_{vr}]) (x_{vw} - E[x_{vw}]) (x_{sw} - E[x_{sw}]) \right] \\
&= \frac{16}{p^2} \sum_{i,j,r,w=1}^p \text{cov}(x_i, x_j) \text{cov}(x_j, x_r) \text{cov}(x_r, x_w) \text{cov}(x_w, x_i) \\
&= \frac{16}{p^2} \text{Tr}(\Sigma_X^4) \\
&\asymp \frac{m^3 p}{p^2} \sup_i E^4(x_i^2) \\
&= O\left(\frac{m^3}{p}\right).
\end{aligned}$$

□

1.7.16 Proof of Theorem 6

Proof. Firstly, the following lemma would be useful.

Lemma 27. *Under null, we have*

$$\frac{1}{\mathcal{S}} uCov_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{\binom{n}{2} \mathcal{S}} \sum_{1 \leq s < t \leq n} H(Z_s, Z_t) + \mathcal{R}_n,$$

where $\sqrt{\binom{n}{2}} \mathcal{R}_{n,p,q} = o_p(1)$ as $n \wedge p \wedge q \rightarrow \infty$, $Z_s = (X_s, Y_s)$ and $H(\cdot, \cdot)$ is defined as

$$H(Z_s, Z_t) := U(X_s, X_t)V(Y_s, Y_t).$$

Proof. Firstly, sample $uCov$ can be written as

$$\begin{aligned}
uCov_n^2(\mathbf{X}, \mathbf{Y}) &= \frac{1}{\sqrt{pq}} \sum_{i=1}^p \sum_{j=1}^q (\tilde{\mathbf{K}}(i) \cdot \tilde{\mathbf{L}}(j)) \\
&= \left(\frac{1}{\sqrt{p}} \sum_{i=1}^p \tilde{\mathbf{K}}(i) \right) \cdot \frac{1}{\sqrt{q}} \sum_{j=1}^q \tilde{\mathbf{L}}(j) \\
&= (\tilde{\mathbf{K}} \cdot \tilde{\mathbf{L}}),
\end{aligned}$$

where $\tilde{\mathbf{K}} = (\bar{k}_{st})_{s,t=1}^n$, $\tilde{\mathbf{L}} = (\bar{l}_{st})_{s,t=1}^n$, $\bar{k}_{st} = \frac{1}{\sqrt{p}} \sum_{i=1}^p k(x_{si}, x_{ti})$ and $\bar{l}_{st} = \frac{1}{\sqrt{q}} \sum_{i=1}^q l(y_{si}, y_{ti})$. Thus, $uCov_n^2(\mathbf{X}, \mathbf{Y})$

is just $dCov_n^2(\mathbf{X}, \mathbf{Y})$ with kernel \bar{K} defines as $\bar{K}(X_s, X_t) = \bar{k}_{st}$ and $\bar{L}(Y_s, Y_t) = \bar{l}_{st}$. Notice that

$$\begin{aligned}\bar{K}(X_s, X_t) - E[\bar{K}(X_s, X_t)|X_s] - E[\bar{K}(X_s, X_t)|X_t] + E[\bar{K}(X_s, X_t)] &= \frac{1}{\sqrt{p}} \sum_{i=1}^p k_{st}(i), \\ \bar{L}(Y_s, Y_t) - E[\bar{L}(Y_s, Y_t)|Y_s] - E[\bar{L}(Y_s, Y_t)|Y_t] + E[\bar{L}(Y_s, Y_t)] &= \frac{1}{\sqrt{q}} \sum_{i=1}^q l_{st}(i),\end{aligned}$$

where $k_{st}(i)$ and $l_{st}(i)$ are the double centered kernel distance defined in Section 1.2.2. By Proposition 2.1 of Yao et al. (2018), we have

$$\begin{aligned}\frac{1}{\mathcal{S}}(\tilde{\mathbf{K}} \cdot \tilde{\mathbf{L}}) &= \frac{1}{\binom{n}{2}\mathcal{S}} \sum_{1 \leq s < t \leq n} U(X_{s,n}, X_{t,n})V(Y_{s,n}, Y_{t,n}) + \mathcal{R}_{n,p,q} \\ &= \frac{1}{\binom{n}{2}\mathcal{S}} \sum_{1 \leq s < t \leq n} \frac{1}{\sqrt{p}} \sum_{i=1}^p k_{st}(i) \frac{1}{\sqrt{q}} \sum_{i=1}^q l_{st}(i) + \mathcal{R}_{n,p,q},\end{aligned}$$

where $\sqrt{\binom{n}{2}}\mathcal{R}_{n,p,q} = o_p(1)$ as $n \wedge p \wedge q \rightarrow \infty$. □

By Lemma 27, we have

$$\sqrt{\binom{n}{2}} \frac{uCov_n^2(\mathbf{X}, \mathbf{Y})}{\mathcal{S}} = \frac{1}{\sqrt{\binom{n}{2}}\mathcal{S}} \sum_{1 \leq s < t \leq n} H(Z_s, Z_t) + \sqrt{\binom{n}{2}}\mathcal{R}_{n,p,q},$$

where $\sqrt{\binom{n}{2}}\mathcal{R}_{n,p,q} = o_p(1)$. By similar proof of Theorem 2.1 in Zhang et al. (2018), under H_0 , we have

$$\frac{1}{\sqrt{\binom{n}{2}}\mathcal{S}} \sum_{1 \leq s < t \leq n} H(Z_s, Z_t) \xrightarrow{d} N(0, 1).$$

□

1.7.17 Proof of Proposition 21

Proof. Notice that by the proof of Theorem 2.2 in Zhang et al. (2018), under null

$$\frac{uCov_n^2(\mathbf{X}, \mathbf{X})}{E[U(X, X')^2]} \xrightarrow{p} 1, \quad \frac{uCov_n^2(\mathbf{Y}, \mathbf{Y})}{E[V(Y, Y')^2]} \xrightarrow{p} 1. \quad (1.19)$$

So, by Theorem 6

$$\sqrt{\binom{n}{2}} \frac{uCov_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{uCov_n^2(\mathbf{X}, \mathbf{X})uCov_n^2(\mathbf{Y}, \mathbf{Y})}} \xrightarrow{d} N(0, 1),$$

and also

$$\frac{uCov_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{uCov_n^2(\mathbf{X}, \mathbf{X})uCov_n^2(\mathbf{Y}, \mathbf{Y})}} \xrightarrow{p} 0.$$

As a consequence, we have $T_u \xrightarrow{d} N(0, 1)$. □

1.7.18 Proof of Proposition 22

Proof. Based on Theorem 4 and 5, the results follow similarly from the proof of Proposition 13. \square

1.7.19 Proof of Corollary 2

Proof. (i) If $R = mdCov$, the result follows from Proposition 21 and the following observation

$$\sqrt{\binom{n}{2}} \frac{R_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{R_n^2(\mathbf{X}, \mathbf{X})R_n^2(\mathbf{Y}, \mathbf{Y})}} = \sqrt{\binom{n}{2}} \frac{uCov_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{uCov_n^2(\mathbf{X}, \mathbf{X})uCov_n^2(\mathbf{Y}, \mathbf{Y})}}.$$

(ii) Recall that when $k(x, y) = l(x, y) = |x - y|^2$, $E[U(X_s, X_t)^2] = \frac{4}{p}Tr(\Sigma_X^2)$ and $E[V(Y_s, Y_t)^2] = \frac{4}{q}Tr(\Sigma_Y^2)$. If $R = hCov$, by Proposition 22, we have

$$\sqrt{\binom{n}{2}} \tau \times \frac{R_n^2(\mathbf{X}, \mathbf{Y})}{S} = A_p B_q \sqrt{\binom{n}{2}} \frac{uCov_n^2(\mathbf{X}, \mathbf{Y})}{S} + \sqrt{\binom{n}{2}} \frac{\mathcal{R}_n''}{S},$$

where $A_p = \frac{\sqrt{p}}{2\tau_X} f^{(1)}\left(\frac{\tau_X}{\gamma_X}\right) \frac{\tau_X}{\gamma_X}$ and $B_q = \frac{\sqrt{q}}{2\tau_Y} g^{(1)}\left(\frac{\tau_Y}{\gamma_Y}\right) \frac{\tau_Y}{\gamma_Y}$. By Theorem 6,

$$A_p B_q \sqrt{\binom{n}{2}} \frac{uCov_n^2(\mathbf{X}, \mathbf{Y})}{S} \xrightarrow{d} cN(0, 1),$$

where c is some constant. Also notice that

$$\left| \sqrt{\binom{n}{2}} \frac{\mathcal{R}_n''}{S} \right| \leq \left| \frac{n\mathcal{R}_n''}{4\sqrt{\frac{1}{p}Tr(\Sigma_X^2)\frac{1}{q}Tr(\Sigma_Y^2)}} \right| = o_p(1).$$

Thus, we have

$$\sqrt{\binom{n}{2}} \tau \times \frac{R_n^2(\mathbf{X}, \mathbf{Y})}{S} \xrightarrow{d} cN(0, 1).$$

Next, under Assumption 16, by Equation (1.19) and Proposition 22

$$\tau \times \frac{\sqrt{R_n^2(\mathbf{X}, \mathbf{X})R_n^2(\mathbf{Y}, \mathbf{Y})}}{S} = \sqrt{\left(\frac{A_p^2 uCov_n^2(\mathbf{X}, \mathbf{X}) + \mathcal{R}'''}{E[U(X, X')^2]}\right) \left(\frac{B_q^2 uCov_n^2(\mathbf{Y}, \mathbf{Y}) + \mathcal{R}'''}{E[U(Y, Y')^2]}\right)} \xrightarrow{p} c.$$

Notice that Under Assumptions 1 and 16, Proposition 22 also holds similarly when $\mathbf{X} = \mathbf{Y}$ or $\mathbf{Y} = \mathbf{X}$. So \mathcal{R}''' and \mathcal{R}'''' are both negligible. Thus, we have

$$\sqrt{\binom{n}{2}} \frac{R_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{R_n^2(\mathbf{X}, \mathbf{X})R_n^2(\mathbf{Y}, \mathbf{Y})}} \xrightarrow{d} N(0, 1)$$

and consequently $T_R \xrightarrow{d} N(0, 1)$. Similarly, it can be proved for $R = dCov$. \square

Chapter 2

Interpoint Distance Based Two Sample Tests in High Dimension

2.1 Introduction

In many statistical and machine learning applications, we need inference about the two populations or distributions based on the data samples collected. For example, we need to compare the effectiveness of two newly developed drugs in clinical research, the higher educational level between two countries in a social study and the global warming effects on two regions in environmental science. Two sample hypothesis testing is a statistical procedure to deal with such problems. Formally speaking, having i.i.d. p -dimensional samples $X_1, \dots, X_n \stackrel{d}{=} X \sim F$ and $Y_1, \dots, Y_m \stackrel{d}{=} Y \sim G$, we are interested in knowing whether the underlining distributions F and G which generate the two samples are the same, i.e. to test the following hypothesis,

$$H_0 : F = G \text{ versus } H_A : F \neq G.$$

The study of two-sample testing has a long history and dates back to Kolmogorov-Smirnov's test Kolmogorov (1933); Smirnov (1948), where the empirical CDFs are compared using the sup-norm. Related work for univariate two-sample tests includes Cramer von-Mises criterion Cramér (1928); Von Mises (1928) and Anderson-Darling test Anderson and Darling (1952). Extensions to comparison of multivariate distributions and also the k -sample problem can be found in Bickel (1969); Bickel and Breiman (1983); Friedman and Rafsky (1979); Henze (1988); Schilling (1986) among others. Some other interesting work focusing on the “trimmed” comparison of distributions can be found in Alvarez-Esteban et al. (2008, 2012); Freitag et al. (2007); Munk and Czado (1998).

However, all the afore-mentioned work focuses on the fixed dimensional case. If the dimension exceeds the sample size or is allowed to grow, some of the above methods are expected to fail. For example, the density-based methods suffer from the curse of high dimensionality in particular. In this paper, we study the two sample tests based on certain dissimilarity metrics that can be expressed as functions of the interpoint distances. Two of the most popular high dimensional two-sample tests that fall into this category are based on the Energy Distance (ED) Székely and Rizzo (2004) and the Maximum Mean Discrepancy (MMD) Gretton et al. (2012b). The former is based on the Euclidean distance between sample elements; while the latter is a kernel based method and is basically a variant of ED with a user-specified kernel as distance metric. To be more specific, both ED and MMD take the following form

$$\text{ED}^k(F, G) = 2E[k(X, Y)] - E[k(X, X')] - E[k(Y, Y')], \quad (2.1)$$

where k is a user-specified kernel, X', Y' are i.i.d copies of X, Y respectively. For instance, k can be chosen

as

$$\begin{aligned}
L^2\text{-norm (Euclidean distance)} : \quad & k(X, Y) = \|X - Y\|_2 = \sqrt{\sum_{u=1}^p (x_u - y_u)^2}, \\
\text{Gaussian kernel} : \quad & k(X, Y) = \exp\left(-\frac{\|X - Y\|_2^2}{2\gamma_p^2}\right), \\
\text{Laplacian kernel} : \quad & k(X, Y) = \exp\left(-\frac{\|X - Y\|_2}{\gamma_p}\right), \\
L^1\text{-norm} : \quad & k(X, Y) = \|X - Y\|_1 = \sum_{u=1}^p |x_u - y_u|,
\end{aligned}$$

where $X = (x_1, \dots, x_p)^T$, $Y = (y_1, \dots, y_p)^T$ and γ_p is a user-specified bandwidth parameter. Then, the population version of ED is given by Equation (2.1) with k being the L^2 -norm and the population version of MMD multiplied by -1 is given by Equation (2.1) with k being Gaussian or Laplacian kernel. When k is L^2 -norm, Gaussian or Laplacian kernel, $\text{ED}^k(F, G)$ enjoys the property that $\text{ED}^k(F, G) = 0 \Leftrightarrow F = G$. In fact, $\text{ED}^k(F, G) = 0 \Leftrightarrow F = G$ holds as long as k is a strongly negative definite kernel Klebanov et al. (2005). ED and MMD based tests are both nonparametric without any assumption on the underlying distributions and can be implemented conveniently in practice using permutations. In this work, we aim to address the following questions:

- 1, Can ED^k based permutation test maintain its power against all kinds of alternatives in the high dimensional setting?
- 2, What are the impact of different distance metrics?

To answer the above questions, we conduct rigorous theoretical analysis on the power of $\text{ED}^k(F, G)$ based permutation test in the high dimensional low sample setting (HDLSS) Hall et al. (2005) as well as high dimensional medium sample size setting (HDMSS) Aoshima et al. (2018). Naturally, we say a test is *consistent* if its power goes to 1 under either HDLSS or HDMSS regime. Here, we study the power property of the permutation based tests because they are frequently implemented for Energy Distance and its variants in real life applications.

Let $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$, $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)^T$, $\mathbf{Z} = (\mathbf{X}^T, \mathbf{Y}^T)^T$ denote the sample matrices and $\text{ED}_n^k(\mathbf{Z})$ be a U-statistic based unbiased estimator of $\text{ED}^k(F, G)$. Our main results include: (i) Derivation of the limiting distribution of $\text{ED}_n^k(\mathbf{\Gamma}\mathbf{Z})$ under both low and medium sample size setting, where $\mathbf{\Gamma} \sim \text{Uniform}(\mathbb{P}_{n+m})$ and \mathbb{P}_{n+m} is the set of permutation matrices of dimension $(n+m) \times (n+m)$. (ii) Based on the asymptotic results, we formulate different local alternatives, under which the power behavior of ED_n^k based permutation tests are discussed in detail. (iii) Our theories are applied to existing kernels and statistics, for example

- 1, Under both HDLSS and HDMSS, ED^k based permutation test w.r.t. L^2 -norm, Gaussian and Laplacian kernel are consistent if the sum of component-wise mean or variance differences are not so small, i.e, $\lim_{p \rightarrow \infty} \sum_{u=1}^p (E(x_u) - E(y_u))^2/p \neq 0$ or $\lim_{p \rightarrow \infty} |\sum_{u=1}^p (\text{var}(x_u) - \text{var}(y_u))/p| \neq 0$. In addition, if the sum of component-wise mean and variance differences are both of order $o(\sqrt{p}/\sqrt{nm})$, i.e.,

$$\sum_{u=1}^p (E(x_u) - E(y_u))^2 = o\left(\frac{\sqrt{p}}{\sqrt{nm}}\right) \text{ and } \left| \sum_{u=1}^p (\text{var}(x_u) - \text{var}(y_u)) \right| = o\left(\frac{\sqrt{p}}{\sqrt{nm}}\right),$$

these tests suffer substantial power loss (the limits of their power are derived) under HDLSS and have trivial power (power no larger than the significance level) under HDMSS. Furthermore, under HDLSS, the afore-mentioned tests have trivial power if additionally we have $\sum_{u,v=1}^p (\text{cov}(x_u, x_v) - \text{cov}(y_u, y_v))^2 = o(p)$.

- 2, When k is chosen as L^1 -norm, ED^k based permutation test experiences a power drop under HDLSS and trivial power under HDMSS if X, Y have the same univariate marginal distribution, i.e. $x_u =^d y_u$ for $u = 1, 2, \dots, p$. This phenomenon is consistent with the fact that ED^k with L^1 -norm can characterise the discrepancies between the marginal univariate distributions. In addition, Under HDLSS, we show that the L^1 -norm based test has trivial power when X and Y have the same bivariate marginal distribution, i.e., $(x_u, x_v) =^d (y_u, y_v), u, v = 1, \dots, p$.

These findings are further corroborated in our simulation study. It is worth mentioning that Chakraborty and Zhang (2019b) investigate the energy distance, maximum mean discrepancy, distance covariance and Hilbert-Schmidt Independence Criterion in the high dimensional setting. They propose a new class of metrics which can detect/measure the equality of low-dimensional marginal distributions and a computational efficient t -test is further proposed based on the new metric. By contrast, our focus is on kernel-based permutation test and their asymptotic power properties in the high dimensional setting. In the following we introduce some notation and define some frequently used operators for later convenience.

2.1.1 Notation

Here, random data samples are denoted as, for each $i = 1, 2, \dots, n$, $X_i =^d X = (x_1, \dots, x_p)^T \in \mathbb{R}^p$ and for each $j = 1, 2, \dots, m$, $Y_j =^d Y = (y_1, \dots, y_p)^T \in \mathbb{R}^p$. Next, let $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$, $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)^T$ and $\mathbf{Z} = (\mathbf{X}^T, \mathbf{Y}^T)^T = (Z_1, Z_2, \dots, Z_{n+m})^T$ denote the random sample matrices. Furthermore, let $\mathbb{P}_{n+m} = \{\Gamma_1, \Gamma_2, \dots, \Gamma_{(n+m)!}\}$ be the group containing all permutation matrices of dimension $(n+m) \times (n+m)$ and for each i , let π_i be the permutation that corresponds to Γ_i via

$$\Gamma_i \mathbf{Z} = (Z_{(\pi_i(1))}, Z_{(\pi_i(2))}, \dots, Z_{(\pi_i(n+m))})^T$$

where $(\pi_i(1)) < \dots < (\pi_i(n+m))$ is the ranked sequence of $\{\pi_i(1), \pi_i(2), \dots, \pi_i(n+m)\}$. For a random permutation matrix $\Gamma \sim \text{uniform}(\mathbb{P}_{n+m})$, we use $\boldsymbol{\pi}$ to represent its corresponding permutation. Next, given any function φ , $\varphi^{(i)}$ is used to denote its i -th order derivative. Finally, calligraphic letters $(\mathcal{K}, \mathcal{L}, \mathcal{R}, \mathcal{W}, \mathcal{G})$ are used to denote self-defined operators that act on random variables to produce random variables.

2.2 Interpoint Distance Based Two Sample Tests

In this paper, we limit our attention to $\text{ED}^k(F, G)$, where k is a user specified dissimilarity metric Sarkar et al. (2018) of the following form

$$k(X, Y) = \varphi \left\{ \frac{1}{p} \sum_{u=1}^p \psi(x_u, y_u) \right\}, \quad (2.2)$$

where $\psi \geq 0$ and φ has continuous second order derivative on $(0, +\infty)$. The reason we focus on $\text{ED}^k(F, G)$ of the above form is that the metric k encompasses many well-known distance metrics such as L^2 -norm, L^1 -norm, Gaussian and Laplacian kernel. Consequently, Energy Distance (ED) and Maximum Mean Discrepancy (MMD) are just special cases of $\text{ED}^k(F, G)$. We summarize the commonly used distance metrics in Table 2.1. Following the literatures Gretton et al. (2012b, 2008), we consider the bandwidth parameter γ in Gaussian and Laplacian kernel as a fixed constant. Notice that if k is some well-known distance metrics such as L^2 -norm, Gaussian kernel (multiplied by -1) and Laplacian kernel (multiplied by -1), a nice property

$\psi(x, y)$	$\varphi(x)$	k	$\text{ED}^k(F, G)$
$(x - y)^2$	\sqrt{x}	L^2 -norm	Energy distance (ED) Székely and Rizzo (2004)
	$-e^{-\frac{x}{2\gamma^2}}$	Gaussian kernel (multiplied by -1)	Maximum Mean Discrepancy (MMD) Gretton et al. (2012b)
	$-e^{-\frac{\sqrt{x}}{\gamma}}$	Laplacian kernel (multiplied by -1)	
$ x - y $	x	L^1 -norm	Used for some graph-based tests Sarkar et al. (2018)

Table 2.1: Correspondence between different choices of ψ, φ and existing distance metrics as well as two sample test statistics in the literature.

for ED^k is that

$$\text{ED}^k(F, G) \geq 0 \text{ and } \text{ED}^k(F, G) = 0 \Leftrightarrow F = G. \quad (2.3)$$

Here, it is just for the ease of presentation and notational simplicity that k is set to be Gaussian or Laplacian kernel multiplied by -1. In fact, if k is a universal kernel (see Theorem 5 and Lemma 1 of Gretton et al. (2012b)) or k is a strongly negative definite kernel (see Theorem 1.9 Klebanov et al. (2005)), Property (2.3) still holds. On the other hand, using $\text{ED}^1(F, G)$ to denote $\text{ED}^k(F, G)$ when k is the L^1 -distance, we observe that $\text{ED}^1(F, G) = \sum_{u=1}^p \text{ED}(F_u, G_u)$, from which it easily follows that

$$\text{ED}^1(F, G) \geq 0 \text{ and } \text{ED}^1(F, G) = 0 \Leftrightarrow F_u = G_u \text{ for all } u = 1, 2, \dots, p.$$

Notice that it is possible to have $F_u = G_u$ for all $u = 1, 2, \dots, p$ but $F \neq G$, under which we have $\text{ED}^1(F, G) = 0$ while $\text{ED}^k > 0$ if k is L^2 -norm, Gaussian kernel (multiplied by -1) or Laplacian kernel (multiplied by -1). Thus, L^2 -norm, Gaussian kernel or Laplacian kernel based test statistics have advantage over L^1 -norm based test statistic in the low dimensional setting, but we will see later that the story is in a sense reversed under the high dimensional setting. Next, an unbiased estimator of ED^k is given as

$$\text{ED}_n^k(\mathbf{Z}) = \frac{2}{mn} \sum_{i=1}^n \sum_{j=1}^m k(X_i, Y_j) - \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} k(X_i, X_j) - \frac{2}{m(m-1)} \sum_{1 \leq i < j \leq m} k(Y_i, Y_j).$$

2.3 Power Analysis for Permutation Test

As permutation tests are commonly used for Energy Distance and kernel variants in practice due to their implementational convenience and accurate size, we study their asymptotic behavior under the high dimensional setting in this subsection. Since we have i.i.d samples, after we permute the data, i.e., shuffle the rows of \mathbf{Z} as $\Gamma_i \mathbf{Z}$ by some permutation matrix Γ_i , what really matters to the distribution of $\text{ED}_n^k(\Gamma_i \mathbf{Z})$ is how many X samples stay in the first n rows. Formally, let $|\mathbb{A}|$ be the cardinality of the set \mathbb{A} and given a

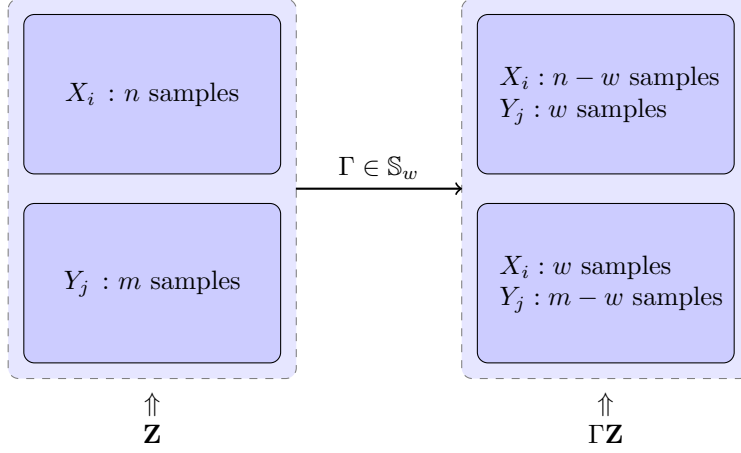


Figure 2.1: Illustration of the Permutation Procedure.

permutation matrix Γ_i with the corresponding permutation π_i , set

$$N(\Gamma_i) = |\{j \in \{1, 2, \dots, n\} : 1 \leq j \leq n, n+1 \leq \pi_i(j) \leq n+m\}|.$$

The integer $n - N(\Gamma_i)$ actually counts the number of samples which belong to the first n rows of \mathbf{Z} both before and after the permutation Γ_i . Notice that it is possible that $N(\Gamma_i) = N(\Gamma_j)$ for different permutations Γ_i and Γ_j . The set \mathbb{S}_w collects all the permutations Γ_i such that $N(\Gamma_i) = w$. Mathematically, fix $0 \leq w \leq \min\{n, m\}$, set $\mathbb{S}_w = \{\Gamma_i : N(\Gamma_i) = w, i = 1, 2, \dots, (n+m)!\}$, then

$$|\mathbb{S}_w| = \binom{m}{w} \binom{n}{n-w} n!m!.$$

To differentiate from Γ_i , we use italic symbol Γ_w to represent an element in \mathbb{S}_w . Intuitively, $|\mathbb{S}_w|$ is the number of permutations that would have $n - w$ samples stay in the first n rows of \mathbf{Z} after we apply the corresponding permutation. The above process is further illustrated in Figure 2.1.

For the inter-point distance based two sample tests, we can equivalently permute the weights on the pair-wise distances instead of permuting data points, i.e., for a fixed permutation matrix $\Gamma_s \in \mathbb{P}_{n+m}$ that corresponds to π_s , we can write $\text{ED}_n^k(\Gamma_s \mathbf{Z})$ as

$$\text{ED}_n^k(\Gamma_s \mathbf{Z}) = \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{s,ij} k(Z_i, Z_j), \quad (2.4)$$

where $\Pi_{s,ij}$ is defined as

$$\Pi_{s,ij} = \begin{cases} -\frac{2}{n(n-1)}, & 1 \leq \pi_s(i), \pi_s(j) \leq n, \\ -\frac{2}{m(m-1)}, & n+1 \leq \pi_s(i), \pi_s(j) \leq n+m, \\ \frac{2}{mn}, & 1 \leq \pi_s(i) \leq n, n+1 \leq \pi_s(j) \leq n+m, \\ \frac{2}{mn}, & n+1 \leq \pi_s(i) \leq n+m, 1 \leq \pi_s(j) \leq m. \end{cases}$$

To formally define the permutation test for $\text{ED}_n^k(\mathbf{Z})$, let \hat{R} denote the randomization distribution of $\text{ED}_n^k(\mathbf{Z})$,

which is defined by

$$\widehat{R}(t) = \frac{1}{(n+m)!} \sum_{i=1}^{(n+m)!} \mathbb{I}_{\{\text{ED}_n^k(\Gamma_i \mathbf{Z}) \leq t\}} = \frac{1}{(m+n)!} \sum_{w=0}^{\min\{n,m\}} \sum_{\Gamma \in \mathbb{S}_w} \mathbb{I}_{\{\text{ED}_n^k(\Gamma \mathbf{Z}) \leq t\}}.$$

For any distribution F , let the $(1-\alpha)$ -th quantile of F be denoted by $Q_{F,1-\alpha}$. In particular, the $(1-\alpha)$ th quantile of \widehat{R} is $Q_{\widehat{R},1-\alpha}$, i.e.

$$Q_{\widehat{R},1-\alpha} = \widehat{R}^{-1}(1-\alpha) = \inf \left\{ t : \widehat{R}(t) \geq 1-\alpha \right\}. \quad (2.5)$$

Then, the level- α permutation test w.r.t. $\text{ED}_n^k(\mathbf{Z})$ is defined as

$$\text{Reject } H_0 \text{ if } \text{ED}_n^k(\mathbf{Z}) > Q_{\widehat{R},1-\alpha}.$$

In real life applications, $(n+m)!$ might be large, we thus resort to an approximation of $Q_{\widehat{R},1-\alpha}$. Let $\Gamma_1, \dots, \Gamma_S$ be i.i.d and uniformly sampled from \mathbb{P}_{n+m} and we approximate the critical value by $Q_{\widetilde{R},1-\alpha}$, where

$$\widetilde{R}(t) := \frac{1}{S} \left(\mathbb{I}_{\{\text{ED}_n^k(\mathbf{Z}) \leq t\}} + \sum_{i=1}^{S-1} \mathbb{I}_{\{\text{ED}_n^k(\Gamma_i \mathbf{Z}) \leq t\}} \right).$$

2.3.1 Local Alternatives

In this subsection, we define different local alternatives, under which the $\text{ED}_n^k(\mathbf{Z})$ based permutation test will be consistent, have a nontrivial power limit and exhibit trivial power (power no larger than the significance level α) in the limit. To formally define the local alternative hypothesis, let the operator \mathcal{K} be defined as

$$\mathcal{K}(Z_i, Z_j) = \frac{1}{\sqrt{p}} \sum_{u=1}^p \left\{ \psi(z_{iu}, z_{ju}) - E[\psi(z_{iu}, z_{ju}) | z_{iu}] - E[\psi(z_{iu}, z_{ju}) | z_{ju}] + E[\psi(z_{iu}, z_{ju})] \right\}, \quad (2.6)$$

It follows from Proposition 2.2.1 of Zhu et al. (2019) that $E[\mathcal{K}(Z_i, Z_j)\mathcal{K}(Z_{i'}, Z_{j'})] = 0$ if $\{i, j\} \neq \{i', j'\}$. Next, denote the average distance over components as

$$\overline{\psi}(Z_i, Z_j) = \frac{1}{p} \sum_{u=1}^p \psi(z_{iu}, z_{ju}).$$

In addition, we need to assume the existence of some constants to properly define the local alternatives. These constants will also appear in the limiting distribution of our test statistics.

Assumption 28. *As $p \rightarrow \infty$, assume the existence of the limiting mean*

$$e_x = \lim_{p \rightarrow \infty} E[\overline{\psi}(X, X')], e_y = \lim_{p \rightarrow \infty} E[\overline{\psi}(Y, Y')] \text{ and } e_{xy} = \lim_{p \rightarrow \infty} E[\overline{\psi}(X, Y)]$$

and also the limiting variances

$$v_x = \lim_{p \rightarrow \infty} \text{var}[\mathcal{K}(X, X')], v_y = \lim_{p \rightarrow \infty} \text{var}[\mathcal{K}(Y, Y')] \text{ and } v_{xy} = \lim_{p \rightarrow \infty} \text{var}[\mathcal{K}(X, Y)].$$

k	H_{A_c} Characterization
L^2 -norm	$H_{A_c} = \left\{ (F, G) \left \begin{array}{l} \sum_{u=1}^p (E(x_u) - E(y_u))^2 = o(p) \text{ and} \\ \sum_{u=1}^p (\text{var}(x_u) - \text{var}(y_u)) = o(p) \end{array} \right. \right\}^c$
Gaussian kernel	
Laplacian kernel	
L^1 -norm	$H_{A_c} = \{(F, G) \sum_{u=1}^p \text{ED}(F_u, G_u) = o(p)\}^c$

Table 2.2: Characterization of H_{A_c} for some specific metrics.

Then, we are ready to define the consistency space H_{A_c} , under which the $\text{ED}_n^k(\mathbf{Z})$ implemented as permutation test can be shown to be consistent under both HDLSS and HDMSS settings.

$$H_{A_c} := \{(F, G) \mid 2\varphi(e_{xy}) \neq \varphi(e_x) + \varphi(e_y)\}.$$

We use \mathbb{A}^c to denote the complement of any given set \mathbb{A} and denote $F = (F_1, F_2, \dots, F_p)$ and $G = (G_1, G_2, \dots, G_p)$, where $F_u, G_u, u = 1, 2, \dots, p$ are the marginal univariate distributions. For commonly used kernels, we have Table 2.2 characterizing H_{A_c} and the proof is postponed to subsection 2.6.1. Then, we present the space H_{A_t} , under which the normal limit of $\text{ED}_n^k(\mathbf{Z})$ can be derived under both HDLSS and HDMSS.

$$H_{A_t} := \left\{ (F, G) \left| \begin{array}{l} e_{xy} = e_x = e_y, \\ |2E[\bar{\psi}(X, Y)] - E[\bar{\psi}(X, X')] - E[\bar{\psi}(Y, Y')]| = o(\sqrt{\frac{1}{nmp}}), \\ E[|E[\bar{\psi}(X, Y)|X] - E[\bar{\psi}(X, X')|X]|] = o(\sqrt{\frac{1}{nmp}}) \text{ and} \\ E[|E[\bar{\psi}(X, Y)|Y] - E[\bar{\psi}(Y, Y')|Y]|] = o(\sqrt{\frac{1}{nmp}}). \end{array} \right. \right\}.$$

Under H_{A_t} , a limit for the power of $\text{ED}_n^k(\mathbf{Z})$ (implemented as permutation test) is derived under HDLSS. On the other hand, its power is shown to be trivial (no larger than the significance level α) under HDMSS and H_{A_t} . Next, we provide sufficient conditions for $(F, G) \in H_{A_t}$ with respect to some well known kernels in Table 2.3. Then, the set of distributions H_{A_t} is defined as

$$H_{A_t} := \{(F, G) \mid (F, G) \in H_{A_t}, v_{xy} = v_x = v_y\}.$$

It can be shown that under H_{A_t} , the $\text{ED}_n^k(\mathbf{Z})$ based permutation test has power no larger than the significance level α for both HDLSS and HDMSS settings. Table 2.4 provide sufficient conditions for being in H_{A_t} with

k	Sufficient conditions for H_{A_t}
L^2 -norm	$\left\{ (F, G) \left \begin{array}{l} \sum_{u=1}^p (E(x_u) - E(y_u))^2 = o(\sqrt{\frac{p}{nm}}) \text{ and} \\ \sum_{u=1}^p (\text{var}(x_u) - \text{var}(y_u)) = o(\sqrt{\frac{p}{nm}}) \end{array} \right. \right\} \subseteq H_{A_t}$
Gaussian kernel	
Laplacian kernel	
L^1 -norm	$\{(F, G) \mid F_u = G_u, u = 1, 2, \dots, p\} \subseteq H_{A_t}$

Table 2.3: Sufficient conditions for H_{A_t} with respect to some specific metrics.

k	Sufficient conditions for H_{A_t}	
L^2 -norm	$\left\{ (F, G) \mid \begin{array}{l} \sum_{u=1}^p (E(x_u) - E(y_u))^2 = o(\sqrt{\frac{p}{nm}}), \\ \sum_{u=1}^p (\text{var}(x_u) - \text{var}(y_u)) = o(\sqrt{\frac{p}{nm}}) \text{ and } \\ \sum_{u,v=1}^p (\text{cov}(x_u, x_v) - \text{cov}(y_u, y_v))^2 = o(p) \end{array} \right\} \subseteq H_{A_t}$	
Gaussian kernel		
Laplacian kernel		
L^1 -norm	$\{(F, G) \mid (x_u, x_v) =^d (y_u, y_v), u, v = 1, \dots, p\} \subseteq H_{A_t}$	

Table 2.4: Sufficient conditions for H_{A_t} with respect to some specific metrics.

respect to some specific dissimilarity metrics. Comparing the three local alternatives, it follows from the definition of $H_{A_c}, H_{A_l}, H_{A_t}$ that $H_{A_c}^c \supseteq H_{A_l} \supseteq H_{A_t}$. We also want to remark that it holds for arbitrary function φ and ψ that

$$\begin{aligned} \{(F, G) \mid F_u = G_u, u = 1, 2, \dots, p\} &\subseteq H_{A_t}, \\ \{(F, G) \mid (x_u, x_v) =^d (y_u, y_v), u, v = 1, \dots, p\} &\subseteq H_{A_t}. \end{aligned}$$

2.3.2 High Dimensional Low Sample Size (HDLSS)

The analysis in this subsection is conducted under the high dimensional low sample size setting (HDLSS), i.e., n, m are fixed constants and we let $p \rightarrow \infty$. Our final goal is to study the power of $\text{ED}_n^k(\mathbf{Z})$ based permutation test under various local alternatives. To this end, we need the following assumption. Recall the operator \mathcal{K} is defined in (2.6).

Assumption 29. For fixed n, m , as $p \rightarrow \infty$,

$$\begin{pmatrix} \mathcal{K}(X_i, Y_j) \\ \mathcal{K}(X_{i_1}, X_{i_2}) \\ \mathcal{K}(Y_{j_1}, Y_{j_2}) \end{pmatrix}_{i,j,i_1 < i_2, j_1 < j_2} \xrightarrow{d} \begin{pmatrix} b_{ij} \\ c_{i_1 i_2} \\ d_{j_1 j_2} \end{pmatrix}_{i,j,i_1 < i_2, j_1 < j_2},$$

where $\{b_{ij}, c_{i_1 i_2}, d_{j_1 j_2}\}_{i,j,i_1 < i_2, j_1 < j_2}$ are uncorrelated and jointly Gaussian with mean 0 and variances $\text{var}(b_{ij}) = v_{xy}$, $\text{var}(c_{i_1 i_2}) = v_x$, $\text{var}(d_{j_1 j_2}) = v_y$.

Remark 30. The above multi-dimensional CLT result is classical and can be derived under suitable moment and weak dependence assumptions on the components of X and Y .

In the above assumption, it is due to the use of double centered distance $\mathcal{K}(Z_i, Z_j)$ that the asymptotic covariance matrix is diagonal. Then, to provide some insights, the first step of our power analysis is the Taylor expansion w.r.t φ up to the second order, i.e., for $i \neq j$

$$k(Z_i, Z_j) = \varphi(e_{ij}) + \varphi^{(1)}(e_{ij})\mathcal{L}(Z_i, Z_j) + \mathcal{R}_2(Z_i, Z_j),$$

where $\mathcal{L}(Z_i, Z_j) := \bar{\psi}(Z_i, Z_j) - e_{ij}$ is an operator that acts on random variables, $\mathcal{R}_2(Z_i, Z_j)$ is the remainder

and

$$e_{ij} = \begin{cases} e_x, & \text{if } 1 \leq i, j \leq n, \\ e_y, & \text{if } n+1 \leq i, j \leq n+m, \\ e_{xy}, & \text{otherwise.} \end{cases}$$

In order to control the remainder term, we need assumptions about the decay rate of $E[\mathcal{L}^2(Z_i, Z_j)]$. Thus, we set

$$\alpha_x^2 = E[\mathcal{L}^2(X, X')], \quad \alpha_y^2 = E[\mathcal{L}^2(Y, Y')] \quad \text{and} \quad \alpha_{xy}^2 = E[\mathcal{L}^2(X, Y)].$$

It then follows from Markov's inequality that $\mathcal{L}(X, Y) = O_p(\alpha_{xy})$, $\mathcal{L}(X, X') = O_p(\alpha_x)$ and $\mathcal{L}(Y, Y') = O_p(\alpha_y)$. Then, our next two assumptions are used to control the remainder terms induced by taking the Taylor expansion.

Assumption 31. $\alpha_{xy}^2 = o(1)$, $\alpha_x^2 = o(1)$ and $\alpha_y^2 = o(1)$.

Assumption 32. $\sqrt{p}\alpha_{xy}^2 = o(1)$, $\sqrt{p}\alpha_x^2 = o(1)$ and $\sqrt{p}\alpha_y^2 = o(1)$.

Remark 33. To gain some insights into the above assumptions, a straightforward calculation yields

$$\alpha_{xy}^2 = \frac{1}{p^2} \sum_{u,v=1}^p \text{cov}(\psi(x_u, y_u), \psi(x_v, y_v)) + \left(\frac{\sum_{u=1}^p E[\psi(x_u, y_u)]}{p} - e_{xy} \right)^2.$$

Therefore, we have $\sqrt{p}\alpha_{xy}^2 = o(1)$ if the component-wise dependencies of both X and Y are not so strong. For illustration purpose, suppose X and Y are κ -dependent weak stationary time series, i.e., $x_u \perp x_v$ and $y_u \perp y_v$ if $|u - v| > \kappa$. Then, if $\max_u E[\psi^2(x_u, y_u)] < \infty$, it is easy to see that $\alpha_{xy}^2 = O(\kappa/p)$ and as a consequence, Assumption 32 is satisfied as long as $\kappa/\sqrt{p} = o(1)$. In addition, it is indeed fairly straightforward to verify the above result when the sequence $\{(x_u, y_u)\}_{u=1}^p$ is α -mixing with geometrically decaying coefficients.

Remark 34. When $\psi(x, y) = (x - y)^2$, some algebra shows that

$$\begin{aligned} \sum_{u,v=1}^p \text{cov}(\psi(x_u, y_u), \psi(x_v, y_v)) &= \text{var}(X^T X) + \text{var}(Y^T Y) \\ &\quad + 4\text{var}(X^T Y) - 4\text{cov}(X^T X, X^T E[Y]) - 4\text{cov}(Y^T Y, Y^T E[X]). \end{aligned}$$

Thus, suppose $\sum_{u=1}^p E[\psi(x_u, y_u)]/p - e_{xy} = o(p^{-1/4})$ and if $\text{var}(X^T X)$, $\text{var}(Y^T Y)$, $\text{var}(X^T Y)$, $\text{var}(X^T E[Y])$, $\text{var}(E[X]^T Y)$ all have order $o(p^{1.5})$, we have $\sqrt{p}\alpha_{xy}^2 = o(1)$.

In the next theorem, we state the asymptotic behavior of $\text{ED}_n^k(\Gamma_w \mathbf{Z})$ for each fixed permutation matrix $\Gamma_w \in \mathbb{S}_w$. Here, we use the italic gamma $\Gamma_w \in \mathbb{S}_w$ to differentiate from $\Gamma_i \in \mathbb{P}_{n+m}$.

Theorem 7. For fixed $\Gamma_w \in \mathbb{S}_w$,

(i) Under Assumptions 28 and 31,

$$\text{ED}_n^k(\Gamma_w \mathbf{Z}) \xrightarrow{P} \mu_{n,w},$$

where $\mu_{n,w}$ is defined as

$$\mu_{n,w} := \mu_n(\Gamma_w \mathbf{Z}) = (2\varphi(e_{xy}) - \varphi(e_x) - \varphi(e_y)) \times \left\{ 1 - \left(\frac{2m-1}{m(m-1)} + \frac{2n-1}{n(n-1)} \right) w + \left(\frac{2}{mn} + \frac{1}{n(n-1)} + \frac{1}{m(m-1)} \right) w^2 \right\}.$$

(ii) Under Assumptions 28, 29, 32 and local alternative H_{A_1} ,

$$\sqrt{p} \left(\text{ED}_n^k(\Gamma_w \mathbf{Z}) - \mu_n(\Gamma_w \mathbf{Z}) \right) \xrightarrow{d} N(0, \sigma_{n,w}^2).$$

where $\sigma_{n,w}^2$ is given as

$$\begin{aligned} \sigma_{n,w}^2 &:= \sigma_n(\Gamma_w \mathbf{Z}) \\ &= \left\{ \frac{4}{nm} - 4 \left(\frac{n+m}{n^2 m^2} - \frac{n}{n^2(n-1)^2} - \frac{m}{m^2(m-1)^2} \right) w \right. \\ &\quad \left. + 4 \left(\frac{2}{n^2 m^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) w^2 \right\} v_{xy}[\varphi^{(1)}(e_{xy})]^2 \\ &\quad + \left\{ \frac{2}{n(n-1)} + 2 \left(\frac{2n}{n^2 m^2} - \frac{2n-1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) w \right. \\ &\quad \left. - 2 \left(\frac{2}{m^2 n^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) w^2 \right\} v_x[\varphi^{(1)}(e_x)]^2 \\ &\quad + \left\{ \frac{2}{m(m-1)} + 2 \left(\frac{2m}{n^2 m^2} - \frac{1}{n^2(n-1)^2} - \frac{2m-1}{m^2(m-1)^2} \right) w \right. \\ &\quad \left. - 2 \left(\frac{2}{n^2 m^2} - \frac{1}{m^2(m-1)^2} - \frac{1}{n^2(n-1)^2} \right) w^2 \right\} v_y[\varphi^{(1)}(e_y)]^2. \end{aligned}$$

We use $W \sim \text{Hypergeometric}(n+m, m, n)$ to denote that W follows the hypergeometric distribution, which describes the probability of n draws from a union of two groups (one group has m elements, the other has n elements) such that w of them are chosen from the group of size m . To be precise, W has probability mass function

$$P(W = w) = \frac{\binom{m}{w} \binom{n}{m-w}}{\binom{n+m}{n}} \text{ for } w \in \{0, 1, \dots, \min\{n, m\}\}.$$

Then, the limiting distribution of $\text{ED}_n^k(\Gamma \mathbf{Z})$ is derived in the following proposition.

Proposition 35. For $\Gamma \sim \text{Uniform}(\mathbb{P}_{n+m})$, which is independent of the data, let

$$W := N(\Gamma) \sim \text{Hypergeometric}(n+m, m, n).$$

(i) Under Assumptions 28 and 31,

$$\text{ED}_n^k(\Gamma \mathbf{Z}) \xrightarrow{p} \mu_{n,W}.$$

(ii) Under Assumptions 28, 29, 32 and local alternative H_{A_1} ,

$$\sqrt{p} \left(\text{ED}_n^k(\Gamma \mathbf{Z}) - \mu_{n,W} \right) \xrightarrow{d} N(0, \sigma_{n,W}^2).$$

In the above proposition, $N(0, \sigma_{n,W}^2)$ should be understood as a mixture of Gaussian with probability distribution

$$P(N(0, \sigma_{n,W}^2) \leq a) = \sum_{w=1}^{\min\{n,m\}} P(W=w) P(N(0, \sigma_{n,w}^2) \leq a).$$

Next, let Γ_0 corresponds to the identity permutation map, we present the power behavior of $\text{ED}^k(\mathbf{Z})$ when the critical values are obtained via permutations.

Theorem 8. Assume that $2\varphi(e_{xy}) \geq \varphi(e_x) + \varphi(e_y)$.

1, **[Consistency]** Suppose Assumptions 28 and 31 hold.

(i) If the critical value is chosen as $Q_{\hat{R}, 1-\alpha}$. Let n, m be large enough such that $n!m!/(n+m)! < 1-\alpha$ if $m \neq n$ and $2(n!)^2/(2n)! < 1-\alpha$ if $m = n$. Then, we have

$$\lim_{p \rightarrow \infty} P_{H_{A_c}} \left(\text{ED}_n^k(\mathbf{Z}) > Q_{\hat{R}, 1-\alpha} \right) = 1,$$

which means that the asymptotic power of ED^k based permutation test is 1 as p goes to infinity.

(ii) If the critical value is chosen as $Q_{\tilde{R}, 1-\alpha}$. Then, we have

$$\lim_{p \rightarrow \infty} P_{H_{A_c}} \left(\text{ED}_n^k(\mathbf{Z}) > Q_{\tilde{R}, 1-\alpha} \right) \geq \begin{cases} 1 - \frac{S-1}{[\alpha S]} \frac{n!m!}{(n+m)!}, & \text{if } n \neq m, \\ 1 - \frac{S-1}{[\alpha S]} \frac{2(n!)^2}{(n+m)!}, & \text{if } n = m. \end{cases}$$

2, **[Power Limit]** Suppose Assumptions 28, 29, 32 hold.

(i) If the critical value is chosen as $Q_{\hat{R}, 1-\alpha}$. Then, we have

$$\lim_{p \rightarrow \infty} P_{H_{A_l}} \left(\text{ED}_n^k(\mathbf{Z}) > Q_{\hat{R}, 1-\alpha} \right) = P(V(\Gamma_0) > Q_{\hat{T}, 1-\alpha}),$$

where

$$\hat{T}(t) := \frac{1}{(n+m)!} \sum_{i=1}^{(n+m)!} \mathbb{I}_{\{V(\Gamma_i) \leq t\}}$$

and

$$V(\Gamma_s) = \sum_{i=1}^n \sum_{j=1}^m \Pi_{s,ij} b_{ij} - \sum_{1 \leq i < j \leq n} \Pi_{s,ij} c_{ij} - \sum_{1 \leq i < j \leq m} \Pi_{s,ij} d_{ij}.$$

(ii) If the critical value is chosen as $Q_{\tilde{R}, 1-\alpha}$.

$$\lim_{p \rightarrow \infty} P_{H_{A_l}} \left(\text{ED}_n^k(\mathbf{Z}) > Q_{\tilde{R}, 1-\alpha} \right) = P \left(V(\Gamma_0) > Q_{\tilde{T}, 1-\alpha} \right),$$

where

$$\tilde{T} := \frac{1}{S} \left(\mathbb{I}_{\{V(\Gamma_0) \leq t\}} + \sum_{i=1}^{S-1} \mathbb{I}_{\{V(\Gamma_i) \leq t\}} \right).$$

3, [Trivial Power] Suppose Assumptions 28, 29, 32 hold. Then, we have

$$\lim_{p \rightarrow \infty} P_{H_{A_t}} \left(\text{ED}_n^k(\mathbf{Z}) > c \right) \leq \alpha \text{ where } c = Q_{\hat{R}, 1-\alpha} \text{ or } c = Q_{\tilde{R}, 1-\alpha},$$

which means that the asymptotic power of ED^k based permutation test is no more than the level α when p goes to infinity.

Remark 36. The above theorem and discussions in subsection 2.3.1 indicate that

- 1, L^1 -norm can be more advantageous than L^2 -norm, Gaussian kernel and Laplacian kernel when the dimension is high, since L^1 -distance leads to high power provided that the summation of discrepancies between marginal univariate distributions is not so small, while L^2 -norm, Gaussian kernel and Laplacian kernel would result in power loss when the total of marginal univariate mean and variance differences between X and Y is of order $o(\sqrt{p})$. Notice that the distributions of X and Y can differ in other aspects of the marginal distribution even if they have the same marginal univariate mean and variance.
- 2, All the tests under examination are only capable of detecting the discrepancies of marginal distributions. If the two high dimensional distributions $F \neq G$, but $F_u = G_u$ for $u = 1, 2, \dots, p$, then none of them have consistent power.

2.3.3 High Dimensional Medium Sample Size (HDMSS)

In this subsection, the theories are developed under the high dimensional medium sample size setting (HDMSS), i.e., as $p \rightarrow \infty$, $n := n(p) \rightarrow \infty$ at a slower rate compared to p and $n/m = \rho$, where $\rho \in (0, \infty)$ is a fixed constant. Though the proofs are quite different, most results and phenomena under the HDLSS setting have their similar counterparts under the HDMSS setting. Now that we have n, m growing to infinity, we need a stronger version of Assumptions 31 and 32.

Assumption 37. $nm\alpha_{xy}^2 = o(1)$, $n^2\alpha_x^2 = o(1)$ and $m^2\alpha_y^2 = o(1)$.

Assumption 38. $\sqrt{nm\rho}\alpha_{xy}^2 = o(1)$, $n\sqrt{p}\alpha_x^2 = o(1)$ and $m\sqrt{p}\alpha_y^2 = o(1)$.

Remark 39. Following Remark 33, for κ -dependent stationary time series, $\alpha_{xy}^2 = O(\kappa/p)$. Thus, Assumptions 37 and 38 both require that $nm\kappa = o(p)$.

To derive the asymptotic distribution under the HDMSS, we note that the leading term of $\text{ED}_n^k(\mathbf{Z})$ is a martingale, the following assumption is used to ensure the conditional Lindeberg condition and the requirements on the conditional variance in classic martingale central limit theorem.

Assumption 40. For any $\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4 \in \{X, Y\}$, suppose

$$\begin{aligned} E \left[\mathcal{K}^4(\Lambda_1, \Lambda_2') \right] &= o(n^2), \\ E \left[\mathcal{K}^2(\Lambda_1, \Lambda_3'') \mathcal{K}^2(\Lambda_2', \Lambda_3'') \right] &= o(n), \\ E \left[\mathcal{K}(\Lambda_1, \Lambda_3'') \mathcal{K}(\Lambda_1, \Lambda_4''') \mathcal{K}(\Lambda_2', \Lambda_4''') \mathcal{K}(\Lambda_2', \Lambda_3'') \right] &= o(1), \end{aligned}$$

where $(\Lambda_1', \Lambda_2', \Lambda_3', \Lambda_4')$, $(\Lambda_1'', \Lambda_2'', \Lambda_3'', \Lambda_4'')$ and $(\Lambda_1''', \Lambda_2''', \Lambda_3''', \Lambda_4''')$ are independent copies of $(\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4)$.

Remark 41. For any function φ and ψ , suppose X and Y are κ dependent sequences, i.e., $x_u \perp x_v$ and $y_u \perp y_v$ if $|u - v| > \kappa$. If there exists some constant $C > 0$ such that

$$\max \left\{ \sup_u E [\psi^4(x_u, y_u)], \sup_u E [\psi^4(x_u, x'_u)], \sup_u E [\psi^4(y_u, y'_u)] \right\} \leq C.$$

Then, for notational convenience, let

$$\phi_{xy,u} = \psi(x_u, y_u) - E[\psi(x_u, y_{ju})|x_u] - E[\psi(x_u, y_u)|y_u] + E[\psi(x_u, y_u)],$$

we see that $\sup_u E[\phi_{xy,u}^4] \leq 4^4 C$ and thus $E[\mathcal{K}^4(X, Y)]$ can be bounded as following

$$E[\mathcal{K}^4(X, Y)] = \frac{1}{p^2} \sum_{s=1}^p \sum_{t,u,v=s-3\kappa}^{s+3\kappa} E[\phi_{xy,s} \phi_{xy,t} \phi_{xy,u} \phi_{xy,v}] = O\left(\frac{\kappa^3}{p}\right).$$

and similar results can be shown for $E[\mathcal{K}^4(X, X')]$ and $E[\mathcal{K}^4(Y, Y')]$. Thus, Assumption 40 is satisfied if $\kappa^3/p = o(1)$.

Let Φ denote the cdf of $N(0, 1)$. We shall show that $\text{ED}^k(\Gamma_w \mathbf{Z})$ converges uniformly with respect to w under the HDMSS setting.

Theorem 9. For $w = 0, 1, 2, \dots, \min\{n, m\}$, fix $\Gamma_w \in \mathbb{S}_w$,

(i) Under Assumptions 28 and 37,

$$\sup_w \left| \text{ED}_n^k(\Gamma_w \mathbf{Z}) - \mu_{n,w} \right| = o_p(1).$$

where $\mu_{n,w}$ is the same as that in Theorem 7.

(ii) Under Assumptions 28, 37, 38, 40 and local alternative H_{A_l} ,

$$\sup_w \left| P\left(\sqrt{nmp} \left(\text{ED}_n^k(\Gamma_w \mathbf{Z}) - \mu_{n,w} \right) \leq a\right) - \Phi\left(\frac{a}{\sqrt{nm\sigma_{n,w}^2}}\right) \right| = o(1)$$

where a is a fixed constant and $\sigma_{n,w}^2$ is the same as in Theorem 7.

Then, the following theorem states the asymptotic distribution of $\text{ED}_n^k(\mathbf{\Gamma} \mathbf{Z})$.

Theorem 10. For $\mathbf{\Gamma} \sim \text{Uniform}(\mathbb{P}_{n+m})$, which is independent of the data,

(i) Under Assumptions 28 and 37,

$$\text{ED}_n^k(\mathbf{\Gamma} \mathbf{Z}) \xrightarrow{P} 0.$$

(ii) Under Assumptions 28, 37, 38, 40 and local alternative H_{A_l} ,

$$\sqrt{nmp} \begin{pmatrix} \text{ED}_n^k(\mathbf{\Gamma} \mathbf{Z}) \\ \text{ED}_n^k(\mathbf{\Gamma}' \mathbf{Z}) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$$

where $\mathbf{\Gamma}'$ is an independent copy of $\mathbf{\Gamma}$ and σ^2 is the asymptotic variance defined as

$$\sigma^2 := 4v_{xy}[\varphi^{(1)}(e_{xy})]^2 + 2\rho v_x[\varphi^{(1)}(e_x)]^2 + \frac{2}{\rho}v_y[\varphi^{(1)}(e_y)]^2.$$

We need the limiting distribution of $(\text{ED}_n^k(\mathbf{\Gamma}\mathbf{Z}), \text{ED}_n^k(\mathbf{\Gamma}'\mathbf{Z}))$ to show that the variance of randomization distribution go to 0, from which it follows that the randomization distribution converges in probability to the limit of its mean. Furthermore, we can show that the critical values are concentrating on some constants.

Corollary 3. *Let $\mathbf{\Gamma}_1, \dots, \mathbf{\Gamma}_S$ be i.i.d and uniformly sampled from \mathbb{P}_{n+m} .*

(i) *Under Assumptions 28 and 37, as $n \wedge m \wedge p \wedge S \rightarrow \infty$,*

$$\widehat{R}(t) \xrightarrow{P} \mathbb{I}_{\{t \geq 0\}} \text{ and } \widetilde{R}(t) \xrightarrow{P} \mathbb{I}_{\{t \geq 0\}}.$$

Consequently, we have $Q_{\widehat{R}, 1-\alpha} \xrightarrow{P} 0$ and $Q_{\widetilde{R}, 1-\alpha} \xrightarrow{P} 0$.

(ii) *Under Assumptions 28, 37, 38, 40 and local alternative H_{A_l} , as $n \wedge m \wedge p \wedge S \rightarrow \infty$,*

$$\begin{aligned} \frac{1}{(n+m)!} \sum_{i=1}^{(n+m)!} \mathbb{I}_{\{\sqrt{nm}p \text{ED}_n^k(\mathbf{\Gamma}_i \mathbf{Z}) \leq t\}} &\xrightarrow{P} \Phi(t/\sigma), \\ \frac{1}{S} \sum_{i=1}^S \mathbb{I}_{\{\sqrt{nm}p \text{ED}_n^k(\mathbf{\Gamma}_i \mathbf{Z}) \leq t\}} &\xrightarrow{P} \Phi(t/\sigma) \end{aligned}$$

Consequently, we have $\sqrt{nm}p Q_{\widehat{R}, 1-\alpha} \xrightarrow{P} \sigma Q_{\Phi, 1-\alpha}$ and $\sqrt{nm}p Q_{\widetilde{R}, 1-\alpha} \xrightarrow{P} \sigma Q_{\Phi, 1-\alpha}$, where σ^2 is defined in Theorem 10.

The power behavior of $\text{ED}_n^k(\mathbf{Z})$ w.r.t permutation test under the HDMSS is stated in the following theorem.

Theorem 11. *Assume that $2\varphi(e_{xy}) \geq \varphi(e_x) + \varphi(e_y)$. For any $c \in \{Q_{\widehat{R}, 1-\alpha}, Q_{\widetilde{R}, 1-\alpha}\}$, the following holds.*

1, **[Consistency]** *Under Assumptions 28, 37. Then, we have*

$$\lim_{p \rightarrow \infty} P_{H_{A_c}} \left(\text{ED}_n^k(\mathbf{Z}) > c \right) = 1,$$

which means that the asymptotic power of ED^k based permutation test is 1 as $p \wedge n \wedge m \rightarrow \infty$.

2, **[Trivial Power]** *Under Assumptions 28, 37, 38, 40. Then, we have*

$$\lim_{p \rightarrow \infty} P_{H_{A_l}} \left(\text{ED}_n^k(\mathbf{Z}) > c \right) \leq \alpha,$$

Thus, we have the asymptotic power of ED^k based permutation test is no more than the level α when $p \wedge n \wedge m \rightarrow \infty$.

Comparing with Theorem 8, the $\text{ED}_n^k(\mathbf{Z})$ based permutation test have trivial power under H_{A_l} and the HDMSS setting. This is due to the interesting facts that $nm\sigma_{n,W}^2$ converges in probability to σ^2 , which is also the limit of $nm\sigma_{n,0}^2$ as $n \rightarrow \infty$ and

$$\text{cov} \left(\text{ED}_n^k(\mathbf{\Gamma}\mathbf{Z}), \text{ED}_n^k(\mathbf{\Gamma}'\mathbf{Z}) \right) \rightarrow 0 \text{ as } n \rightarrow \infty,$$

which ensures that the randomization distribution converges in probability to its mean limit.

2.4 Numerical Studies

In this section, we consider several examples to demonstrate the finite sample performance of ED^k based permutation test for different distance metrics. In our numerical comparison, we include the tests of Li (2018) (denoted as JL) and Biswas and Ghosh (2014) (denoted as BG) as these two were shown to have higher power over others in Li (2018). The critical values of JL test are determined by its asymptotic distribution, whereas BG test is also implemented as a permutation test.

2.4.1 Performance on simulated data

In all our simulations, we set $\alpha = 0.05$ and perform 1000 Monte Carlo replications with 300 permutations for each test. The first example is adopted from the simulation setting of Li (2018) to study the size accuracy.

Example 7. *Generate samples as*

$$X = (V^{1/2}RV^{1/2})^{1/2}Z_1,$$

$$Y = (V^{1/2}RV^{1/2})^{1/2}Z_2,$$

where $R = (r_{ij})_{i,j=1}^p$, $r_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$ or 0.8 ; V is a diagonal matrix with $V_{ii}^{1/2} = 1$ or uniformly drawn from $(1,5)$. Z_1, Z_2 are i.i.d copies of Z with

$$Z = \underbrace{(z_1, z_2, \dots, z_p)}_{\sim N(0,1)} \text{ or } Z = \underbrace{(z_1, z_2, \dots, z_p)}_{\sim \text{Exponential}(1)} - \mathbf{1}_p.$$

In Example 7, X and Y follow the same distribution and we consider cases that $n = m = 50$ or $n = 70, m = 30$. From Table 2.5, we can see that all the tests have quite accurate size. To compare the power,

Table 2.5: Size comparison from Example 7 for $p = 500$

	ρ	$V_{ii}^{1/2}$	n	m	$ED^{L^2\text{-norm}}$	ED^{Gaussian}	$ED^{\text{Laplacian}}$	$ED^{L^1\text{-norm}}$	BG	JL
Normal	0.5	1	50	50	0.06	0.06	0.058	0.059	0.053	0.053
	0.5	1	70	30	0.07	0.07	0.068	0.073	0.047	0.057
	0.5	Un(1,5)	50	50	0.052	0.052	0.05	0.051	0.056	0.057
	0.5	Un(1,5)	70	30	0.059	0.059	0.061	0.05	0.049	0.045
	0.8	1	50	50	0.053	0.053	0.052	0.059	0.054	0.055
	0.8	1	70	30	0.045	0.046	0.046	0.05	0.052	0.055
	0.8	Un(1,5)	50	50	0.045	0.045	0.049	0.048	0.054	0.054
	0.8	Un(1,5)	70	30	0.05	0.05	0.049	0.046	0.051	0.051
Exponential	0.5	1	50	50	0.06	0.06	0.058	0.059	0.053	0.053
	0.5	1	70	30	0.063	0.063	0.063	0.058	0.048	0.053
	0.5	Un(1,5)	50	50	0.057	0.057	0.058	0.055	0.049	0.06
	0.5	Un(1,5)	70	30	0.056	0.056	0.06	0.058	0.059	0.058
	0.8	1	50	50	0.054	0.054	0.051	0.047	0.065	0.062
	0.8	1	70	30	0.061	0.061	0.062	0.065	0.057	0.06
	0.8	Un(1,5)	50	50	0.051	0.05	0.052	0.046	0.045	0.057
	0.8	Un(1,5)	70	30	0.062	0.062	0.062	0.062	0.06	0.064

we first use an example from Li (2018), which include the situation when X and Y only differ in their means or only differ in their covariance matrices or differ in both, where $\beta \in [0, 1]$ is the percentage of the p components that differ in their distributions.

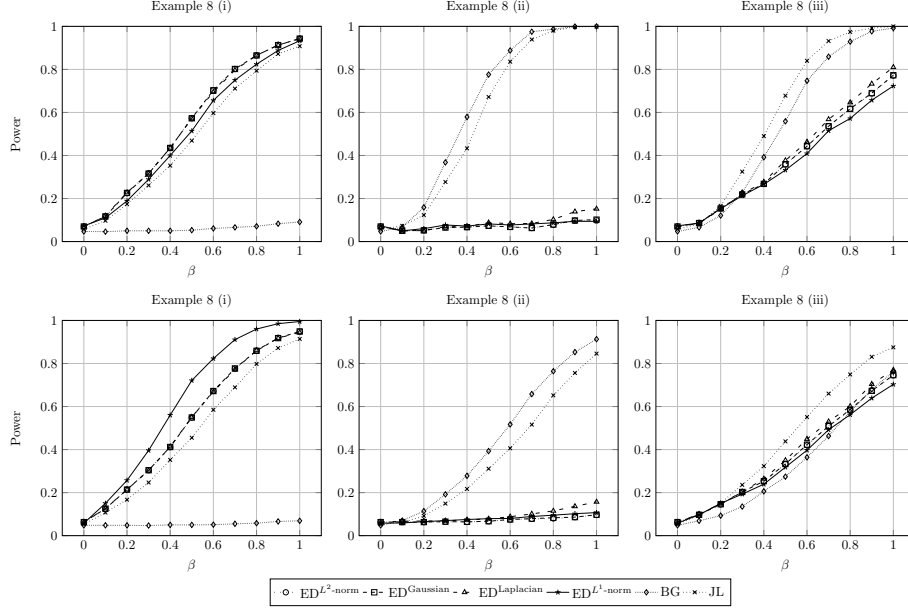


Figure 2.2: Power comparison for example 8 and $n = 70$, $m = 30$, $p = 500$, where in the top 3 figures Z_1, Z_2 are generated from normal distribution and in the bottom 3 figures, Z_1, Z_2 are generated from exponential distribution.

Example 8. Let R, V, Z_1, Z_2 be defined the same as in Example 7 and we choose $\rho = 0.5$ here. Generate samples as

(i)

$$X = (V^{1/2}RV^{1/2})^{1/2}Z_1,$$

$$Y = (0.125 \times \mathbf{1}_{\beta p}, \mathbf{0}_{(1-\beta)p}) + (V^{1/2}RV^{1/2})^{1/2}Z_2.$$

(ii) Let V^* be a diagonal matrix with $V_{ii}^{*1/2} = 1.05$ for $i = 1, 2, \dots, \beta p$ and $V_{ii}^{*1/2} = 1$ for $i = \beta p + 1, \dots, \beta p$.

$$X = (V^{1/2}RV^{1/2})^{1/2}Z_1,$$

$$Y = (V^{*1/2}RV^{*1/2})^{1/2}Z_2.$$

(iii) Let $V_{ii}^{*1/2} = 1.04$ for $i = 1, 2, \dots, \beta p$ and $V_{ii}^{*1/2} = 1$ for $i = \beta p + 1, \dots, \beta p$.

$$X = (V^{1/2}RV^{1/2})^{1/2}Z_1,$$

$$Y = (0.1 \times \mathbf{1}_{\beta p}, \mathbf{0}_{(1-\beta)p}) + (V^{*1/2}RV^{*1/2})^{1/2}Z_2.$$

From Figure 2.2, we can see that (1) when there is a small difference in the means, ED^k-based tests and JL perform similarly, while BG barely show any power. (2) when there is a small difference in the scales, JL and BG are consistent and ED^k-based tests have very little power. Similar phenomenon by Li (2018) were also observed, i.e., ED^k based permutation test is not sensitive to small scale differences and the method proposed by Li (2018) and Biswas and Ghosh (2014) have dominant power in this case. Note that there is a tuning parameter involved in JL test and its choice could have a big impact on the size and power; results

not shown. (3) when there are differences for both the means and scales, all the tests performs comparably.

Next, Example 9 examines the situation when X and Y have the same marginal univariate mean and variance, but different marginal univariate distributions.

Example 9. *Generate samples as*

- (i) Let $\text{Rademacher}(0.5)$ be the Rademacher distribution with success probability 0.5, e.g. $P(y_{iu} = -1) = P(y_{iu} = 1) = 0.5$.

$$X = (x_1, \dots, x_p) \stackrel{iid}{\sim} N(0, 1),$$

$$Y = (\underbrace{y_1, y_2, \dots, y_{\beta p}}_{\stackrel{iid}{\sim} \text{Rademacher}(0.5)}, \underbrace{y_{\beta p+1}, y_{\beta p+2}, \dots, y_p}_{\stackrel{iid}{\sim} N(0, 1)}).$$

(ii)

$$X = (x_1, \dots, x_p) \stackrel{iid}{\sim} N(0, 1),$$

$$Y = (\underbrace{y_1, y_2, \dots, y_{\beta p}}_{\stackrel{iid}{\sim} \text{Uniform}(-\sqrt{3}, \sqrt{3})}, \underbrace{y_{\beta p+1}, y_{\beta p+2}, \dots, y_p}_{\stackrel{iid}{\sim} N(0, 1)}).$$

From Figure 2.3, we see that only $\text{ED}^{L^1\text{-norm}}$ based permutation test has power growing as β elevates (p fixed) or p increases (β fixed). This phenomenon matches with our theories, which indicate that L^2 -norm, Gaussian and Laplacian kernel can detect only marginal mean and variance differences. For $\text{ED}^{L^1\text{-norm}}$ based permutation test, the power is growing more rapidly for Example 9 (i) than Example 9 (ii), which might suggest that L^1 -distance is more sensitive for the difference between continuous and discrete distributions. It is also apparent that the JL and BG tests show little power in this example. The next example examines the case where X and Y have the same marginal univariate distributions.

Example 10. *Generate samples as*

- (i) Let $(y'_1, y'_2, \dots, y'_{\beta p/2}) \stackrel{iid}{\sim} \text{Bernoulli}(0.5)$

$$X = (x_1, \dots, x_p) \stackrel{iid}{\sim} \text{Bernoulli}(0.5),$$

$$Y = (y'_1, \mathbb{I}_{\{y'_1=1\}}, y'_2, \mathbb{I}_{\{y'_2=1\}}, \dots, y'_{\beta p/2}, \mathbb{I}_{\{y'_{\beta p/2}=1\}}, \underbrace{y_1, y_2, \dots, y_{(1-\beta)p}}_{\stackrel{iid}{\sim} \text{Bernoulli}(0.5)}).$$

- (ii) Let $(y'_1, y'_2, \dots, y'_{\beta p/3}) \stackrel{iid}{\sim} \text{Bernoulli}(0.5)$ and $(y''_1, y''_2, \dots, y''_{\beta p/3}) \stackrel{iid}{\sim} \text{Bernoulli}(0.5)$

$$X = (x_1, \dots, x_p) \stackrel{iid}{\sim} \text{Bernoulli}(0.5),$$

$$Y = (y'_1, y''_1, \mathbb{I}_{\{y'_1=y''_1\}}, \dots, y'_{\beta p/3}, y''_{\beta p/3}, \mathbb{I}_{\{y'_{\beta p/3}=y''_{\beta p/3}\}}, \underbrace{y_1, y_2, \dots, y_{(1-\beta)p}}_{\stackrel{iid}{\sim} \text{Bernoulli}(0.5)}).$$

Notice that in Example 10 (i) X, Y have the same marginal univariate distribution, but different marginal bivariate distributions and in Example 10 (ii) X, Y have the same marginal bivariate distribution, but

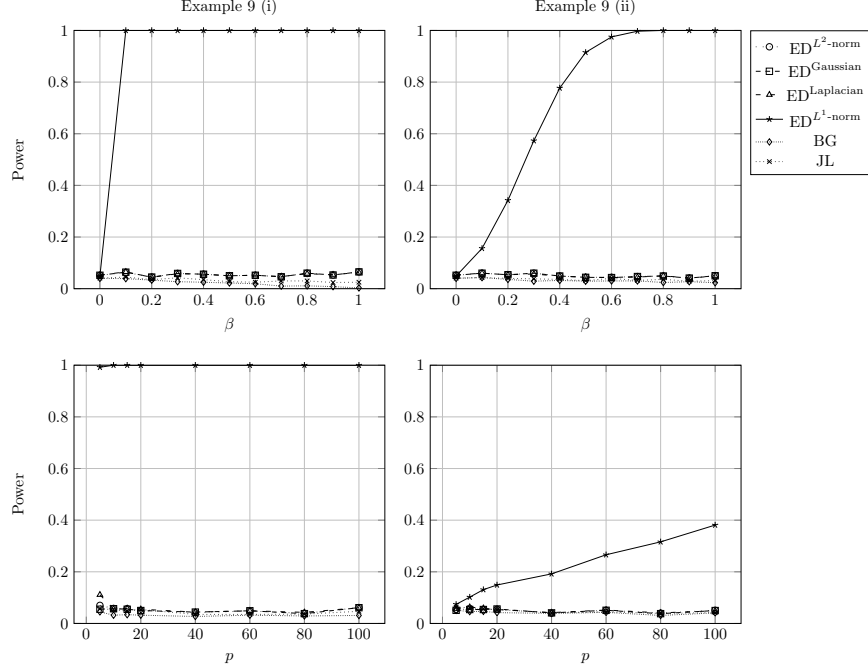


Figure 2.3: Power comparison for Example 9 and $n = 70$, $m = 30$. For the top two figures, the dimension p is equal to 500 and we plot the power as β ranges from 0 to 1. For the bottom two figures, β is fixed to be 0.5 and the power is plotted with respect to p .

different joint distribution. Theorem 8 (ii) and Theorem 11 (ii) both provide insights that L^2 -norm, L^1 -norm, Gaussian or Laplacian kernel based tests all suffer substantial power loss under Example 10 (i). On the other hand, Theorem 8 (iii) suggests us that since Example 10 (ii) belong to class H_{A_t} , all these tests have trivial power. The simulation results of Example 10 are in Figure 2.4 and they again corroborate our theoretical findings.

2.4.2 Performance on real data

We also compare the power of the above tests on the following real data sets.

- Strawberry data: this data set contains the spectrographs of fruit purees. There are totally two classes: one is strawberry purees (authentic samples) and the other one is non-strawberry purees (adulterated strawberries and other fruits). Each data point is of length 235.
- SmallKitchenAppliances data: this data sets contains records of the electricity usage of some kitchen appliances. We only use classes Kettle and Microwave. Each data point has readings taken every 2 minutes over 24 hours.
- Earthquakes data: this data set is from Northern California Earthquake Data Center and has classes of positive and negative major earthquake events. There are 368 negative and 93 positive cases and each data point is of length 512.

All the above data sets are downloaded from UCR Time Series Classification Archive Dau et al. (2018) (https://www.cs.ucr.edu/~eamonn/time_series_data_2018/) and a glance of these data sets is provided

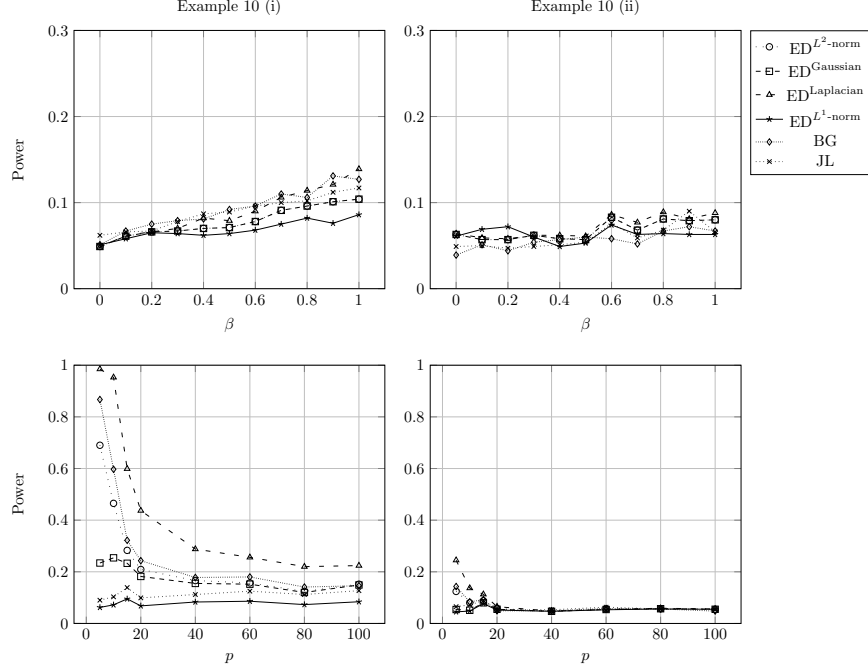


Figure 2.4: Power comparison for Example 10 and $n = 70$, $m = 30$. For the top two figures, the dimension p is equal to 500 and we plot the power as β ranges from 0 to 1. For the bottom two figures, β is fixed to be 1 and the power is plotted with respect to p .

in Figure 2.5. For each of the three data sets, the data points have two classes and we want to compare the underlining distributions of the two classes. Following the procedures of Biswas and Ghosh (2014) and Sarkar et al. (2018), for each $m = n \in \{10, 20, 30, 40, 50, 60\}$, we randomly sample n points from each class and test whether the two distributions are the same using the afore-mentioned tests. The same procedure is repeated 1000 times to calculate the power.

The experimental results for these data sets are shown in Figure 2.6, from which we see that all the tests have very high power for the Strawberry data with relatively low sample size. As for the SmallKitchenAppliances and Earthquakes data sets, the L^1 -norm based test demonstrates superior power compared to other tests. It is also worth noting that BG and JL barely exhibit any power for the Earthquakes data.

2.5 Discussions & Conclusion

In this paper, we study the two-sample hypothesis testing problem in a high dimension and low/medium sample size setting. Our focus is on the interpoint distance based permutation tests, such as those based on Energy Distance (ED) and Maximum Mean Discrepancy (MMD). Our theory demonstrates that all these tests under examination are unable to detect the difference between two high dimensional distributions beyond univariate marginal distributions. In particular, the ED test with L^2 -norm and MMD with Gaussian or Laplacian kernels suffer substantial power loss under the HDLSS and have trivial power under the HDMSS when the average of component-wise mean and variance discrepancies between two distributions are both asymptotically zero at the rate of $o(1/\sqrt{nm})$. Thus these tests mainly target mean and variance differences in marginal distributions. By contrast, if we use L^1 -norm in ED test, then the non-negligible difference in marginal univariate distributions, as quantified by cumulative energy distance of marginal distributions, can

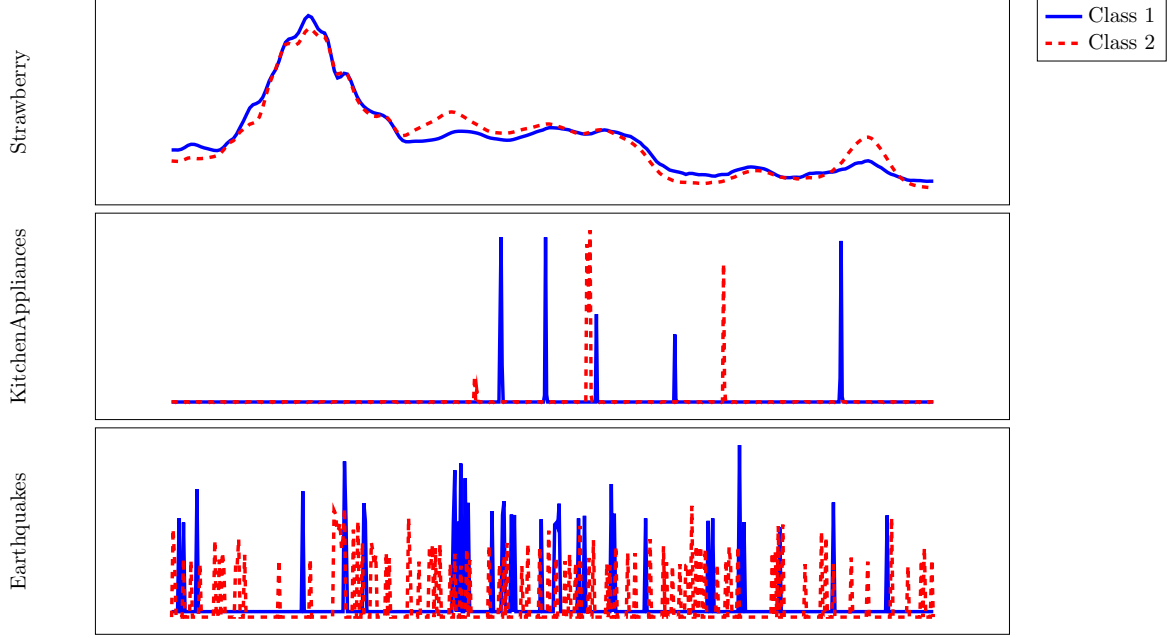


Figure 2.5: A glance of the data in Section 2.4.2, where we plot one point from each of the two classes for each data set.

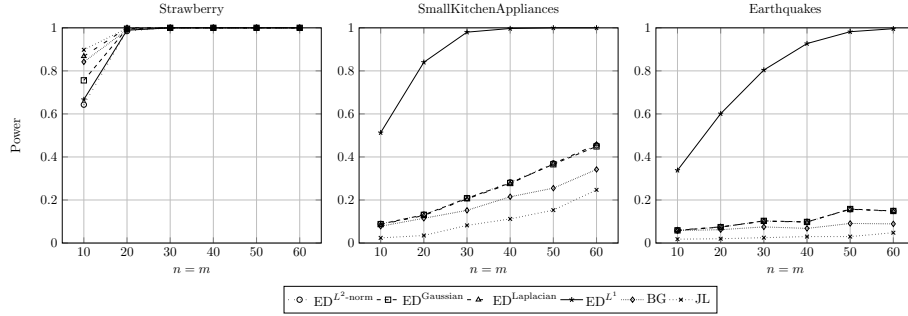


Figure 2.6: Power comparison for real data examples in Section 2.4.2.

be detected with high power. Thus the theory suggests that

1), The ED with L^2 -norm, and MMD with Gaussian and Laplacian kernels are of the same category, as they all depend on the interpoint distance as measured by Euclidean distance, which leads to undesirable power limitation.

2), Although in a low dimensional setting the use of L^1 -norm in ED is not preferred due to the fact that it does not completely characterize the difference between two distributions since $ED^1(F, G) = 0$ does not necessarily imply $F = G$, it seems to have some advantage over the ED with L^2 -norm and MMD with Gaussian and Laplacian kernels in the high dimensional setting, as shown in both theory and numerical studies.

3), As shown in our simulations and data illustration, the existing interpoint distance test by Li (2018) and Biswas and Ghosh (2014) also suffer from low power when the two distributions have the same marginal mean and variances but different marginal distributions. So in this sense, they are also inferior to the ED test with L^1 -norm.

4), The difference in marginal distributions of two high dimensional distributions can be interpreted as the main effect of the distribution differences. It is a standard statistical practice to test for the nullity of main effects first, before proceeding to the higher-order interactions. Thus we advocate the use of L^1 -norm based test to test for the presence of main differences in two high dimensional distributions.

To conclude the paper, we shall mention a few future directions. First, we are holding the bandwidth parameter in Gaussian and Laplacian kernels fixed for theoretical convenience, and it would be interesting to relax this restriction by allowing it to be data-dependent. Second, there might be some intrinsic difficulty of capturing all kinds of differences in two high dimensional distributions with limited sample sizes, so it seems natural to ask whether it is possible to detect any difference beyond marginal univariate distributions. If possible, what would be the form of the new tests? We leave these topics for future investigation.

2.6 Technical Details

2.6.1 Proof of Sufficient Conditions for Local Alternatives

When $\psi(x, y) = (x - y)^2$, φ is strictly concave, strictly increasing on $(0, +\infty)$ (e.g. L^2 -norm, Gaussian kernel multiplied by -1 and Laplacian kernel multiplied by -1), we first note that $2e_{xy} - e_x - e_y = 2 \lim_{p \rightarrow \infty} \sum_{u=1}^p (E(x_u) - E(y_u))^2 / p \geq 0$ and

$$\varphi(e_{xy}) - \frac{\varphi(e_x) + \varphi(e_y)}{2} \geq \varphi(e_{xy}) - \varphi\left(\frac{e_x + e_y}{2}\right) \geq 0,$$

where the equality holds iff $e_{xy} = e_x = e_y$. Also, some algebra shows that

$$\begin{aligned} e_{xy} &= e_x + \lim_{p \rightarrow \infty} \frac{1}{p} \sum_{u=1}^p (E(x_u) - E(y_u))^2 + \lim_{p \rightarrow \infty} \frac{1}{p} \sum_{u=1}^p (\text{var}(y_u) - \text{var}(x_u)) \\ &= e_y + \lim_{p \rightarrow \infty} \frac{1}{p} \sum_{u=1}^p (E(x_u) - E(y_u))^2 + \lim_{p \rightarrow \infty} \frac{1}{p} \sum_{u=1}^p (\text{var}(x_u) - \text{var}(y_u)). \end{aligned}$$

Thus, in summary we have

$$\begin{aligned} 2\varphi(e_{xy}) &= \varphi(e_x) + \varphi(e_y) \Leftrightarrow e_{xy} = e_x = e_y \\ &\Leftrightarrow \sum_{u=1}^p (E(x_u) - E(y_u))^2 = o(p) \text{ and } \left| \sum_{u=1}^p (\text{var}(x_u) - \text{var}(y_u)) \right| = o(p). \end{aligned}$$

This proves the result for H_{A_c} characterization. Next, for sufficient conditions of H_{A_t} , if we have

$$\sum_{u=1}^p (E(x_u) - E(y_u))^2 = o\left(\frac{\sqrt{p}}{\sqrt{nm}}\right) \text{ and } \left| \sum_{u=1}^p (\text{var}(x_u) - \text{var}(y_u)) \right| = o\left(\frac{\sqrt{p}}{\sqrt{nm}}\right),$$

then it holds that $e_{xy} = e_x = e_y$ and

$$\begin{aligned} & E \left[|E[\bar{\psi}(X, Y)|X] - E[\bar{\psi}(X, X')|X]| \right] \\ & \leq \frac{2}{p} \sqrt{\sum_{u=1}^p E(x_u^2) \sum_{u=1}^p (E(x_u) - E(y_u))^2} + \frac{1}{p} \left| \sum_{u=1}^p (\text{var}(y_u) - \text{var}(x_u)) \right| \\ & \quad + \frac{1}{p} \sqrt{\sum_{u=1}^p (E(x_u) + E(y_u))^2 \sum_{u=1}^p (E(x_u) - E(y_u))^2} = o\left(\frac{1}{\sqrt{nm p}}\right). \end{aligned}$$

For H_{A_t} , a straight forward calculation shows that

$$\psi(x, y) - E[\psi(x, y)|x] - E[\psi(x, y)|y] + E[\psi(x, y)] = -2(x - E(x))(y - E(y))$$

and $v_{xy} = \sum_{u,v=1}^p 4\text{cov}(x_u, x_v)\text{cov}(y_u, y_v)/p$. Thus, from CauchySchwarz inequality, we have

$$\sum_{u,v=1}^p (\text{cov}(x_u, x_v) - \text{cov}(y_u, y_v))^2 = o(p) \Rightarrow v_{xy} = v_x = v_y.$$

When $\psi(x, y) = |x - y|$, $\varphi(x) = x$, the results follow from the following equality.

$$\begin{aligned} & 2\varphi(e_{xy}) - \varphi(e_x) - \varphi(e_y) \\ & = 2 \lim_{p \rightarrow \infty} \frac{1}{p} \sum_{u=1}^p E[|x_{1u} - y_{1u}|] - \lim_{p \rightarrow \infty} \frac{1}{p} \sum_{u=1}^p E[|x_{1u} - x_{2u}|] - \lim_{p \rightarrow \infty} \frac{1}{p} \sum_{u=1}^p E[|y_{1u} - y_{2u}|] \\ & = \lim_{p \rightarrow \infty} \frac{1}{p} \sum_{u=1}^p (2E[|x_{1u} - y_{1u}|] - E[|x_{1u} - x_{2u}|] - E[|y_{1u} - y_{2u}|]) \\ & = \lim_{p \rightarrow \infty} \frac{1}{p} \sum_{u=1}^p \text{ED}(F_u, G_u). \end{aligned}$$

2.6.2 Proof of Theorem 7

Proof. (i) Taking a first order Taylor expansion w.r.t φ gives

$$k(Z_i, Z_j) = \varphi(e_{ij}) + \mathcal{R}_1(Z_i, Z_j),$$

where $\mathcal{R}_1(Z_i, Z_j)$ is an operator that acts on random variables

$$\mathcal{R}_1(Z_i, Z_j) = \mathcal{L}(Z_i, Z_j) \int_0^1 \varphi^{(1)}(e_{ij} + v\mathcal{L}(Z_i, Z_j)) dv$$

For each fixed permutation matrix $\Gamma_w \in \mathbb{S}_w$

$$\text{ED}_n^k(\Gamma_w \mathbf{Z}) = \underbrace{\sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{w,ij} \varphi(e_{ij})}_{:=\mu_n(\Gamma_w \mathbf{Z})} + \underbrace{\sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{w,ij} \mathcal{R}_1(Z_i, Z_j)}_{:=R_1(\Gamma_w \mathbf{Z})}, \quad (2.7)$$

where $\mu_n(\Gamma_w \mathbf{Z})$ is the asymptotic mean for the permuted data and equals

$$\begin{aligned} \frac{2}{mn} & \left((w^2 + (n-w)(m-w)) \varphi(e_{xy}) + (n-w)w\varphi(e_x) + (m-w)w\varphi(e_y) \right) \\ & - \frac{1}{n(n-1)} (2w(n-w)\varphi(e_{xy}) + (n-w)(n-w-1)\varphi(e_x) + w(w-1)\varphi(e_y)) \\ & - \frac{1}{m(m-1)} (2w(m-w)\varphi(e_{xy}) + w(w-1)\varphi(e_x) + (m-w)(m-w-1)\varphi(e_y)). \end{aligned}$$

Then, after re-arranging the terms according to the powers of w , we have

$$\mu_n(\Gamma_w \mathbf{Z}) = (2\varphi(e_{xy}) - \varphi(e_x) - \varphi(e_y)) f(w),$$

where $f(w)$ is a second order polynomial with respect to w

$$f(w) = 1 - \left(\frac{2m-1}{m(m-1)} + \frac{2n-1}{n(n-1)} \right) w + \left(\frac{2}{mn} + \frac{1}{n(n-1)} + \frac{1}{m(m-1)} \right) w^2.$$

For the remainder term $R_1(\Gamma_w \mathbf{Z})$, notice that $\mathcal{L}(Z_i, Z_j) \xrightarrow{p} 0$ for any $1 \leq i < j < n+m$. By the continuous mapping theorem, we know

$$\int_0^1 \varphi^{(1)}(e_{ij} + v\mathcal{L}(Z_i, Z_j)) dv \xrightarrow{p} \int_0^1 \varphi^{(1)}(e_{ij}) dv.$$

Thus, it holds that $\mathcal{R}_1(Z_i, Z_j) \asymp_p \mathcal{L}(Z_i, Z_j)$ and $R_1(\Gamma_w \mathbf{Z}) = O_p(\alpha_{xy} + \alpha_x + \alpha_y) = o_p(1)$.

(ii) Taking a second order Taylor expansion w.r.t φ gives

$$k(Z_i, Z_j) = \varphi(e_{ij}) + \varphi^{(1)}(e_{ij}) \frac{\mathcal{K}(Z_i, Z_j) + \mathcal{W}(Z_i, Z_j)}{\sqrt{p}} + \mathcal{R}_2(Z_i, Z_j),$$

where \mathcal{R}_2 and \mathcal{W} are defined as

$$\begin{aligned} \mathcal{R}_2(Z_i, Z_j) &= \mathcal{L}^2(Z_i, Z_j) \int_0^1 \int_0^1 u \varphi^{(2)}(e_{ij} + uv\mathcal{L}(Z_i, Z_j)) dv du, \\ \mathcal{W}(Z_i, Z_j) &= \frac{1}{\sqrt{p}} \sum_{u=1}^p (E[\psi(z_{iu}, z_{ju}) | z_{iu}] + E[\psi(z_{iu}, z_{ju}) | z_{ju}] - E[\psi(z_{iu}, z_{ju})] - e_{ij}). \end{aligned}$$

Accordingly, we can decompose the sample energy distance as

$$\begin{aligned} \sqrt{p} \left(\text{ED}_n^k(\Gamma_w \mathbf{Z}) - \mu_n(\Gamma_w \mathbf{Z}) \right) &= \underbrace{\sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{w,ij} \varphi^{(1)}(e_{ij}) \mathcal{K}(Z_i, Z_j)}_{:=L(\Gamma_w \mathbf{Z})} \\ &+ \underbrace{\sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{w,ij} \varphi^{(1)}(e_{ij}) \mathcal{W}(Z_i, Z_j)}_{:=R_1(\Gamma_w \mathbf{Z})} + \underbrace{\sqrt{p} \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{w,ij} \mathcal{R}_2(Z_i, Z_j)}_{:=R_2(\Gamma_w \mathbf{Z})}. \quad (2.8) \end{aligned}$$

For the leading term $L(\Gamma_w \mathbf{Z})$, notice that under Assumption 29, $(\mathcal{K}(Z_i, Z_j))_{i < j}$ converges jointly to a multivariate normal with mean 0 and a diagonal covariance matrix. Thus, given a permutation matrix Γ_w , we

are able to obtain $L(\Gamma_w \mathbf{Z}) \xrightarrow{d} N(0, \sigma_n^2(\Gamma_w \mathbf{Z}))$, where

$$\begin{aligned} \sigma_n^2(\Gamma_w \mathbf{Z}) = & \frac{4}{n^2(n-1)^2} \left\{ \frac{(n-w)(n-w-1)}{2} v_x[\varphi^{(1)}(e_x)]^2 \right. \\ & \left. + \frac{w(w-1)}{2} v_y[\varphi^{(1)}(e_y)]^2 + (n-w)wv_{xy}[\varphi^{(1)}(e_{xy})]^2 \right\} \\ & + \frac{4}{m^2(m-1)^2} \left\{ \frac{w(w-1)}{2} v_x[\varphi^{(1)}(e_x)]^2 \right. \\ & \left. + \frac{(m-w)(m-w-1)}{2} v_y[\varphi^{(1)}(e_y)]^2 + w(m-w)v_{xy}[\varphi^{(1)}(e_{xy})]^2 \right\} \\ & + \frac{4}{n^2m^2} \left\{ (n-w)wv_x[\varphi^{(1)}(e_x)]^2 \right. \\ & \left. + w(m-w)v_y[\varphi^{(1)}(e_y)]^2 + ((n-w)(m-w) + w^2)v_{xy}[\varphi^{(1)}(e_{xy})]^2 \right\}. \end{aligned}$$

By collecting terms with respect to v_{xy}, v_x, v_y , we obtain

$$\begin{aligned} \sigma_{n,w}^2 = & \left\{ \frac{4}{nm} - 4 \left(\frac{n+m}{n^2m^2} - \frac{n}{n^2(n-1)^2} - \frac{m}{m^2(m-1)^2} \right) w \right. \\ & \left. + 4 \left(\frac{2}{n^2m^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) w^2 \right\} v_{xy}[\varphi^{(1)}(e_{xy})]^2 \\ & + \left\{ \frac{2}{n(n-1)} + 2 \left(\frac{2n}{n^2m^2} - \frac{2n-1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) w \right. \\ & \left. - 2 \left(\frac{2}{m^2n^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) w^2 \right\} v_x[\varphi^{(1)}(e_x)]^2 \\ & + \left\{ \frac{2}{m(m-1)} + 2 \left(\frac{2m}{n^2m^2} - \frac{1}{n^2(n-1)^2} - \frac{2m-1}{m^2(m-1)^2} \right) w \right. \\ & \left. - 2 \left(\frac{2}{n^2m^2} - \frac{1}{m^2(m-1)^2} - \frac{1}{n^2(n-1)^2} \right) w^2 \right\} v_y[\varphi^{(1)}(e_y)]^2. \end{aligned}$$

We then conclude the result by showing that the remainder terms are negligible. $R_l(\Gamma_w \mathbf{Z}) = o_p(1)$ is proved in lemma 44. For the $R_2(\Gamma_w \mathbf{Z})$ term, it can be shown similarly that $\mathcal{R}_2(Z_i, Z_j) \asymp_p \mathcal{L}^2(Z_i, Z_j) = O_p(\alpha_{xy}^2 + \alpha_x^2 + \alpha_y^2)$, which implies that $R_2(\Gamma_w \mathbf{Z}) = O_p(\sqrt{p}(\alpha_{xy}^2 + \alpha_x^2 + \alpha_y^2)) = o_p(1)$ under Assumption 32. \square

2.6.3 Proof of Proposition 35

Proof. (i) From Equation 2.7, we obtain

$$\text{ED}_n^k(\Gamma \mathbf{Z}) = \underbrace{\sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \mathbf{\Pi}_{ij} \varphi(e_{ij})}_{:= \mu_{n,W}} + o_p(1),$$

where $\mathbf{\Pi}_{ij}$ corresponds to Γ and $W = N(\Gamma) \sim \text{Hypergeometric}(m+n, m, n)$.

(ii) It follows from Equation 2.8, Lemma 44 and the proof of Theorem 7 that

$$\sqrt{p} \left(\text{ED}_n^k(\Gamma \mathbf{Z}) - \mu_n(\Gamma \mathbf{Z}) \right) = \underbrace{\sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{ij} \varphi^{(1)}(e_{ij}) \mathcal{K}(Z_i, Z_j)}_{:=L(\Gamma \mathbf{Z})} + o_p(1).$$

Then, under Assumption 29, it is not hard to see that $L(\Gamma \mathbf{Z}) \xrightarrow{d} N(0, \sigma_{n,W}^2)$, where $W = N(\Gamma) \sim \text{Hypergeometric}(m+n, m, n)$. This concludes the proposition. \square

2.6.4 Proof of Theorem 8

1, For any $\mathbf{a} \in \mathbb{R}^{(n+m)!}$, we define the α -th quantile of the set $\{a_1, \dots, a_{(n+m)!}\}$ as

$$Q_{1-\alpha} \{a_1, \dots, a_{(n+m)!}\} = \min \left\{ a_i : \frac{1}{(n+m)!} \sum_{i=1}^{(n+m)!} \mathbb{I}_{\{a_i \leq t\}} \geq 1 - \alpha \right\}.$$

Then, we can view $Q_{1-\alpha}$ as a continuous function on $\mathbb{R}^{(n+m)!}$.

(i) By Theorem 7, for any fixed $\Gamma_i \in \mathbb{P}_{n+m}$, we have $\text{ED}_n^k(\Gamma_i \mathbf{Z}) \xrightarrow{p} \mu_n(\Gamma_i \mathbf{Z})$. The continuous mapping theorem implies

$$Q_{1-\alpha} \left\{ \text{ED}_n^k(\Gamma_1 \mathbf{Z}), \dots, \text{ED}_n^k(\Gamma_{(n+m)!} \mathbf{Z}) \right\} \xrightarrow{p} Q_{1-\alpha} \left\{ \mu_n(\Gamma_1 \mathbf{Z}), \dots, \mu_n(\Gamma_{(n+m)!} \mathbf{Z}) \right\}.$$

Then, it follows from the definition of $\mu_n(\cdot)$ in Theorem 7 that

$$\mu_n(\mathbf{Z}) = \mu_n(\Gamma_0 \mathbf{Z}) = \max \left\{ \mu_n(\Gamma_1 \mathbf{Z}), \dots, \mu_n(\Gamma_{(n+m)!} \mathbf{Z}) \right\}.$$

Notice that

$$\begin{cases} \frac{n!m!}{(n+m)!} < 1 - \alpha & \text{if } m \neq n, \\ \frac{2(n!)^2}{(2n)!} < 1 - \alpha & \text{if } m = n, \end{cases} \text{ implies } \begin{cases} \frac{|\mathbb{S}_0|}{(n+m)!} < 1 - \alpha & \text{if } m \neq n, \\ \frac{|\mathbb{S}_0| + |\mathbb{S}_{\min\{n,m\}}|}{(n+m)!} < 1 - \alpha & \text{if } m = n, \end{cases}$$

and so $\mu_n(\Gamma_0 \mathbf{Z}) > Q_{1-\alpha} \left\{ \mu_n(\Gamma_1 \mathbf{Z}), \dots, \mu_n(\Gamma_{(n+m)!} \mathbf{Z}) \right\}$. Thus, as $p \rightarrow \infty$, we conclude

$$\begin{aligned} P \left(\text{ED}_n^k(\mathbf{Z}) > Q_{1-\alpha} \left\{ \text{ED}_n^k(\Gamma_1 \mathbf{Z}), \dots, \text{ED}_n^k(\Gamma_{(n+m)!} \mathbf{Z}) \right\} \right) \\ \rightarrow P \left(\mu_n(\Gamma_0 \mathbf{Z}) > Q_{1-\alpha} \left\{ \mu_n(\Gamma_1 \mathbf{Z}), \dots, \mu_n(\Gamma_{(n+m)!} \mathbf{Z}) \right\} \right) = 1. \end{aligned}$$

(ii) For any random permutation matrix Γ_s , by Proposition 35, we have $\text{ED}_n^k(\Gamma_s \mathbf{Z}) \xrightarrow{p} \mu_{n,W_s}$, where $W_s = N(\Gamma_s) \sim \text{Hypergeometric}(n+m, m, n)$. Then, the continuous mapping theorem implies that

$$\begin{aligned} P \left(\text{ED}_n^k(\mathbf{Z}) > Q_{1-\alpha} \left\{ \text{ED}_n^k(\mathbf{Z}), \text{ED}_n^k(\Gamma_1 \mathbf{Z}), \dots, \text{ED}_n^k(\Gamma_{S-1} \mathbf{Z}) \right\} \right) \\ \rightarrow P \left(\mu_{n,0} > Q_{1-\alpha} \left\{ \mu_{n,0}, \mu_{n,W_1}, \dots, \mu_{n,W_{S-1}} \right\} \right). \end{aligned}$$

Since $\mu_{n,0} = \max_w \mu_{n,w}$, in order to have $\mu_{n,0} = Q_{1-\alpha} \left\{ \mu_{n,0}, \mu_{n,W_1}, \dots, \mu_{n,W_{S-1}} \right\}$, at least $\lfloor \alpha S \rfloor + 1$ elements

of $\{\mu_{n,0}, \mu_{n,W_1}, \dots, \mu_{n,W_{S-1}}\}$ should be equal to $\mu_{n,0}$. Thus, we get

$$\begin{aligned} P(\mu_{n,0} > Q_{1-\alpha} \{\mu_{n,0}, \mu_{n,W_1}, \dots, \mu_{n,W_{S-1}}\}) \\ = 1 - P(\mu_{n,0} = Q_{1-\alpha} \{\mu_{n,0}, \mu_{n,W_1}, \dots, \mu_{n,W_{S-1}}\}) \\ \geq \begin{cases} 1 - \frac{S-1}{[\alpha S]} \frac{n!m!}{(n+m)!}, & \text{if } n \neq m, \\ 1 - \frac{S-1}{[\alpha S]} \frac{2(n!)^2}{(n+m)!}, & \text{if } n = m. \end{cases} \end{aligned}$$

2, (i) Since $\mu_n(\Gamma_u \mathbf{Z}) = 0$ for all $u = 1, 2, \dots, (n+m)!$ under H_{A_t} , Assumption 29 implies that

$$\sqrt{p} \text{ED}_n^k(\Gamma_u \mathbf{Z}) \xrightarrow{d} \sum_{i=1}^n \sum_{j=1}^m \Pi_{u,ij} b_{ij} - \sum_{1 \leq i < j \leq n} \Pi_{u,ij} c_{ij} - \sum_{1 \leq i < j \leq m} \Pi_{u,ij} d_{ij},$$

where $\Pi_{u,ij}$ corresponds to Γ_u . Then, the continuous mapping theorem entails

$$\begin{aligned} P(\text{ED}_n^k(\mathbf{Z}) > Q_{1-\alpha} \{\text{ED}_n^k(\Gamma_1 \mathbf{Z}), \dots, \text{ED}_n^k(\Gamma_{(n+m)!} \mathbf{Z})\}) \\ = P(\sqrt{p} \text{ED}_n^k(\mathbf{Z}) > Q_{1-\alpha} \{\sqrt{p} \text{ED}_n^k(\Gamma_1 \mathbf{Z}), \dots, \sqrt{p} \text{ED}_n^k(\Gamma_{(n+m)!} \mathbf{Z})\}) \\ \rightarrow P(V(\Gamma_0) > Q_{\hat{T}, 1-\alpha}). \end{aligned}$$

(ii) Conditioned on $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \dots, \mathbf{\Gamma}_{S-1}$, the result can be shown similarly with part(i). Then, since the number of permutations is fixed and finite, the unconditioned version follows straightforwardly.

3, (i) By construction, we have

$$\frac{1}{(n+m)!} \sum_{u=1}^{(n+m)!} \mathbb{I}_{\{V(\Gamma_u) > Q_{1-\alpha} \{V(\Gamma_1), \dots, V(\Gamma_{(n+m)!})\}\}} \leq \alpha.$$

If $\varphi'(e_{xy}) = \varphi'(e_x) = \varphi'(e_y)$ and $v_{xy} = v_x = v_y$, then $V(\Gamma_u) =^d V(\Gamma_0)$ for any $u = 1, 2, \dots, (n+m)!$ and so

$$(V(\Gamma_u), Q_{1-\alpha} \{V(\Gamma_1), \dots, V(\Gamma_{(n+m)!})\}) =^d (V(\Gamma_0), Q_{1-\alpha} \{V(\Gamma_1), \dots, V(\Gamma_{(n+m)!})\}).$$

Thus, we have

$$\begin{aligned} P(V(\Gamma_0) > Q_{1-\alpha} \{V(\Gamma_1), \dots, V(\Gamma_{(n+m)!})\}) = \\ E \left[\frac{1}{(n+m)!} \sum_{u=1}^{(n+m)!} \mathbb{I}_{\{V(\Gamma_u) > Q_{1-\alpha} \{V(\Gamma_1), \dots, V(\Gamma_{(n+m)!})\}\}} \right] \leq \alpha. \end{aligned}$$

(ii) The proof follows similarly from part (i) by observing that for any $s = 1, 2, \dots, S$

$$(V(\mathbf{\Gamma}_s), Q_{1-\alpha} \{V(\Gamma_0), V(\mathbf{\Gamma}_1), \dots, V(\mathbf{\Gamma}_{S-1})\}) =^d (V(\Gamma_0), Q_{1-\alpha} \{V(\Gamma_0), V(\mathbf{\Gamma}_1), \dots, V(\mathbf{\Gamma}_{S-1})\}).$$

2.6.5 Proof of Theorem 9

(i) Recall that for a fixed permutation matrix $\Gamma_w \in \mathbb{S}_w$ that corresponds to $\Pi_{w,ij}$,

$$\text{ED}_n^k(\Gamma_w \mathbf{Z}) = \underbrace{\sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{w,ij} \varphi(e_{ij})}_{:=\mu_n(\Gamma_w \mathbf{Z})} + \underbrace{\sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{w,ij} \mathcal{R}_1(Z_i, Z_j)}_{:=R_1(\Gamma_w \mathbf{Z})}.$$

Under the HDMSS setting, part (i) follows from Lemma 42.

Lemma 42. *Under Assumption 37, $\sup_{\Gamma} |R_1(\Gamma \mathbf{Z})| = o_p(1)$.*

Proof. Consider the events $B_{\mathbf{XY}}, B_{\mathbf{X}}, B_{\mathbf{Y}}$ and their complements $B_{\mathbf{XY}}^c, B_{\mathbf{X}}^c, B_{\mathbf{Y}}^c$, where

$$\begin{aligned} B_{\mathbf{XY}} &= \left\{ \min_{1 \leq s \leq n, 1 \leq t \leq m} \mathcal{L}(X_s, Y_t) \leq -\frac{1}{2}e_{xy} \text{ or } \max_{1 \leq s \leq n, 1 \leq t \leq m} \mathcal{L}(X_s, Y_t) \geq \frac{1}{2}e_{xy} \right\}, \\ B_{\mathbf{X}} &= \left\{ \min_{1 \leq s \neq t \leq n} \mathcal{L}(X_s, X_t) \leq -\frac{1}{2}e_x \text{ or } \max_{1 \leq s \neq t \leq n} \mathcal{L}(X_s, X_t) \geq \frac{1}{2}e_x \right\}, \\ B_{\mathbf{Y}} &= \left\{ \min_{1 \leq s \neq t \leq m} \mathcal{L}(Y_s, Y_t) \leq -\frac{1}{2}e_y \text{ or } \max_{1 \leq s \neq t \leq m} \mathcal{L}(Y_s, Y_t) \geq \frac{1}{2}e_y \right\}. \end{aligned}$$

Then, under assumption 37, as $n \wedge m \wedge p \rightarrow \infty$

$$\begin{aligned} P(B_{\mathbf{XY}}) &= P \left(\bigcup_{1 \leq s \leq n, 1 \leq t \leq m} \left\{ \mathcal{L}(X_s, Y_t) \leq -\frac{1}{2}e_{xy} \text{ or } \mathcal{L}(X_s, Y_t) \geq \frac{1}{2}e_{xy} \right\} \right) \\ &\leq \sum_{1 \leq s \leq n, 1 \leq t \leq m} P \left(|\mathcal{L}(X_s, Y_t)| \geq \frac{1}{2}e_{xy} \right) \\ &\leq nmP \left(|\mathcal{L}(X, Y)| \geq \frac{1}{2}e_{xy} \right) \\ &\leq \frac{4nmE[\mathcal{L}(X, Y)^2]}{e_{xy}^2} \\ &= o(1). \end{aligned}$$

Similarly, we can show that $P(B_{\mathbf{X}}) = o(1)$ and $P(B_{\mathbf{Y}}) = o(1)$. Conditioned on event $B_{\mathbf{XY}}^c B_{\mathbf{X}}^c B_{\mathbf{Y}}^c$, we have $e_{ij} \leq e_{ij} + v\mathcal{L}(Z_i, Z_j) \leq 3e_{ij}/2$ for any $0 \leq v \leq 1$. Suppose $\varphi^{(1)}(\cdot)$ is a continuous function on $(0, +\infty)$, we know there exist a constant C such that $|\varphi^{(1)}(e_{ij} + v\mathcal{L}(Z_i, Z_j))| \leq C$ and consequently, we have

$$|\mathcal{R}_1(Z_i, Z_j)| \leq C' |\mathcal{L}(Z_i, Z_j)|,$$

where C' is a constant depends only on φ, e_{xy}, e_x and e_y . Let Π_{ij} corresponds to Γ ,

$$\begin{aligned}
& \sup_{\Gamma} \left| \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{ij} \mathcal{R}_1(Z_i, Z_j) \right| \\
& \leq \sup_{\Gamma} \left| \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{ij} \mathcal{R}_1(Z_i, Z_j) \left\{ \mathbb{I}_{B_{\mathbf{XY}}^c} \mathbb{I}_{B_{\mathbf{X}}^c} \mathbb{I}_{B_{\mathbf{Y}}^c} + \mathbb{I}_{B_{\mathbf{XY}}} + \mathbb{I}_{B_{\mathbf{X}}} + \mathbb{I}_{B_{\mathbf{Y}}} \right\} \right| \\
& \leq \sup_{\Gamma} \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} |\Pi_{ij} \mathcal{R}_1(Z_i, Z_j)| \mathbb{I}_{B_{\mathbf{XY}}^c} \mathbb{I}_{B_{\mathbf{X}}^c} \mathbb{I}_{B_{\mathbf{Y}}^c} + o_p(1) \\
& \leq \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \frac{C''}{nm} |\mathcal{R}_1(Z_i, Z_j)| \mathbb{I}_{B_{\mathbf{XY}}^c} \mathbb{I}_{B_{\mathbf{X}}^c} \mathbb{I}_{B_{\mathbf{Y}}^c} + o_p(1),
\end{aligned}$$

where C'' is a constant depends only on ρ . Then, for any $\epsilon > 0$, by Markov's inequality

$$\begin{aligned}
P \left(\frac{1}{mn} \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} |\mathcal{R}_1(Z_i, Z_j)| \mathbb{I}_{B_{\mathbf{XY}}^c} \mathbb{I}_{B_{\mathbf{X}}^c} \mathbb{I}_{B_{\mathbf{Y}}^c} > \epsilon \right) \\
\leq \frac{1}{\epsilon mn} \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} E[|\mathcal{R}_1(Z_i, Z_j)| \mathbb{I}_{B_{\mathbf{XY}}^c} \mathbb{I}_{B_{\mathbf{X}}^c} \mathbb{I}_{B_{\mathbf{Y}}^c}] \leq C''' \frac{(\alpha_{xy} + \alpha_x + \alpha_y)}{\epsilon},
\end{aligned}$$

where C''' is a constant depends only on $\rho, \varphi, e_{xy}, e_x$ and e_y . □

(ii) Similar to the HDLSS setting, we consider the following decomposition

$$\begin{aligned}
\sqrt{nm} p \text{ED}_n^k(\Gamma_w \mathbf{Z}) &= \underbrace{\sqrt{nm} \mu_n(\Gamma_w \mathbf{Z}) + \sqrt{nm} \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{w,ij} \varphi^{(1)}(e_{ij}) \mathcal{K}(Z_i, Z_j)}_{:= \sqrt{nm} L(\Gamma_w \mathbf{Z})} \\
&+ \underbrace{\sqrt{nm} \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{w,ij} \varphi^{(1)}(e_{ij}) \mathcal{W}(Z_i, Z_j)}_{:= \sqrt{nm} R_l(\Gamma_w \mathbf{Z})} + \underbrace{\sqrt{nm} p \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{w,ij} \mathcal{R}_2(Z_i, Z_j)}_{:= \sqrt{nm} R_2(\Gamma_w \mathbf{Z})}.
\end{aligned}$$

Next, for any $-\infty < a < \infty$ and $\epsilon > 0$, using the inequality $P(X \leq a) \leq P(Y \leq a + \epsilon) + P(|X - Y| > \epsilon)$, we can show that

$$\begin{aligned}
P(\sqrt{nm} L(\Gamma_w \mathbf{Z}) \leq a - \epsilon) &= P(|\sqrt{nm} R_l(\Gamma_w \mathbf{Z}) + \sqrt{nm} R_2(\Gamma_w \mathbf{Z})| > \epsilon) \\
&\leq P(\sqrt{nm} p [\text{ED}_n^k(\Gamma_w \mathbf{Z}) - \mu_n(\Gamma_w \mathbf{Z})] \leq a) \\
&\leq P(\sqrt{nm} L(\Gamma_w \mathbf{Z}) \leq a + \epsilon) + P(|\sqrt{nm} R_l(\Gamma_w \mathbf{Z}) + \sqrt{nm} R_2(\Gamma_w \mathbf{Z})| > \epsilon).
\end{aligned}$$

Then, some algebra shows that

$$\begin{aligned}
& \sup_w \left| P(\sqrt{nm}p[\text{ED}_n^k(\Gamma_w \mathbf{Z}) - \mu_n(\Gamma_w \mathbf{Z})] \leq a) - \Phi\left(a/\sqrt{nm\sigma_n^2(\Gamma_w \mathbf{Z})}\right) \right| \\
& \leq \sup_w \left| P(\sqrt{nm}L(\Gamma_w \mathbf{Z}) \leq a - \epsilon) - \Phi\left((a - \epsilon)/\sqrt{nm\sigma_n^2(\Gamma_w \mathbf{Z})}\right) \right| \\
& \quad + \sup_w \left| \Phi\left(a/\sqrt{nm\sigma_n^2(\Gamma_w \mathbf{Z})}\right) - \Phi\left((a - \epsilon)/\sqrt{nm\sigma_n^2(\Gamma_w \mathbf{Z})}\right) \right| \\
& + \sup_w \left| P(\sqrt{nm}L(\Gamma_w \mathbf{Z}) \leq a + \epsilon) - \Phi\left((a + \epsilon)/\sqrt{nm\sigma_n^2(\Gamma_w \mathbf{Z})}\right) \right| \\
& \quad + \sup_w \left| \Phi\left(a/\sqrt{nm\sigma_n^2(\Gamma_w \mathbf{Z})}\right) - \Phi\left((a + \epsilon)/\sqrt{nm\sigma_n^2(\Gamma_w \mathbf{Z})}\right) \right| \\
& + 2 \sup_w P(|\sqrt{nm}R_l(\Gamma_w \mathbf{Z}) + \sqrt{nm}R_2(\Gamma_w \mathbf{Z})| > \epsilon).
\end{aligned}$$

Next, by Lemma 43, 44 and 45. the right hand side can be made arbitrarily small by first choose ϵ small enough, then n, m, p large enough.

Lemma 43. Under Assumption 37 and 38, $\sup_{\Gamma} |\sqrt{nm}R_2(\Gamma \mathbf{Z})| = o_p(1)$.

Proof. The proof is similar with Lemma 42 by observing that conditioned on event $B_{\mathbf{X}\mathbf{Y}}^c B_{\mathbf{X}}^c B_{\mathbf{Y}}^c$, it holds for some constant C that $|\mathcal{R}_2(Z_i, Z_j)| \leq C |\mathcal{L}^2(Z_i, Z_j)|$. \square

Lemma 44. Under H_{A_l} , $\sup_{\Gamma \in \mathbb{P}_{n+m}} |\sqrt{nm}R_l(\Gamma \mathbf{Z})| = o_p(1)$.

Proof. For any fixed permutaiton marix $\Gamma \in \mathbb{P}_{n+m}$, we have

$$\begin{aligned}
\sqrt{nm}R_l(\Gamma \mathbf{Z}) &= \sqrt{nm}p \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{ij} \varphi^{(1)}(e_{ij}) (E[\bar{\psi}(Z_i, Z_j)|Z_i] + E[\bar{\psi}(Z_i, Z_j)|Z_j]) \\
&\quad - \sqrt{nm}p \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{ij} \varphi^{(1)}(e_{ij}) (E[\bar{\psi}(Z_i, Z_j)] + e_{ij}).
\end{aligned}$$

Let $w = N(\Gamma)$, similar to the computation of $\mu_{n,w}$, we obtain

$$\begin{aligned}
\sqrt{nm}p \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{ij} \varphi^{(1)}(e_{ij}) E[\bar{\psi}(Z_i, Z_j)] &= \sqrt{nm}p \left\{ 2\varphi^{(1)}(e_{xy}) E[\bar{\psi}(X, Y)] \right. \\
&\quad \left. - \varphi^{(1)}(e_x) E[\bar{\psi}(X, X')] - \varphi^{(1)}(e_y) E[\bar{\psi}(Y, Y')] \right\} f(w),
\end{aligned}$$

where the right hand side is of order $o_p(1)$ under H_{A_l} . Let π corresponds to Γ , then for each $1 \leq i \leq n$ such that $1 \leq \pi(i) \leq n$, it follows from the definition of Π_{ij} that

$$\begin{aligned}
& \frac{1}{4} \sum_{1 \leq j \leq n+m}^{j \neq i} \Pi_{ij} \varphi^{(1)}(e_{ij}) E[\bar{\psi}(X_i, Z_j)|X_i] \\
&= -\frac{(n-w-1)}{n(n-1)} \varphi^{(1)}(e_x) E[\bar{\psi}(X_i, X)|X_i] - \frac{w}{n(n-1)} \varphi^{(1)}(e_{xy}) E[\bar{\psi}(X_i, Y)|X_i] \\
&\quad + \frac{w}{nm} \varphi^{(1)}(e_x) E[\bar{\psi}(X_i, X)|X_i] + \frac{m-w}{nm} \varphi^{(1)}(e_{xy}) E[\bar{\psi}(X_i, Y)|X_i] \\
&= \left(\frac{1}{n} - \frac{w}{nm} - \frac{w}{n(n-1)} \right) \left\{ \varphi^{(1)}(e_{xy}) E[\bar{\psi}(X_i, Y)|X_i] - \varphi^{(1)}(e_x) E[\bar{\psi}(X_i, X)|X_i] \right\},
\end{aligned}$$

which entails

$$\sup_{\Gamma} \left| \frac{1}{4} \sum_{1 \leq j \leq n+m}^{j \neq i} \Pi_{ij} \varphi^{(1)}(e_{ij}) E [\bar{\psi}(X_i, Z_j) | X_i] \right| \leq \frac{C}{\sqrt{nm}} \left| \varphi^{(1)}(e_{xy}) E [\bar{\psi}(X_i, Y) | X_i] - \varphi^{(1)}(e_x) E [\bar{\psi}(X_i, X) | X_i] \right|,$$

where C is a constant that only depends on ρ . Using the same approach, the above bound can be shown to hold for each $1 \leq i \leq n$ such that $n+1 \leq \pi(i) \leq n+m$. Similarly, we can show that for each $n+1 \leq i \leq n+m$,

$$\sup_{\Gamma} \left| \frac{1}{4} \sum_{1 \leq j \leq n+m}^{j \neq i} \Pi_{ij} \varphi^{(1)}(e_{ij}) E [\bar{\psi}(Y_i, Z_j) | Y_i] \right| \leq \frac{C}{\sqrt{nm}} \left| \varphi^{(1)}(e_{xy}) E [\bar{\psi}(Y_i, X) | Y_i] - \varphi^{(1)}(e_y) E [\bar{\psi}(Y_i, Y) | Y_i] \right|.$$

Consequently, the following bound holds

$$\begin{aligned} \sup_{\Gamma} \left| \sqrt{nm} p \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \Pi_{ij} \varphi^{(1)}(e_{ij}) (E [\bar{\psi}(Z_i, Z_j) | Z_i] + E [\bar{\psi}(Z_i, Z_j) | Z_j]) \right| \\ \leq C' \sqrt{p} \sum_{i=1}^n \left| \varphi^{(1)}(e_{xy}) E [\bar{\psi}(X_i, Y) | X_i] - \varphi^{(1)}(e_x) E [\bar{\psi}(X_i, X) | X_i] \right| \\ + C' \sqrt{p} \sum_{i=n+1}^{n+m} \left| \varphi^{(1)}(e_{xy}) E [\bar{\psi}(Y_i, X) | Y_i] - \varphi^{(1)}(e_y) E [\bar{\psi}(Y_i, Y) | Y_i] \right|, \end{aligned}$$

where C' is a constant. Finally, an application of Markov's inequality shows that the right hand side is of order $o_p(1)$ under H_{A_1} . \square

Lemma 45. *Under Assumptions 28. Let $\Gamma_1, \Gamma_2 \in \mathbb{P}_{n+m}$. Then, for any constants a_1, a_2, b , we have*

$$\begin{aligned} \sup_{\Gamma_1, \Gamma_2} \left| P(a_1 \sqrt{nm} L(\Gamma_1 \mathbf{Z}) + a_2 \sqrt{nm} L(\Gamma_2 \mathbf{Z}) \leq b) - \Phi \left(\frac{b}{\eta_{a_1, a_2}(\Gamma_1, \Gamma_2)} \right) \right| \\ \leq C \left\{ \max_{\Lambda_1, \Lambda_2 \in \{X, Y\}} E [\mathcal{K}^4(\Lambda_1, \Lambda_2')] / n^2 \right. \\ + \max_{\Lambda_1, \Lambda_2, \Lambda_3 \in \{X, Y\}} E [\mathcal{K}^2(\Lambda_1, \Lambda_3'') \mathcal{K}^2(\Lambda_2', \Lambda_3'')] / n \\ + \max_{\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4 \in \{X, Y\}} E [\mathcal{K}(\Lambda_1, \Lambda_3'') \mathcal{K}(\Lambda_1, \Lambda_4''') \mathcal{K}(\Lambda_2', \Lambda_4''') \mathcal{K}(\Lambda_2', \Lambda_3'')] \\ \left. + (v_x - \text{var}[\mathcal{K}(X, X')])^2 + (v_y - \text{var}[\mathcal{K}(Y, Y')])^2 + (v_{xy} - \text{var}[\mathcal{K}(X, Y)])^2 \right\}^{1/5}, \end{aligned}$$

where C is a constant depend on φ, ρ, e_x, e_y and e_{xy} only; $\eta_{a_1, a_2}(\Gamma_1, \Gamma_2)$ is defined as

$$[\eta_{a_1, a_2}(\Gamma_1, \Gamma_2)]^2 = nm \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} (a_1 \Pi_{1, ij} + a_2 \Pi_{2, ij})^2 [\varphi^{(1)}(e_{ij})]^2 v_{ij},$$

where $\Pi_{1,ij}, \Pi_{2,ij}$ correspond to Γ_1, Γ_2 respectively and

$$v_{ij} = \begin{cases} v_x, & \text{if } 1 \leq i, j \leq n, \\ v_y, & \text{if } n+1 \leq i, j \leq n+m, \\ v_{xy}, & \text{otherwise.} \end{cases}$$

Proof. Notice that

$$\sqrt{nm} (a_1 L(\Gamma_1 \mathbf{Z}) + a_2 L(\Gamma_2 \mathbf{Z})) = \sqrt{nm} \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} (a_1 \Pi_{1,ij} + a_2 \Pi_{2,ij}) \varphi^{(1)}(e_{ij}) \mathcal{K}(Z_i, Z_j).$$

Then, for notational convenience, set

$$\mathcal{H}(Z_i, Z_j) = \sqrt{nm} (a_1 \Pi_{1,ij} + a_2 \Pi_{2,ij}) \varphi^{(1)}(e_{ij}) \mathcal{K}(Z_i, Z_j),$$

and $S_{n+m,l} = \sum_{i=2}^l \xi_{n+m,i}$, where $\xi_{n+m,i} = \sum_{j=1}^{i-1} \mathcal{H}(Z_i, Z_j)$. Next, let

$$\mathcal{F}_{n+m,l} = \sigma(Z_1, Z_2, \dots, Z_l)$$

be the σ -algebra generated by Z_1, \dots, Z_l , we have $\{S_{n+m,l}, \mathcal{F}_{n+m,l}, 1 \leq l \leq n+m\}$ is a martingale array and thus we can apply the Berry-Esseen type bound for martingale sequences [Theorem 1 of Hall and Heyde (1981)]. By setting $m = 0$ and $\delta = 1$ in Theorem 1 of Hall and Heyde (1981), we compute

$$\sum_{i=2}^{n+m} E [\xi_{n+m,i}^2], \text{var} \left[\sum_{i=2}^{n+m} E [\xi_{n+m,i}^2 | \mathcal{F}_{n+m,i-1}] \right] \text{ and } \sum_{i=2}^{n+m} E [\xi_{n+m,i}^4]$$

Firstly, due to the property of double centering, $E[\mathcal{H}(Z_i, Z_j) \mathcal{H}(Z_{i'}, Z_{j'})] \neq 0$ only when $\{i, j\} = \{i', j'\}$. Then, let $\eta_{a_1, a_2}(\Gamma_1, \Gamma_2)$ be in Theorem 1 of Hall and Heyde (1981)

$$\frac{\eta_{a_1, a_2}^2(\Gamma_1, \Gamma_2)}{nm} := \lim_{p \rightarrow \infty} \frac{\sum_{i=2}^{n+m} E [\xi_{n+m,i}^2]}{nm} = \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} (a_1 \Pi_{1,ij} + a_2 \Pi_{2,ij})^2 [\varphi^{(1)}(e_{ij})]^2 v_{ij}.$$

Then, to calculate the variance, notice that

$$\text{var} \left[\sum_{i=2}^{n+m} E [\xi_{n+m,i}^2 | \mathcal{F}_{n+m,i-1}] \right] = \sum_{i_1, i_2=2}^{n+m} \sum_{j_1, j_2=1}^{i_1-1} \sum_{j_3, j_4=1}^{i_2-1} \Theta(i_1, i_2; j_1, j_2, j_3, j_4),$$

where $\Theta(i_1, i_2; j_1, j_2, j_3, j_4)$ is defined as

$$\Theta(i_1, i_2; j_1, j_2, j_3, j_4) = \text{cov} [E [\mathcal{H}(Z_{i_1}, Z_{j_1}) \mathcal{H}(Z_{i_1}, Z_{j_2}) | Z_{j_1}, Z_{j_2}], E [\mathcal{H}(Z_{i_2}, Z_{j_3}) \mathcal{H}(Z_{i_2}, Z_{j_4}) | Z_{j_3}, Z_{j_4}]].$$

Next, for any $1 \leq j_1, j_2 \leq n+m$, $\Lambda \in \{X, Y\}$, denote

$$\mathcal{G}_\Lambda(Z_{j_1}, Z_{j_2}) = E[\mathcal{K}(\Lambda, Z_{j_1}) \mathcal{K}(\Lambda, Z_{j_2}) | Z_{j_1}, Z_{j_2}].$$

To bound each $\Theta(i_1, i_2; j_1, j_2, j_3, j_4)$, we need to study the covariance between $\mathcal{G}_{\Lambda_1}(Z_{j_1}, Z_{j_2})$ and $\mathcal{G}_{\Lambda_2}(Z_{j'_1}, Z_{j'_2})$.

Lemma 46. *Then, for any $1 \leq j_1, j_2, j'_1, j'_2 \leq n + m$, $\Lambda_1, \Lambda_2 \in \{X, Y\}$, we have*

$$\begin{aligned} & \text{cov} [\mathcal{G}_{\Lambda_1}(Z_{j_1}, Z_{j_2}), \mathcal{G}_{\Lambda_2}(Z_{j'_1}, Z_{j'_2})] \\ &= \begin{cases} E [\mathcal{K}^2(\Lambda_1, Z_{j_1}) \mathcal{K}^2(\Lambda'_2, Z_{j'_1})] - E [\mathcal{K}^2(\Lambda_1, Z_{j_1})] E [\mathcal{K}^2(\Lambda_2, Z_{j'_1})], & j_1 = j_2 = j'_1 = j'_2; \\ E [\mathcal{K}(\Lambda_1, Z_{j_1}) \mathcal{K}(\Lambda_1, Z_{j_2}) \mathcal{K}(\Lambda'_2, Z_{j'_2}) \mathcal{K}(\Lambda'_2, Z_{j'_1})], & j_1 = j'_1 \neq j_2 = j'_2; \\ E [\mathcal{K}(\Lambda_1, Z_{j_1}) \mathcal{K}(\Lambda_1, Z_{j_2}) \mathcal{K}(\Lambda'_2, Z_{j'_2}) \mathcal{K}(\Lambda'_2, Z_{j'_1})], & j_1 = j'_2 \neq j_2 = j'_1; \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Proof. If $j_1 = j'_2 \neq j_2 = j'_1$,

$$\begin{aligned} & E [\mathcal{G}_{\Lambda_1}(Z_{j_1}, Z_{j_2}) \mathcal{G}_{\Lambda_2}(Z_{j'_1}, Z_{j'_2})] \\ &= E [E [\mathcal{K}(\Lambda_1, Z_{j_1}) \mathcal{K}(\Lambda_1, Z_{j_2}) | Z_{j_1}, Z_{j_2}] E [\mathcal{K}(\Lambda'_2, Z_{j'_1}) \mathcal{K}(\Lambda'_2, Z_{j'_2}) | Z_{j_1}, Z_{j_2}]] \\ &= E [E [\mathcal{K}(\Lambda_1, Z_{j_1}) \mathcal{K}(\Lambda_1, Z_{j_2}) \mathcal{K}(\Lambda'_2, Z_{j'_2}) \mathcal{K}(\Lambda'_2, Z_{j'_1}) | Z_{j_1}, Z_{j_2}]] \\ &= E [\mathcal{K}(\Lambda_1, Z_{j_1}) \mathcal{K}(\Lambda_1, Z_{j_2}) \mathcal{K}(\Lambda'_2, Z_{j'_2}) \mathcal{K}(\Lambda'_2, Z_{j'_1})] \end{aligned}$$

It can be shown similarly for cases $j_1 = j_2 = j_3 = j_4$ and $j_1 = j'_1 \neq j_2 = j'_2$. Next, we show that for other cases, the covariance is 0. We take $j_1 = j'_1, j_1 \neq j_2, j_1 \neq j'_2, j_2 \neq j'_2$ as an example

$$\begin{aligned} & E [\mathcal{G}_{\Lambda_1}(Z_{j_1}, Z_{j_2}) \mathcal{G}_{\Lambda_2}(Z_{j'_1}, Z_{j'_2})] \\ &= E [E [\mathcal{K}(\Lambda_1, Z_{j_1}) \mathcal{K}(\Lambda_1, Z_{j_2}) | Z_{j_1}, Z_{j_2}] E [\mathcal{K}(\Lambda'_2, Z_{j'_1}) \mathcal{K}(\Lambda'_2, Z_{j'_2}) | Z_{j_1}, Z_{j_2}]] \\ &= E [E [\mathcal{K}(\Lambda_1, Z_{j_1}) \mathcal{K}(\Lambda_1, Z_{j_2}) \mathcal{K}(\Lambda'_2, Z_{j'_1}) \mathcal{K}(\Lambda'_2, Z_{j'_2}) | Z_{j_1}, Z_{j_2}, Z_{j'_2}]] \\ &= E [\mathcal{K}(\Lambda_1, Z_{j_1}) \mathcal{K}(\Lambda_1, Z_{j_2}) \mathcal{K}(\Lambda'_2, Z_{j'_1}) \mathcal{K}(\Lambda'_2, Z_{j'_2})] \\ &= E [\mathcal{K}(\Lambda_1, Z_{j_1}) \mathcal{K}(\Lambda'_2, Z_{j'_1}) E [\mathcal{K}(\Lambda_1, Z_{j_2}) | \Lambda_1, \Lambda'_2, Z_{j_1}] E [\mathcal{K}(\Lambda'_2, Z_{j'_2}) | \Lambda_1, \Lambda'_2, Z_{j_1}]] \\ &= 0. \end{aligned}$$

□

Next, we can bound $\text{var} \left[\sum_{i=2}^{n+m} E \left[\xi_{n+m,i}^2 \mid \mathcal{F}_{n+m,i-1} \right] \right]$ as

$$\begin{aligned}
& \sum_{i_1, i_2=1}^{n+m} \sum_{j_1, j_2=1}^{i_1-1} \sum_{j_3, j_4=1}^{i_2-1} \Theta(i_1, i_2; j_1, j_2, j_3, j_4) \\
&= \sum_{i=1}^{n+m} \left\{ \sum_{j=1}^{i-1} \Theta(i, i; j, j, j, j) + 2 \sum_{1 \leq j_1 \neq j_2 \leq i-1} \Theta(i, i; j_1, j_2, j_1, j_2) \right\} \\
&\quad + 2 \sum_{1 \leq i_1 < i_2 \leq n+m} \left\{ \sum_{j=1}^{i_1-1} \Theta(i_1, i_2; j, j, j, j) + 2 \sum_{1 \leq j_1 \neq j_2 \leq i_1-1} \Theta(i_1, i_2; j_1, j_2, j_1, j_2) \right\} \\
&= O \left(\max_{\Lambda_1, \Lambda_2, \Lambda_3 \in \{X, Y\}} E \left[\mathcal{K}^2(\Lambda_1, \Lambda_3'') \mathcal{K}^2(\Lambda_2', \Lambda_3'') \right] / n \right. \\
&\quad \left. + \max_{\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4 \in \{X, Y\}} E \left[\mathcal{K}(\Lambda_1, \Lambda_3'') \mathcal{K}(\Lambda_1, \Lambda_4''') \mathcal{K}(\Lambda_2', \Lambda_4''') \mathcal{K}(\Lambda_2', \Lambda_3'') \right] \right)
\end{aligned}$$

Finally, to find the upper bound of $\sum_{i=2}^{n+m} E \left(\xi_{n+m,i}^4 \right)$,

$$\begin{aligned}
& \sum_{i=2}^{n+m} E \left(\xi_{n+m,i}^4 \right) \\
&= \sum_{i=2}^{n+m} \sum_{j_1, j_2, j_3, j_4=1}^{i-1} E \left[\mathcal{H}(Z_i, Z_{j_1}) \mathcal{H}(Z_i, Z_{j_2}) \mathcal{H}(Z_i, Z_{j_3}) \mathcal{H}(Z_i, Z_{j_4}) \right] \\
&= \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} E \left[\mathcal{H}^4(Z_i, Z_j) \right] + 6 \sum_{i=2}^{n+m} \sum_{1 \leq j_1 < j_2 \leq i-1} E \left[\mathcal{H}^2(Z_i, Z_{j_1}) \mathcal{H}^2(Z_i, Z_{j_2}) \right] \\
&= O \left(\max_{\Lambda_1, \Lambda_2 \in \{X, Y\}} E \left[\mathcal{K}^4(\Lambda_1, \Lambda_2') \right] / n^2 \right) \\
&\quad + O \left(\max_{\Lambda_1, \Lambda_2, \Lambda_3 \in \{X, Y\}} E \left[\mathcal{K}^2(\Lambda_1, \Lambda_3'') \mathcal{K}^2(\Lambda_2', \Lambda_3'') \right] / n \right).
\end{aligned}$$

Combining the above bounds, the lemma is a consequence of Theorem 1 in Hall and Heyde (1981). \square

2.6.6 Proof of Theorem 10

(i) For a random permutation matrix $\mathbf{\Gamma} \sim \text{Uniform}(\mathbb{P}_{n+m})$, it follows from Lemma 42 that

$$\text{ED}_n^k(\mathbf{\Gamma}\mathbf{Z}) = \mu_n(\mathbf{\Gamma}\mathbf{Z}) + R_1(\mathbf{\Gamma}\mathbf{Z}) = (2\varphi(e_{xy}) - \varphi(e_x) - \varphi(e_y))f(W) + o_p(1)$$

where $W = N(\mathbf{\Gamma}) \sim \text{Hypergeometric}(m+n, m, n)$. From the normal limit of hypergeometric distribution Lahiri et al. (2006), we know that

$$\frac{W}{\sqrt{nm}} \xrightarrow{p} \frac{\sqrt{\rho}}{1+\rho}.$$

Next, some algebra shows that $f(W) \xrightarrow{p} 0$ and so the result is proved.

(ii) Recall that we can decompose the sample energy distance as

$$\begin{aligned}\sqrt{nm}p\text{ED}_n^k(\mathbf{\Gamma}\mathbf{Z}) &= \underbrace{\sqrt{nm}p\mu_n(\mathbf{\Gamma}\mathbf{Z}) + \sqrt{nm} \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \mathbf{\Pi}_{ij} \varphi^{(1)}(e_{ij}) \mathcal{K}(Z_i, Z_j)}_{:=\sqrt{nm}L(\mathbf{\Gamma}\mathbf{Z})} \\ &\quad + \underbrace{\sqrt{nm} \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \mathbf{\Pi}_{ij} \varphi^{(1)}(e_{ij}) \mathcal{W}(Z_i, Z_j)}_{:=\sqrt{nm}R_l(\mathbf{\Gamma}\mathbf{Z})} + \underbrace{\sqrt{nm}p \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \mathbf{\Pi}_{ij} \mathcal{R}_2(Z_i, Z_j)}_{:=\sqrt{nm}R_2(\mathbf{\Gamma}\mathbf{Z})}.\end{aligned}$$

The result is a consequence of Lemma 44, 43 and 47.

Lemma 47. *Under Assumptions 28 and 40,*

$$\sqrt{nm} \begin{pmatrix} L(\mathbf{\Gamma}\mathbf{Z}) \\ L(\mathbf{\Gamma}'\mathbf{Z}) \end{pmatrix} \xrightarrow{d} N \left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right)$$

where $\mathbf{\Gamma}'$ is independent copy of $\mathbf{\Gamma}$ and σ^2 is the asymptotic variance defined as

$$\sigma^2 := 4v_{xy}[\varphi^{(1)}(e_{xy})]^2 + 2\rho v_x[\varphi^{(1)}(e_x)]^2 + \frac{2}{\rho}v_y[\varphi^{(1)}(e_y)]^2.$$

Proof. We apply the Cramér-Wold device. For any constants a_1, a_2 , we have

$$\begin{aligned}\eta_{a_1, a_2}^2(\mathbf{\Gamma}, \mathbf{\Gamma}') &= nm \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} (a_1 \mathbf{\Pi}_{ij} + a_2 \mathbf{\Pi}'_{ij})^2 [\varphi^{(1)}(e_{ij})]^2 v_{ij} \\ &= nm \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} (a_1^2 \mathbf{\Pi}_{ij}^2 + a_2^2 (\mathbf{\Pi}'_{ij})^2 + 2a_1 a_2 \mathbf{\Pi}_{ij} \mathbf{\Pi}'_{ij}) [\varphi^{(1)}(e_{ij})]^2 v_{ij}.\end{aligned}$$

Notice that for $\{i_1, j_1\} \cap \{i_2, j_2\} = \emptyset$, it can be shown that $E[\mathbf{\Pi}_{i_1 j_1} \mathbf{\Pi}_{i_2 j_2}] = O(1/n^5)$. Then, denote $c_{ij} = 2a_1 a_2 [\varphi^{(1)}(e_{ij})]^2 v_{ij}$, we have

$$\begin{aligned}E \left[\left(nm \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} c_{ij} \mathbf{\Pi}_{ij} \mathbf{\Pi}'_{ij} \right)^2 \right] &= n^2 m^2 \sum_{i=2}^{n+m} \sum_{j_1=1}^{i-1} \sum_{j_2=1}^{i-1} c_{ij_1} c_{ij_2} E^2[\mathbf{\Pi}_{ij_1} \mathbf{\Pi}_{ij_2}] \\ &\quad + 2n^2 m^2 \sum_{2 \leq i_1 < i_2 \leq n+m} \sum_{j=1}^{i_1-1} c_{i_1 j} c_{i_2 j} E^2[\mathbf{\Pi}_{i_1 j} \mathbf{\Pi}_{i_2 j}] \\ &\quad + n^2 m^2 \sum_{2 \leq i_1 \neq i_2 \leq n+m} \sum_{j_1 \neq j_2} c_{i_1 j_1} c_{i_2 j_2} E^2[\mathbf{\Pi}_{i_1 j_1} \mathbf{\Pi}_{i_2 j_2}] \\ &= O(1/n).\end{aligned}$$

In addition, let $W = N(\mathbf{\Gamma})$, we obtain

$$\begin{aligned}
\sigma_n^2(\mathbf{\Gamma}\mathbf{Z}) &:= \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \mathbf{\Pi}_{ij}^2 [\varphi^{(1)}(e_{ij})]^2 v_{ij} \\
&= \left\{ \frac{4}{nm} - 4 \left(\frac{n+m}{n^2 m^2} - \frac{n}{n^2(n-1)^2} - \frac{m}{m^2(m-1)^2} \right) W \right. \\
&\quad \left. + 4 \left(\frac{2}{n^2 m^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) W^2 \right\} v_{xy} [\varphi^{(1)}(e_{xy})]^2 \\
&+ \left\{ \frac{2}{n(n-1)} + 2 \left(\frac{2n}{n^2 m^2} - \frac{2n-1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) W \right. \\
&\quad \left. - 2 \left(\frac{2}{m^2 n^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) W^2 \right\} v_x [\varphi^{(1)}(e_x)]^2 \\
&+ \left\{ \frac{2}{m(m-1)} + 2 \left(\frac{2m}{n^2 m^2} - \frac{1}{n^2(n-1)^2} - \frac{2m-1}{m^2(m-1)^2} \right) W \right. \\
&\quad \left. - 2 \left(\frac{2}{n^2 m^2} - \frac{1}{m^2(m-1)^2} - \frac{1}{n^2(n-1)^2} \right) W^2 \right\} v_y [\varphi^{(1)}(e_y)]^2.
\end{aligned}$$

Since $W/\sqrt{nm} \xrightarrow{P} \sqrt{\rho}/(1+\rho)$, some algebra shows that

$$\sigma_n^2(\mathbf{\Gamma}\mathbf{Z}) := \sum_{i=2}^{n+m} \sum_{j=1}^{i-1} \mathbf{\Pi}_{ij}^2 [\varphi^{(1)}(e_{ij})]^2 v_{ij} \xrightarrow{P} \sigma^2,$$

which entails that $\eta_{a_1, a_2}^2(\mathbf{\Gamma}, \mathbf{\Gamma}') \xrightarrow{P} a_1^2 \sigma^2 + a_2^2 \sigma^2$. Since $|\Phi(\cdot)| \leq 1$, we have

$$E \left[\left| \Phi \left(\frac{b}{\eta_{a_1, a_2}(\mathbf{\Gamma}, \mathbf{\Gamma}')} \right) - \Phi \left(\frac{b}{\sqrt{a_1^2 \sigma^2 + a_2^2 \sigma^2}} \right) \right| \right] \rightarrow 0.$$

Next, by a simple triangle inequality

$$\begin{aligned}
&\left| P(a_1 \sqrt{nm} L(\mathbf{\Gamma}\mathbf{Z}) + a_2 \sqrt{nm} L(\mathbf{\Gamma}'\mathbf{Z}) \leq b) - \Phi \left(\frac{b}{\sqrt{a_1^2 \sigma^2 + a_2^2 \sigma^2}} \right) \right| \leq \\
&\left| P(a_1 \sqrt{nm} L(\mathbf{\Gamma}\mathbf{Z}) + a_2 \sqrt{nm} L(\mathbf{\Gamma}'\mathbf{Z}) \leq b) - \Phi \left(\frac{b}{\sqrt{\eta_{a_1, a_2}^2(\mathbf{\Gamma}, \mathbf{\Gamma}')}} \right) \right| \\
&\quad + \left| \Phi \left(\frac{b}{\sqrt{\eta_{a_1, a_2}^2(\mathbf{\Gamma}, \mathbf{\Gamma}')}} \right) - \Phi \left(\frac{b}{\sqrt{a_1^2 \sigma^2 + a_2^2 \sigma^2}} \right) \right|.
\end{aligned}$$

Taking expectation with respect to $\mathbf{\Gamma}, \mathbf{\Gamma}'$ on both sides, then it follows from Lemma 45 and Assumption 40

that

$$\begin{aligned}
& \left| P(a_1\sqrt{nm}L(\mathbf{\Gamma}\mathbf{Z}) + a_2\sqrt{nm}L(\mathbf{\Gamma}'\mathbf{Z}) \leq b) - \Phi\left(\frac{b}{\sqrt{a_1^2\sigma^2 + a_2^2\sigma^2}}\right) \right| \leq \\
& E \left[\left| P(a_1\sqrt{nm}L(\mathbf{\Gamma}\mathbf{Z}) + a_2\sqrt{nm}L(\mathbf{\Gamma}'\mathbf{Z}) \leq b) - \Phi\left(\frac{b}{\sqrt{\eta_{a_1, a_2}^2(\mathbf{\Gamma}, \mathbf{\Gamma}')}}\right) \right| \right] \\
& + E \left[\left| \Phi\left(\frac{b}{\sqrt{\eta_{a_1, a_2}^2(\mathbf{\Gamma}, \mathbf{\Gamma}')}}\right) - \Phi\left(\frac{b}{\sqrt{a_1^2\sigma^2 + a_2^2\sigma^2}}\right) \right| \right] = o(1).
\end{aligned}$$

□

2.6.7 Proof of Corollary 3

By using Theorem 15.2.3 of Lehmann and Romano (2006), the result is a consequence of Theorem 10.

2.6.8 Proof of Theorem 11

(i) By Corollary 3 and Theorem 9

$$\text{Power} = P_{H_{A_c}} \left(\text{ED}_n^k(\mathbf{Z}) > c \right) \rightarrow P(2\varphi(e_{xy}) - \varphi(e_x) - \varphi(e_y) > 0) = 1.$$

(ii) By Corollary 3 and Theorem 9

$$\text{Power} = P_{H_{A_l}} \left(\text{ED}_n^k(\mathbf{Z}) > c \right) = P_{H_{A_l}} \left(\sqrt{nm}p\text{ED}_n^k(\mathbf{Z}) > \sqrt{nm}pc \right) \rightarrow P(N(0, \sigma^2) > \sigma Q_{\Phi, 1-\alpha}) = \alpha.$$

Chapter 3

Change Point Detection for High-dimensional Time Series

3.1 Introduction

Given a sequence of observations, change point detection deals with the problem of testing whether there exists a shift or multiple shifts in certain aspects of their underlying distributions as well as identifying the locations of these shifts. Change point detection methodology can be divided into two categories: online monitoring methods [Dette and Gösmann (2019)] are used when the observations arrive sequentially and aim to detect the change as soon as it occurs; On the other hand, retrospective change point detection methods need to collect all the observations first and then perform the change point analysis retrospectively.

In this work, we want to detect the change point in mean retrospectively, i.e., given a set of data $Y_i \in \mathbb{R}^p$, $i = 1, 2, \dots, n$, with mean $E[Y_i] = \mu_i$, we are interested in testing

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_n \text{ versus } H_A : \mu_1 = \dots = \mu_{k^*} \neq \mu_{k^*+1} = \dots = \mu_n,$$

where k^* denotes the unknown location of the change point. Change point detection is a classical topic and dates back to Page (1954, 1955). There is a huge literature under the low dimensional setting, see [Carlstein et al. (1994), Brodsky and Darkhovsky (2013)] for book-length treatments of the topic. For more recent work, we mention Aue et al. (2009), Shao and Zhang (2010), Matteson and James (2014), Kirch et al. (2015), Zhang and Lavitas (2018) among others.

With the advancement of science and technology, high-dimensional data has become ubiquitous in many fields. Testing the structural break for high dimensional time series has attracted a lot of attention from academic community. Xie et al. (2012) tackled the problem by assuming that the underlying data generating distribution has certain low dimensional structure. Li et al. (2019) proposed a method based on U-statistic and a bias correction term is added to account for the weak temporal dependence, but the weak convergence of the empirical process was not provided. Assuming that the mean change happens in a small subset of coordinates, Wang and Samworth (2018) proposed a two stage method by first finding a good projection direction and then applying the existing low-dimensional method. Jirak et al. (2015) took the maximum of all marginal CUSUM and obtained the critical value from an extreme value distribution of Gumble type or bootstrap. For a sequence of data with both cross-sectional and temporal dependence, Cho et al. (2016) constructed a statistic based on the maximum of cumulative ordered CUSUM.

Assuming that the data points are independent, Wang et al. (2019) proposed a U -statistic based approach for high-dimensional data, as motivated by the pioneering work of Chen et al. (2010). To extend to change point detection problem, they derived the weak convergence of this U -statistic based process under the i.i.d assumption. In this paper, we extend the U -statistic based approach in Wang et al. (2019) to weakly dependent high dimensional time series. To this end, we formulate a trimmed version of the U -statistic that

excludes the pair of data points when the absolute difference of their indices is small, to reduce the bias caused by the temporal dependence. In order to derive the asymptotic distribution of our statistic, we assume that the time series is generated from a high-dimensional linear process and apply the Beveridge-Nelson (BN) decomposition [Phillips and Solo (1992)]. By adopting the fixed- b asymptotic framework [Kiefer and Vogelsang (2005)], i.e., the trimming parameter over sample size is fixed as a constant η between 0 and 1, we show that the U -statistic based process converges weakly to a continuous functional of 4 Gaussian processes. The cross-covariance structure of these four Gaussian processes only depend on the η , so the limiting null distribution is pivotal for a given η . We also derive the asymptotic power under local alternatives.

Our method requires mild structural assumptions on the data generating process and involves only one trimming parameter, whose impact is captured to the first order by the limiting null distribution. In addition, we combine our statistic with the wild binary segmentation (WBS) procedure [Fryzlewicz et al. (2014)] to detect multiple change points, since WBS is shown to be advantageous than BS in the presence of multiple non-monotone shifts. Empirical simulations show that our trimmed statistic has advantages over the double CUSUM statistic in Cho et al. (2016) when the dependence within time series is strong and suggests that trimming is necessary and effective to reduce bias. On the other hand, too much trimming can harm the estimation accuracy of the change point locations.

The remainder of the paper is organized as follows. Section 2 provides a detailed review of the U -statistic based method in Wang et al. (2019) and the extension to high-dimensional time series. In Section 3, we present the limiting distributions of our self-normalized statistic under the null and alternatives. Section 4 contains all the simulation results. Section 5 concludes. All technical proofs are included in Section 6.

3.2 Change Point Detection for High-dimensional Time Series

3.2.1 Review of Wang et al. (2019)

Assuming that the observations are $Y_i = \mu_i + X_i$, $i = 1, 2, \dots, n$ where X_1, X_2, \dots, X_n are i.i.d copies of a \mathbb{R}^p -valued random vector X_0 such that $E[X_0] = 0$ and $E[X_0 X_0^T] = \Sigma$. For the sub-sample $X_l, X_{l+1}, \dots, X_k, \dots, X_{l+m}$ with the potential change point k , Wang et al. (2019) propose to use the following U -statistic

$$D(k; l, m) := \sum_{l \leq j_1, j_3 \leq k} \sum_{k+1 \leq j_2, j_4 \leq m}^{|j_1 - j_3| > 0} \sum_{|j_2 - j_4| > 0} (Y_{j_1} - Y_{j_2})^T (Y_{j_3} - Y_{j_4}), \quad (3.1)$$

where $l + 1 \leq k$ and $k + 2 \leq m$. In addition, define the normalization factor as

$$W_n(k; l, m) := \frac{1}{n} \sum_{t=l+1}^{k-2} D^2(t; l, k) + \frac{1}{n} \sum_{t=k+2}^{m-2} D^2(t; k+1, m),$$

where $l + 1 \leq k - 2$ and $k + 2 \leq m - 2$. Then, the self-normalized statistic is defined as

$$T_n := \sup_{k=4, \dots, n-4} \frac{D^2(k; 1, n)}{W_n(k; 1, n)}$$

Under null, $D(k; l, m)$ is a function of statistics with the following form

$$\tilde{S}_n(k, m) = \sum_{i=k}^{m-1} \sum_{j=k}^i X_{i+1}^T X_j.$$

To see this, some algebra shows that

$$\begin{aligned} D(k; l, m) &:= 2(m-k)(m-k-1)\tilde{S}_n(l, k) + 2(k-l)(k-l+1)\tilde{S}_n(k+1, m) \\ &\quad - 2(k-l)(m-k-1)(\tilde{S}_n(l, m) - \tilde{S}_n(l, k) - \tilde{S}_n(k+1, m)) \end{aligned}$$

Wang et al. (2019) derive the weak convergence of the two parameter process $\{\tilde{S}_n(\lfloor an \rfloor, \lfloor bn \rfloor)\}_{(a,b) \in [0,1]^2}$ in $l^\infty([0,1]^2)$ and show that limiting distribution of T_n is pivotal.

3.2.2 Trimmed U -statistic

Now, suppose $\{Y_t\}$ is a realization of \mathbb{R}^p valued time series, we consider the following trimmed statistic instead

$$D(k; l, m|\tau) = \sum_{\substack{|j_1-j_3|>\tau \\ l \leq j_1, j_3 \leq k}} \sum_{\substack{|j_2-j_4|>\tau \\ k+\tau+1 \leq j_2, j_4 \leq m}} (Y_{j_1} - Y_{j_2})^T (Y_{j_3} - Y_{j_4}),$$

where τ is a fixed constant such that $l + \tau + 1 \leq k \leq m - 2\tau - 2$. It is clear that when $\tau = 0$, $D(k; l, m|0) = D(k; l, m)$, where $D(k; l, m)$ is defined in Equation (3.1). Letting

$$W_n(k; l, m|\tau) := \frac{1}{n} \sum_{t=l+\tau+1}^{k-2\tau-2} D^2(t; l, k|\tau) + \frac{1}{n} \sum_{t=k+\tau+2}^{m-2\tau-2} D^2(t; k+1, m|\tau),$$

where $l + \tau + 1 \leq k - 2\tau - 2$ and $k + \tau + 2 \leq m - 2\tau - 2$. The self-normalized statistic is defined as

$$T_n := \sup_{k=3\tau+4, \dots, n-3\tau-4} \frac{D^2(k; 1, n|\tau)}{W_n(k; 1, n|\tau)}.$$

To derive the limiting distribution of T_n , the key observation is that the 3 parameter process $D(k; l, m|\tau)$ can still be written as a continuous functional of 4 two parameter processes. It is due to the trimming that we have 4 processes and the functional is more complex than the one in Wang et al. (2019).

3.2.3 Wild Binary Segmentation

To extend our method to multiple change point detection, we combine our statistics with the wild binary segmentation procedure. Firstly, let $F_n^M = \{(s_m, e_m)\}_{m=1}^M$, where s_m, e_m are sampled uniformly from $\{1, 2, \dots, n\}$ such that $e_m - s_m \geq 6\tau + 7 + \lfloor 0.15n \rfloor$, where $\lfloor 0.15n \rfloor$ is the additional subsample size. Secondly, given F_n^M , the threshold ς_n is chosen as follows: for $r = 1, 2, \dots, R$, generate the n i.i.d multivariate normal random variables, i.e., $\{X_{r,i}\}_{i=1}^n$ and $X_{r,1}, X_{r,2}, \dots, X_{r,n} \stackrel{i.i.d}{\sim} N(0, I_p)$. Based on the r th sample $\{X_{r,i}\}_{i=1}^n$, calculate

$$\hat{\varsigma}_{n,r} = \max_{m \in \mathcal{M}_{s,e}, b \in \{s_m+3\tau+3, \dots, e_m-3\tau-4\}} \frac{D^2(b; s_m, e_m|\tau)}{W_n(b; s_m, e_m|\tau)}$$

The threshold ς_n is set to be the 95% quantile of $\{\hat{\varsigma}_{n,1}, \hat{\varsigma}_{n,2}, \dots, \hat{\varsigma}_{n,R}\}$. Then, to detect multiple changes in Y_1, Y_2, \dots, Y_n , apply the function below as $\text{WBS}(1, n, \varsigma_n)$.

Function $\text{WBS}(s, e, \varsigma_n)$:

```

if  $e - s < 6\tau + 7$  then
  | STOP
else
  |  $\mathcal{M}_{s,e} :=$  set of those indices  $m$  for which  $[s_m, e_m] \in F_n^M$  is such that  $[s_m, e_m] \subseteq [s, e]$ 
  |  $(m_0, b_0) := \arg \max_{m \in \mathcal{M}_{s,e}, b \in \{s_m + 3\tau + 3, \dots, e_m - 3\tau - 4\}} \frac{D^2(b; s_m, e_m | \tau)}{W_n(b; s_m, e_m | \tau)}$ 
  | if  $\frac{D^2(b_0; s_{m_0}, e_{m_0} | \tau)}{W_n(b_0; s_{m_0}, e_{m_0} | \tau)} > \varsigma_n$  then
  |   | add  $b_0$  to the set of estimated change points
  |   |  $\text{WBS}(s, b_0, \varsigma_n)$ 
  |   |  $\text{WBS}(b_0 + 1, e, \varsigma_n)$ 
  | else
  |   | STOP
  | end
end

```

Algorithm 1: WBS

3.3 Asymptotic Theory

For any matrix $A = (a_{i,j})_{i=1, \dots, n; j=1, \dots, m} \in \mathbb{R}^{n \times m}$, its L_1 norm is denoted as $\|A\|_1 := \max_j \sum_{i=1}^n |a_{i,j}|$ and L_∞ norm denoted as $\|A\|_\infty := \max_i \sum_{j=1}^m |a_{i,j}|$. The joint cumulant of n random variables Z_1, \dots, Z_n is denoted as $\text{cum}(Z_1, Z_2, \dots, Z_n)$. Throughout, we assume $\tau = \lfloor \eta n \rfloor, \eta \in (0, 1)$ and fix η in our asymptotics. The following assumption is used to study the asymptotic distribution of T_n .

Assumption 48. For $i = 1, 2, \dots, n$, the observations are $Y_i = \mu_i + X_i$, where $X_i = \sum_{j=0}^\infty c_j \epsilon_{i-j}$ and ϵ_i are i.i.d p -dimensional innovations with mean 0 and c_j are $p \times p$ coefficient matrices. Let $\Gamma = (\sum_{u=0}^\infty c_u) \text{cov}(\epsilon_0) (\sum_{u=0}^\infty c_u)^T$ and suppose

A.1 $\sup_{l=1, \dots, p} \|\epsilon_{0,l}\|_8 < \infty$.

A.2 For any $m \geq 0$,

$$\sum_{u=m}^\infty \|c_u\|_1 \leq C\rho^m \text{ and } \sum_{u=m}^\infty \|c_u\|_\infty \leq C\rho^m,$$

where $C > 0$ and $0 < \rho < 1$ are some constants.

A.3 $\text{tr}(\Gamma^4) = o(\|\Gamma\|_F^4)$.

A.4 $p^6 \rho^{\lfloor \eta n \rfloor} / \|\Gamma\|_F^6 = O(1)$.

A.5 For any $h = 2, 3, 4, 5, 6$,

$$\sum_{k_1, \dots, k_h=1}^p |\text{cum}(\epsilon_{0,k_1}, \dots, \epsilon_{0,k_h})| \leq C' \|\Gamma\|_F^h,$$

where C' is some constant independent of n, p .

Remark 49. Assumptions A.1 and A.2 imply the Uniform Geometric Moment Contraction (UGMC(8)) property in Wang and Shao (2019). The UGMC condition is the generalization of Geometric Moment Contraction property in Hsing et al. (2004) and Wu and Shao (2004) and its equivalent form has been used in Zhang and Cheng (2018). Assumption A.3 is satisfied under some weak cross-sectional dependence conditions, see Wang et al. (2019) for some discussions. Assumption A.4 implies that the bias caused by temporal dependence is asymptotically negligible. Assumption A.5 holds under mild conditions, see Wang and Shao (2019) for some verified examples.

Then, we have the following theorem states the asymptotic distribution of T_n .

Theorem 12. Suppose Assumption 48 is true. Then,

$$T_n \xrightarrow{\mathcal{D}} T := \sup_{r \in (2\eta, 1-2\eta)} \frac{G^2(r; 0, 1|\eta)}{\int_{\eta}^{r-\eta} G^2(u; 0, r|\eta) du + \int_{r+\eta}^{1-\eta} G^2(u; r, 1|\eta) du},$$

where

$$\begin{aligned} G(r; a, b|\eta) : &= 2(b-r-2\eta)^2 V_1(a, r|\eta) + 2(r-a-\eta)^2 V_1(r+\eta, b|\eta) \\ &\quad - (r-\eta)(b-\eta) U_1(a, r-\eta; r+\eta, b-\eta) \\ &\quad + (r-\eta)(r+2\eta) U_1(a, r-\eta; r+2\eta, b) \\ &\quad + (a+\eta)(b-\eta) U_1(a+\eta, r; r+\eta, b-\eta) \\ &\quad - (a+\eta)(r+2\eta) U_1(a+\eta, r; r+2\eta, b) \\ &\quad - (b-\eta) U_2(a, r-\eta; r+\eta, b-\eta) \\ &\quad + (r+2\eta) U_2(a, r-\eta; r+2\eta, b) \\ &\quad + (b-\eta) U_2(a+\eta, r; r+\eta, b-\eta) \\ &\quad - (r+2\eta) U_2(a+\eta, r; r+2\eta, b) \\ &\quad - (r-\eta) U_3(a, r-\eta; r+\eta, b-\eta) \\ &\quad + (r-\eta) U_3(a, r-\eta; r+2\eta, b) \\ &\quad + (a+\eta) U_3(a+\eta, r; r+\eta, b-\eta) \\ &\quad - (a+\eta) U_3(a+\eta, r; r+2\eta, b) \\ &\quad + U_4(a, r-\eta; r+\eta, b-\eta) \\ &\quad - U_4(a, r-\eta; r+2\eta, b) \\ &\quad - U_4(a+\eta, r; r+\eta, b-\eta) \\ &\quad + U_4(a+\eta, r; r+2\eta, b); \end{aligned}$$

For $u, v = 1, 2, 3, 4$,

$$U_u(a_1, a_2; b_1, b_2) = V_u(a_1, b_2|\eta) - V_u(a_1, b_1|\eta) - V_u(a_2, b_2|\eta) + V_u(a_2, b_1|\eta)$$

and V_1, V_2, V_3, V_4 are Gaussian processes with covariance structures

$$\text{cov}(V_u(a_1, b_1|\eta), V_v(a_2, b_2|\eta)) = C_{u,v}(a_1 \vee a_2, b_1 \wedge b_2) \mathbb{I}_{\{b_1 \wedge b_2 - a_1 \vee a_2 - \eta > 0\}}$$

η	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$
0.010	702.018	1034.285	1419.761	2352.446	2988.825
0.020	921.366	1381.689	1882.719	3288.334	3921.601
0.030	1237.174	1882.130	2595.468	4528.448	5519.012
0.040	1721.446	2610.516	3688.366	6373.774	7978.322
0.050	2473.703	3849.287	5260.391	9568.212	11650.292
0.060	3718.168	5657.260	7933.067	14913.855	17543.712
0.070	5521.788	8537.631	11976.984	21751.883	27166.972
0.080	9640.496	14908.465	21325.384	38889.673	49069.849
0.09	17392.21	27479.22	38437.33	72032.59	91588.16
0.10	32394.61	50616.69	71142.16	131662.45	166026.23

Table 3.1: Simulated critical values of T_n .

where let $w_{i,j}^u = \mathbb{I}_{\{u=1\}} + \frac{j}{n}\mathbb{I}_{\{u=2\}} + \frac{i+\lfloor \eta n \rfloor + 1}{n}\mathbb{I}_{\{u=3\}} + \frac{i+\lfloor \eta n \rfloor + 1}{n}\frac{j}{n}\mathbb{I}_{\{u=4\}}$, $C_{u,v}(a, b)$ is defined as

$$C_{u,v}(a, b) = \lim_{n \rightarrow \infty} \frac{2}{n^2} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor an \rfloor}^i w_{i,j}^u w_{i,j}^v.$$

The critical values of T_n are simulated in Table 3.1. For any real-valued vector $\delta = (\delta_1, \delta_2, \dots, \delta_p)^T \in \mathbb{R}^p$, its L^1 -norm and L^2 -norm are denoted as $\|\delta\|_1 := \sum_{i=1}^p |\delta_i|$ and $\|\delta\|_2 := (\sum_{i=2}^p \delta_i^2)^{1/2}$. We then present the asymptotic distribution under some local alternatives.

Theorem 13. *Suppose Assumption 48 hold and $\sqrt{n} \log_2(n) \|\delta_n\|_1 = o(n \|\delta_n\|_2^2)$. Assume that there exists $\phi \in (0, 1)$ such that $\mu_i = \mu^*$ for $i = 1, 2, \dots, \lfloor \phi n \rfloor$ and $\mu_i = \mu^* + \delta_n$ for $i = \lfloor \phi n \rfloor + 1, \dots, n$. Then,*

1, *If $n \|\delta_n\|_2^2 / \|\Gamma\|_F \rightarrow \infty$, then $T_n \rightarrow \infty$ in probability.*

2, *If $n \|\delta_n\|_2^2 / \|\Gamma\|_F \rightarrow 0$, then $T_n \rightarrow T$.*

3, *If $n \|\delta_n\|_2^2 / \|\Gamma\|_F \rightarrow c \in (0, \infty)$, then*

$$T_n \xrightarrow{\mathcal{D}} \sup_{r \in (0, 1)} \frac{\tilde{G}^2(r; 0, 1 | \eta, \phi)}{\int_0^r \tilde{G}^2(u; 0, r | \eta, \phi) du + \int_r^1 \tilde{G}^2(u; r, 1 | \eta, \phi) du},$$

where

$$\tilde{G}(r; a, b | \eta, \phi) := G(r; a, b | \eta) + c \diamond(r; a, b | \eta, \phi),$$

and $\diamond(r; a, b|\eta, \phi)$ is a constant defined as

$$\begin{aligned}
\diamond(r; a, b|\eta, \phi) = & 2(b-r-2\eta)^2 \Delta_1(a, r|\eta, \phi) + 2(r-a-\eta)^2 \Delta_1(r+\eta, b|\eta, \phi) \\
& - (r-\eta)(b-\eta) \square_1(a, r-\eta; r+\eta, b-\eta) \\
& + (r-\eta)(r+2\eta) \square_1(a, r-\eta; r+2\eta, b) \\
& + (a+\eta)(b-\eta) \square_1(a+\eta, r; r+\eta, b-\eta) \\
& - (a+\eta)(r+2\eta) \square_1(a+\eta, r; r+2\eta, b) \\
& - (b-\eta) \square_2(a, r-\eta; r+\eta, b-\eta) \\
& + (r+2\eta) \square_2(a, r-\eta; r+2\eta, b) \\
& + (b-\eta) \square_2(a+\eta, r; r+\eta, b-\eta) \\
& - (r+2\eta) \square_2(a+\eta, r; r+2\eta, b) \\
& - (r-\eta) \square_3(a, r-\eta; r+\eta, b-\eta) \\
& + (r-\eta) \square_3(a, r-\eta; r+2\eta, b) \\
& + (a+\eta) \square_3(a+\eta, r; r+\eta, b-\eta) \\
& - (a+\eta) \square_3(a+\eta, r; r+2\eta, b) \\
& + \square_4(a, r-\eta; r+\eta, b-\eta) \\
& - \square_4(a, r-\eta; r+2\eta, b) \\
& - \square_4(a+\eta, r; r+\eta, b-\eta) \\
& + \square_4(a+\eta, r; r+2\eta, b).
\end{aligned}$$

For $u = 1, 2, 3, 4$,

$$\square_u(a_1, a_2; b_1, b_2) = \Delta_u(a_1, b_2|\eta, \phi) - \Delta_u(a_1, b_1|\eta, \phi) - \Delta_u(a_2, b_2|\eta, \phi) + \Delta_u(a_2, b_1|\eta, \phi).$$

and

$$\Delta_u(a, b|\eta, \phi) = \begin{cases} \lim_{n \rightarrow \infty} \frac{\sqrt{2}}{n^2} \sum_{i=\lfloor \phi n \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor \phi n \rfloor}^i w_{i,j}^u, & \text{if } a < \phi < b - \eta, \\ 0, & \text{otherwise.} \end{cases}$$

Remark 50. If $\eta = 0$, direct calculations show that

$$\begin{aligned}
G(r; a, b|0) = & 2(b-r)^2 V_1(a, r|0) + 2(r-a)^2 V_1(r, b|0) \\
& - 2(r-a)(b-r)(V_1(a, b|0) - V_1(a, r|0) - V_1(r, b|0)),
\end{aligned}$$

which is identical to $G(r; a, b)$ in Wang et al. (2019). In addition, the constant term $\diamond(r; a, b|\eta, \phi)$ can be seen to be equal to $\sqrt{2}\Delta(r, a, b, \cdot)$ in Wang et al. (2019).

3.4 Simulation Study

3.4.1 Size and Power for Change Point Testing

We first consider the following single change point model.

Example 11. Consider the following $VAR(1)$ model,

$$X_i - \mu \mathbb{I}_{\{i > 0.5n\}} = \rho(X_{i-1} - \mu \mathbb{I}_{\{i > 0.5n\}}) + \epsilon_i,$$

where $\mu = E[X_i]$, $\{\epsilon_i\}$ are the random errors and we consider $\rho \in \{0.2, 0.5, 0.8, -0.5\}$. Under the null hypothesis, $\mu = 0$. Under the alternative hypothesis, we examine the following two types of mean shift, i.e.,

(i) Homogeneous alternative: $\mu^T = 0.1(1, 1, \dots, 1)$.

(ii) Inhomogeneous alternative:

$$\mu^T = 0.1(\mu_1, \dots, \mu_p), \text{ where } (\mu_1, \dots, \mu_p) \stackrel{i.i.d}{\sim} \text{Uniform}(0, 1).$$

Also, for the random errors $\{\epsilon_i\}$, we consider the following three scenarios

(a) Gaussian errors with $AR(1)$ type covariance structure: $\epsilon_i \stackrel{i.i.d}{\sim} N(0, \Sigma_\epsilon)$, where $\Sigma_\epsilon = (0.5^{|i-j|})_{i,j=1}^p$.

(b) Non-Gaussian errors: $\{\epsilon_i\}_{i=1}^n$ are i.i.d and each entry of $\epsilon_i = (\epsilon_{i,1}, \dots, \epsilon_{i,p})^T$ is generated independently and uniformly from $[-1, 1]$, i.e., $(\epsilon_{i,1}, \dots, \epsilon_{i,p}) \stackrel{i.i.d}{\sim} \text{Uniform}(-1, 1)$.

The simulation results for Example 11 with dimension $p = 2n$ are presented in Table 3.2. Overall, our method demonstrates satisfactory size and power for scenarios (a) and (b). This is as expected, as both data generating procedures satisfy Assumption 48. For $\rho = 0.2, 0.5, -0.5$, by setting η to be 0.02, we can have accurate size if $n = 400$ and good testing power if $n = 800$. The power is generally higher for non-Gaussian errors, this might due to the cross-sectional independence of the non-Gaussian errors. For $\rho = 0.8$, we observe some size distortions when n is small. This phenomenon is not surprising, as larger ρ would result bigger bias and in return, it can be seen from Assumption A.4 that we need either larger n or larger η to control the bias.

3.4.2 Change Point Estimation

For multiple change point detection and estimation, we compare algorithm 1 with the double CUSUM binary segmentation algorithm (DCBS) [Cho et al. (2016)].

Example 12. Consider the model,

$$X_i = \mu_i + \epsilon_i,$$

where $\epsilon_i = (\epsilon_{i,1}, \epsilon_{i,2}, \dots, \epsilon_{i,p})^T$ is generated from the following two methods.

(i) Gaussian errors with $AR(1)$ type covariance structure: set $\Sigma_u = (0.5^{|i-j|})_{i,j=1}^p$, for $i = 1, 2, \dots, n$, let $u_i \stackrel{i.i.d}{\sim} N(0, \Sigma_u)$ and

$$\epsilon_i = \rho \epsilon_{i-1} + u_i.$$

Table 3.2: Simulation results for Example 11 with $p = 2n$ and significance level 0.05.

ρ	n	η	Null		(i)		(ii)	
			(a)	(b)	(a)	(b)	(a)	(b)
0.2	200	0.02	0.047	0.051	0.700	1	0.199	0.830
	200	0.05	0.056	0.060	0.712	1	0.207	0.831
	400	0.02	0.049	0.049	0.998	1	0.679	1
	400	0.05	0.052	0.050	0.997	1	0.682	1
	800	0.02	0.053	0.046	1	1	0.996	1
	800	0.05	0.052	0.053	1	1	0.995	1
0.5	200	0.02	0.060	0.052	0.267	0.821	0.110	0.290
	200	0.05	0.070	0.074	0.303	0.915	0.104	0.372
	400	0.02	0.051	0.055	0.759	1	0.244	0.878
	400	0.05	0.059	0.055	0.767	1	0.243	0.881
	800	0.02	0.053	0.048	0.999	1	0.727	1
	800	0.05	0.053	0.056	0.999	1	0.736	1
0.8	200	0.02	0	0	0	0	0	0
	200	0.05	0.073	0.040	0.096	0.093	0.080	0.053
	400	0.02	0	0	0.001	0.0004	0.0004	0
	400	0.05	0.086	0.081	0.172	0.557	0.104	0.194
	800	0.02	0.052	0.046	0.305	0.904	0.111	0.319
	800	0.05	0.064	0.065	0.370	0.975	0.112	0.493
-0.5	200	0.02	0.007	0.005	0.986	1	0.314	0.994
	200	0.05	0.016	0.022	0.998	1	0.683	1
	400	0.02	0.038	0.032	1	1	0.999	1
	400	0.05	0.030	0.031	1	1	0.999	1
	800	0.02	0.047	0.044	1	1	1	1
	800	0.05	0.043	0.044	1	1	1	1

(ii) Non-Gaussian errors: for $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$

$$u_{i,j} \stackrel{i.i.d}{\sim} \text{Uniform}(-1.6, 1.6),$$

$$\epsilon_i = \rho \epsilon_{i-1} + u_i.$$

(iii) Motivated by Cho et al. (2016), let $\varrho_k = 0.6(k+1)^{-1}$, for $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$

$$u_{i,j} = \sum_{k=0}^{99} \varrho_k v_{i,j-k}, \text{ where } v_{i,j} \stackrel{i.i.d}{\sim} N(0, 1),$$

$$\epsilon_{i,j} = \rho \epsilon_{i-1,j} - 0.1 \epsilon_{i-2,j} + u_{i,j} + 0.2 u_{i-1,j}.$$

Then, three change points and some sparse structure are considered

$$\mu_i = \delta_1 \mathbb{I}_{i > k_1} + \delta_2 \mathbb{I}_{i > k_2} + \delta_3 \mathbb{I}_{i > k_3},$$

where for $r = 1, 2, 3$, $\delta_r = (\delta_{r,1}, \delta_{r,1}, \dots, \delta_{r,p})^T \in \mathbb{R}^p$. Denote $\Pi_r = \{j | |\delta_{r,j}| > 0, j = 1, 2, \dots, p\}$ and $\delta_{r,j} \stackrel{i.i.d}{\sim} \text{Uniform}(0.75\theta_r, 1.25\theta_r)$ for $j \in \Pi_r$. Here, we set $(k_1, |\Pi_1|, \theta_1) = ([0.3n], [0.75p], 0.4)$, $(k_2, |\Pi_2|, \theta_2) = ([0.6n], [0.25p], 0.696)$ and $(k_3, |\Pi_3|, \theta_3) = ([0.8n], [0.1p], 1.12)$.

Table 3.3: Simulation results for Example 12 with $n = 250$ and significance level 0.05. WBS¹ and WBS² are with respect to trimming $\eta = 0.01$ and $\eta = 0.02$ respectively.

p	ρ		# of change points (%)						ARI	accuracy (%)		
			0	1	2	3	4	5		k_1	k_2	k_3
(i)	250	WBS ¹	0	0	0	99	1	0	0.929	96	95	95
		WBS ²	0	0	0	100	0	0	0.901	85	77	86
		DCBS	0	0	0	100	0	0	0.996	100	100	100
		WBS ¹	0	0	5	88	6	1	0.902	91	86	85
		WBS ²	0	0	1	98	1	0	0.882	76	70	79
		DCBS	0	13	56	31	0	0	0.813	84	97	35
	0.6	WBS ¹	89	11	0	0	0	0	0.041	5	3	2
		WBS ²	1	2	27	70	0	0	0.784	58	55	55
		DCBS	53	46	1	0	0	0	0.222	2	40	1
	500	WBS ¹	0	0	0	96	4	0	0.930	93	96	97
		WBS ²	0	0	0	100	0	0	0.905	87	84	77
		DCBS	0	0	0	100	0	0	1.000	100	100	100
		WBS ¹	0	0	0	97	3	0	0.924	95	92	97
		WBS ²	0	0	0	100	0	0	0.896	85	84	65
		DCBS	0	22	43	35	0	0	0.804	78	100	35
	0.6	WBS ¹	100	0	0	0	0	0	0	0	0	0
		WBS ²	0	0	8	90	2	0	0.861	69	74	67
		DCBS	46	54	0	0	0	0	0.260	2	51	0
(ii)	250	WBS ¹	0	0	0	98	2	0	0.932	97	88	99
		WBS ²	0	0	0	100	0	0	0.904	85	79	89
		DCBS	0	0	0	100	0	0	0.993	100	100	100
		WBS ¹	0	1	23	74	2	0	0.860	86	82	75
		WBS ²	0	0	1	99	0	0	0.893	77	72	82
		DCBS	0	26	33	41	0	0	0.784	73	95	42
	0.6	WBS ¹	100	0	0	0	0	0	0	0	0	0
		WBS ²	0	8	41	51	0	0	0.732	53	53	41
		DCBS	90	10	0	0	0	0	0.046	1	5	0
	500	WBS ¹	0	0	0	99	1	0	0.936	96	95	100
		WBS ²	0	0	0	100	0	0	0.910	86	84	88
		DCBS	0	0	0	100	0	0	1.000	100	100	100
		WBS ¹	0	0	8	89	3	0	0.909	90	92	91
		WBS ²	0	0	0	100	0	0	0.906	84	81	84
		DCBS	0	67	18	15	0	0	0.621	33	100	15
	0.6	WBS ¹	100	0	0	0	0	0	0	0	0	0
		WBS ²	0	3	29	68	0	0	0.798	64	53	57
		DCBS	95	5	0	0	0	0	0.022	0	4	0
(iii)	250	WBS ¹	0	0	0	100	0	0	0.919	91	80	92
		WBS ²	0	0	0	99	1	0	0.895	84	73	84
		DCBS	0	0	0	100	0	0	0.999	100	100	100
		WBS ¹	0	0	2	96	2	0	0.905	86	79	81
		WBS ²	0	0	0	100	0	0	0.885	80	70	74
		DCBS	0	1	20	79	0	0	0.952	98	100	79
	0.6	WBS ¹	0	6	25	62	7	0	0.801	73	68	60
		WBS ²	0	4	23	73	0	0	0.789	60	57	49
		DCBS	7	86	7	0	0	0	0.465	7	86	1
	500	WBS ¹	0	0	0	99	1	0	0.922	87	89	94
		WBS ²	0	0	0	100	0	0	0.899	84	82	78
		DCBS	0	0	0	100	0	0	1	100	100	100
		WBS ¹	0	0	0	98	2	0	0.915	83	86	91
		WBS ²	0	0	0	100	0	0	0.886	76	71	75
		DCBS	0	0	13	87	0	0	0.975	100	100	87
	0.6	WBS ¹	0	0	10	78	12	0	0.876	82	79	80
		WBS ²	0	0	4	96	0	0	0.859	65	59	65
		DCBS	3	95	2	0	0	0	0.478	2	97	0

The simulation result based 100 sample paths is presented in Table 3.3, where we report the estimated # of change points in %, average of Adjusted Rand index (ARI) and the location estimation accuracy in $(|\hat{k}_i - k_i| < \log(n), \text{in } \%)$. It can be seen that for the cases of $\rho = 0.4$, WBS¹ and DCBS does not work well in detecting the three change points. Further more, the performance of WBS¹ and DCBS is terrible when $\rho = 0.6$, while the averaged ARI of WBS² is above 0.73 for all cases. This suggests the necessity of trimming in the presence of strong temporal dependence. When $\rho = 0.2$, all three methods perform very well and DCBS has the highest ARI. In addition, the simulation results suggest that setting trimming parameter to be $\eta = 0.01$ is sufficient to control the bias when $\rho = 0.2$, under which larger trimming can decrease the accuracy for estimating the change point locations. Ideally, we would want to minimize the trimming parameter within the range that the bias is well controlled.

3.5 Conclusion

In this work, we consider the change point detection problem for high dimensional time series. Our statistic is motivated by the U -statistic for high dimensional two sample mean testing [Chen et al. (2010)] and extends the U -statistic based method used in Wang et al. (2019) to high-dimensional time series setting. To account for temporal dependence, we replace the U -statistic with its trimmed version as a way to reduce bias. As an important theoretical contribution, we show that the trimmed U -statistic based process converges weakly to a continuous functional of 4 Gaussian processes under the fixed- b regime for high-dimensional linear processes. Our test statistic is pivotal due to the use of self-normalization [Shao and Zhang (2010)] and its limiting quantiles can be estimated easily by simulations. Finally, empirical studies show that trimming is effective and necessary when the temporal dependence between data points is strong.

The current work can be generalized in many directions. Firstly, it is interesting to extend the asymptotic theory to a more general setting, such as nonlinear causal process Wu (2005b); see Wang and Shao (2019) for recent extension of SN to high-dimensional time series under the framework of nonlinear causal process. Secondly, while we consider a shift in mean in this paper, it is also of great value to do change point detection for other aspects of the underlying high dimensional distribution, such quantiles; see Shao and Zhang (2010) for some discussions under the low dimensional setting Thirdly, selecting the trimming parameter for real applications can be nontrivial and it would be interesting to develop a data-driven procedure to be adaptive to the magnitude of temporal dependence. We leave these topics for future investigations.

3.6 Technical Details

For the technical details, we define the summation $\sum_{i=m_1}^{m_2} *$ to be 0 if $m_1 > m_2$. In addition, given data set $\{X_i\}_{i=1}^n$, we use the convention that $X_j = 0$ if $j = 0$ or $j > n$.

3.6.1 Properties of Linear Process

Firstly, applying Beveridge Nelson (BN) decomposition in Phillips and Solo (1992), we have

$$X_i = D_i - \varepsilon_i,$$

where $D_i = (\sum_{u=0}^{\infty} c_u)\epsilon_i$, $\tilde{D}_i = \sum_{j=0}^{\infty} (\sum_{u=j+1}^{\infty} c_u)\epsilon_{i-j}$ and $\varepsilon_i = \tilde{D}_i - \tilde{D}_{i-1}$. We then state three useful auxiliary lemmas.

Lemma 51. Suppose Assumption 48 (A.1, A.2, A.5) is true. Then, for any $h = 2, 3, 4, 5, 6$ and $j = 0, 1, \dots, h$, we have

$$\begin{aligned} \sum_{l_1, l_2, \dots, l_h=1}^p |\text{cum}(D_{i_1, l_1}, \dots, D_{i_j, l_j}, \tilde{D}_{i_{j+1}, l_{j+1}}, \dots, \tilde{D}_{i_h, l_h})| &\lesssim \|\Gamma\|_F^h, \\ \sum_{l_1, l_2, \dots, l_h=1}^p \text{cum}^2(D_{i_1, l_1}, \dots, D_{i_j, l_j}, \tilde{D}_{i_{j+1}, l_{j+1}}, \dots, \tilde{D}_{i_h, l_h}) &\lesssim \|\Gamma\|_F^h. \end{aligned}$$

Lemma 52. Suppose Assumption 48 (A.1, A.2) is true. Then, for some constant C and $0 < \rho < 1$, we have

$$|\text{cum}(Z_{i_0}, \dots, Z_{i_k})| \leq C\rho^{i_{\max} - i_{\min}},$$

where $i_{\max} = \max\{i_0, \dots, i_k\}$, $i_{\min} = \min\{i_0, \dots, i_k\}$ and for each $i \in \{i_0, \dots, i_k\}$, Z_i can be any element from the set $\{X_{i,j}, D_{i,j}, \tilde{D}_{i,j}, \varepsilon_{i,j}\}_{i=i_0, \dots, i_k, j=1, \dots, p}$.

Lemma 53. Under Assumption 48, for any $i \neq j$, we have

$$E[(D_i^T D_j)^6] \lesssim \|\Gamma\|_F^6.$$

3.6.2 Proof of Theorem 12

Recall that

$$\tilde{H}_{X_n}(k; l, m|\tau) = \sum_{l \leq j_1, j_3 \leq k}^{|j_1 - j_3| > \tau} \sum_{k + \tau + 1 \leq j_2, j_4 \leq m}^{|j_2 - j_4| > \tau} (X_{j_1} - X_{j_2})^T (X_{j_3} - X_{j_4}).$$

For $u = 1, 2, 3, 4$, define

$$\tilde{S}_{X_n}^u(k, m|\tau) = \begin{cases} \sum_{i=k}^{m-\tau-1} \sum_{j=k}^i \tilde{w}_{i,j}^u X_{i+\tau+1}^T X_j, & m - \tau - 1 \geq k \\ 0, & \text{otherwise} \end{cases} \quad (3.2)$$

where

$$\tilde{w}_{i,j}^u = \mathbb{I}_{\{u=1\}} + \frac{j}{n} \mathbb{I}_{\{u=2\}} + \frac{i + \tau + 1}{n} \mathbb{I}_{\{u=3\}} + \frac{i + \tau + 1}{n} \frac{j}{n} \mathbb{I}_{\{u=4\}}.$$

Next, we write $\tilde{H}_{X_n}(k; l, m|\tau)$ as a function of $\tilde{S}_{X_n}^u(k, m|\tau)$.

$$\begin{aligned}
\frac{\tilde{H}_{X_n}(k; l, m|\tau)}{n^2} = & 2 \frac{(m-2\tau-k-1)(m-2\tau-k)}{n^2} \tilde{S}_{X_n}^1(l, k|\tau) \\
& + 2 \frac{(k-\tau-l)(k-\tau-l+1)}{n^2} \tilde{S}_{X_n}^1(k+\tau+1, m|\tau) \\
& - \frac{(k-\tau)(m-\tau)}{n^2} Q_{X_n}^1(l, k-\tau-1; k+\tau+1, m-\tau-1) \\
& + \frac{(k-\tau)(k+2\tau+1)}{n^2} Q_{X_n}^1(l, k-\tau-1; k+2\tau+2, m) \\
& + \frac{(l+\tau)(m-\tau)}{n^2} Q_{X_n}^1(l+\tau+1, k; k+\tau+1, m-\tau-1) \\
& - \frac{(l+\tau)(k+2\tau+1)}{n^2} Q_{X_n}^1(l+\tau+1, k; k+2\tau+2, m) \\
& + \frac{(m-\tau)}{n} Q_{X_n}^2(l, k-\tau-1; k+\tau+1, m-\tau-1) \\
& - \frac{(k+2\tau+1)}{n} Q_{X_n}^2(l, k-\tau-1; k+2\tau+2, m) \\
& - \frac{(m-\tau)}{n} Q_{X_n}^2(l+\tau+1, k; k+\tau+1, m-\tau-1) \\
& + \frac{(k+2\tau+1)}{n} Q_{X_n}^2(l+\tau+1, k; k+2\tau+2, m) \\
& + \frac{(k-\tau)}{n} Q_{X_n}^3(l, k-\tau-1; k+\tau+1, m-\tau-1) \\
& - \frac{(k-\tau)}{n} Q_{X_n}^3(l, k-\tau-1; k+2\tau+2, m) \\
& - \frac{(l+\tau)}{n} Q_{X_n}^3(l+\tau+1, k; k+\tau+1, m-\tau-1) \\
& + \frac{(l+\tau)}{n} Q_{X_n}^3(l+\tau+1, k; k+2\tau+2, m) \\
& - Q_{X_n}^4(l, k-\tau-1; k+\tau+1, m-\tau-1) \\
& + Q_{X_n}^4(l, k-\tau-1; k+2\tau+2, m) \\
& + Q_{X_n}^4(l+\tau+1, k; k+\tau+1, m-\tau-1) \\
& - Q_{X_n}^4(l+\tau+1, k; k+2\tau+2, m)
\end{aligned}$$

where for $w_1 < w_2$, $h_1 < h_2$, $w_2 \leq h_1 - \tau$,

$$Q_{X_n}^u(w_1, w_2; h_1, h_2) = \tilde{S}_{X_n}^u(w_1, h_2|\tau) - \tilde{S}_{X_n}^u(w_1, h_1-1|\tau) - \tilde{S}_{X_n}^u(w_2+1, h_2|\tau) + \tilde{S}_{X_n}^u(w_2+1, h_1-1|\tau),$$

otherwise $Q_{X_n}^u(w_1, w_2; h_1, h_2) = 0$. See Figure 3.1 for an illustration of $Q_{X_n}^u$ and $\tilde{S}_{X_n}^u$. Thus, for fixed $\eta \in (0, 1)$ and $u = 1, 2, 3, 4$, it is key to study the following two parameter processes

$$S_{X_n}^u(a, b|\eta) = \begin{cases} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor an \rfloor}^i w_{i,j}^u X_{i+\lfloor \eta m \rfloor + 1}^T X_j, & 0 < a < b - \eta < 1 - \eta; \\ 0, & \text{otherwise.} \end{cases} \quad (3.3)$$

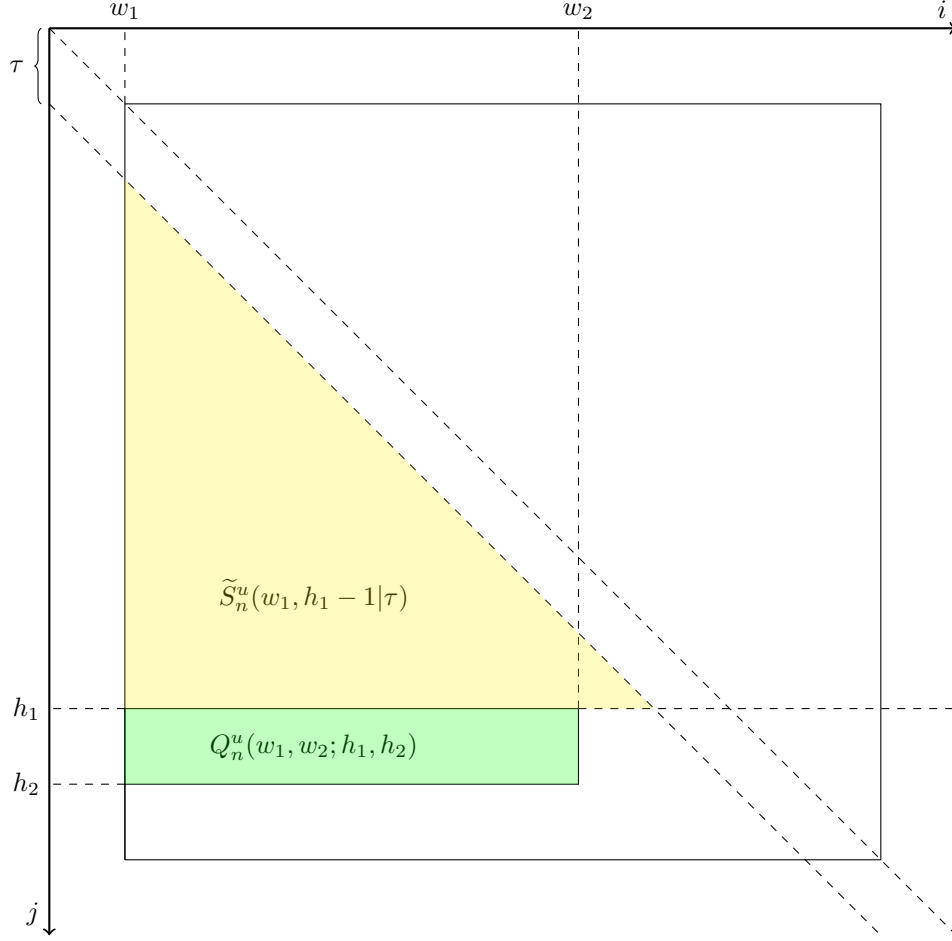


Figure 3.1: Illustration of $Q_{X_n}^u(w_1, w_2; h_1, h_2)$ (green region) and $\tilde{S}_{X_n}^u(w_1, h_1 - 1|\tau)$ (yellow region)

where

$$w_{i,j}^u = \mathbb{I}_{\{u=1\}} + \frac{j}{n} \mathbb{I}_{\{u=2\}} + \frac{i + \lfloor \eta n \rfloor + 1}{n} \mathbb{I}_{\{u=3\}} + \frac{i + \lfloor \eta n \rfloor + 1}{n} \frac{j}{n} \mathbb{I}_{\{u=4\}}.$$

Recall that Beveridge Nelson (BN) decomposition in Phillips and Solo (1992) implies $X_i = D_i - \varepsilon_i$, where $D_i = (\sum_{u=0}^{\infty} c_u) \varepsilon_i$, $\tilde{D}_i = \sum_{j=0}^{\infty} (\sum_{u=j+1}^{\infty} c_u) \varepsilon_{i-j}$ and $\varepsilon_i = \tilde{D}_i - \tilde{D}_{i-1}$. By applying the BN decomposition, we would have for any $u = 1, 2, 3, 4$

$$\frac{\sqrt{2}}{n \|\Gamma\|_F} S_{X_n}^u(a, b|\eta) = \frac{\sqrt{2}}{n \|\Gamma\|_F} S_{D_n}^u(a, b|\eta) + R_u, \quad (3.4)$$

where $S_{D_n}^u(a, b|\eta)$ is defined similarly as in equation (3.3) and it holds in $l^\infty([0, 1]^2)$ that $R_u \rightsquigarrow 0$. The proof is postponed to Section 3.6.2. Consequently, it holds in $l^\infty([0, 1]^3)$ that

$$H_{X_n}(r; a, b|\eta) := \tilde{H}_{X_n}(\lfloor rn \rfloor; \lfloor an \rfloor, \lfloor bn \rfloor | \lfloor \eta n \rfloor) = H_{D_n}(r; a, b|\eta) + o_p(1).$$

The convergence of marginals $(H_{D_n}(r_1; a_1, b_1|\eta), \dots, H_{D_n}(r_K; a_K, b_K|\eta))$ is shown in Section 3.6.2 and the tightness of $H_{D_n}(r; a, b|\eta)$ follows from the tightness of each $\frac{\sqrt{2}}{n \|\Gamma\|_F} S_{D_n}^u(a, b|\eta)$, which can be shown similarly

as in Wang et al. (2019). So, we have $H_{D_n}(r; a, b) \rightsquigarrow G(r; a, b)$ in $l^\infty([0, 1]^3)$. Finally, Theorem 12 can be proved similarly as in Wang et al. (2019).

Convergence of Marginals

It suffices to show that for any fixed intervals $(a_{u,k}, b_{u,k}) \in (0, 1)^2$ and constants $\alpha_{u,k} \in \mathbb{R}$, where $k = 1, 2, \dots, K$, $u = 1, 2, 3, 4$, it holds that

$$\frac{\sqrt{2}}{n\|\Gamma\|_F} \sum_{u=1}^4 \sum_{k=1}^K \alpha_{u,k} S_{D_n}^u(a_{u,k}, b_{u,k}|\eta) \xrightarrow{\mathcal{D}} \sum_{u=1}^4 \sum_{k=1}^K \alpha_{u,k} V_u(a_{u,k}, b_{u,k}|\eta).$$

Some algebra show that

$$\frac{\sqrt{2}}{n\|\Gamma\|_F} \sum_{u=1}^4 \sum_{k=1}^K \alpha_{u,k} S_{D_n}^u(a_{u,k}, b_{u,k}|\eta) = \sum_{i=\lfloor a_{\min} n \rfloor}^{\lfloor b_{\max} n \rfloor - \lfloor \eta n \rfloor - 1} \tilde{\xi}_i,$$

where $a_{\min} = \min_{u,k} a_{u,k}$, $b_{\max} = \max_{u,k} b_{u,k}$ and

$$\begin{aligned} \tilde{\xi}_i &= \sum_{u=1}^4 \sum_{k=1}^K \mathbb{I}_{\{\lfloor a_{u,k} n \rfloor \leq i \leq \lfloor b_{u,k} n \rfloor - \lfloor \eta n \rfloor - 1\}} \alpha_{u,k} \xi_{a_{u,k}, i}^u, \\ \xi_{a_{u,k}, i}^u &= \frac{\sqrt{2}}{n\|\Gamma\|_F} \sum_{j=\lfloor a_{u,k} n \rfloor}^i w_{i,j}^u D_{i+\lfloor \eta n \rfloor + 1}^T D_j. \end{aligned}$$

Then, the conditional variance is calculated as

$$\begin{aligned} \sum_{i=\lfloor a_{\min} n \rfloor}^{\lfloor b_{\max} n \rfloor - \lfloor \eta n \rfloor - 1} E[\tilde{\xi}_i^2 | \mathcal{F}_{i-1}] &= \sum_{u_1=1}^4 \sum_{k_1=1}^K \sum_{u_2=1}^4 \sum_{k_2=1}^K \left\{ \alpha_{u_1, k_1} \alpha_{u_2, k_2} \right. \\ &\quad \left. \sum_{i=\lfloor (a_{u_1, k_1} \vee a_{u_2, k_2}) n \rfloor}^{\lfloor (b_{u_1, k_1} \wedge b_{u_2, k_2}) n \rfloor - \lfloor \eta n \rfloor - 1} E[\xi_{a_{u_1, k_1}, i}^{u_1} \xi_{a_{u_2, k_2}, i}^{u_2} | \mathcal{F}_{i-1}] \right\}. \end{aligned}$$

It can be shown that under Assumption 48, for $a' \leq a \leq b - \eta \leq 1 - \eta$

$$\sum_{i=\lfloor a n \rfloor}^{\lfloor b n \rfloor - \lfloor \eta n \rfloor - 1} E[\xi_{a', i}^u \xi_{a, i}^v | \mathcal{F}_{i-1}] \xrightarrow{L^2} C_{u,v}(a, b), \quad (3.5)$$

where, $C_{u,v}(a, b) = 0$ if $a > b - \eta$; Otherwise, it is given as

$$C_{u,v}(a, b) = \lim_{n \rightarrow \infty} \frac{2}{n^2} \sum_{i=\lfloor a n \rfloor}^{\lfloor b n \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor a n \rfloor}^i w_{i,j}^u w_{i,j}^v.$$

The proof is postponed to Section 3.6.2. Thus, we have

$$\sum_{i=\lfloor a_{\min} n \rfloor}^{\lfloor b_{\max} n \rfloor - \lfloor \eta n \rfloor - 1} E[\tilde{\xi}_i^2 | \mathcal{F}_{i-1}] \xrightarrow{p} \sum_{u_1=1}^4 \sum_{k_1=1}^K \sum_{u_2=1}^4 \sum_{k_2=1}^K \alpha_{u_1, k_1} \alpha_{u_2, k_2} C_{u,v}(a_{u_1, k_1} \vee a_{u_2, k_2}, b_{u_1, k_1} \wedge b_{u_2, k_2}).$$

Next, we check the conditional Lindeberg condition. To this end, it suffices to show that for any fixed interval (a, b) and $u \in \{1, 2, 3, 4\}$

$$\sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} E[(\xi_{a,i}^u)^4] = o(1).$$

Due to the independence of $\{D_i\}_{i=1}^n$, we have

$$E[D_{i+\lfloor \eta n \rfloor + 1}^T D_{j_1} D_{i+\lfloor \eta n \rfloor + 1}^T D_{j_2} D_{i+\lfloor \eta n \rfloor + 1}^T D_{j_3} D_{i+\lfloor \eta n \rfloor + 1}^T D_{j_4}] \neq 0$$

only if j_1, j_2, j_3, j_4 are pair-wise equal. In addition, from Lemma 53

$$E[D_{i+\lfloor \eta n \rfloor + 1}^T D_{j_1} D_{i+\lfloor \eta n \rfloor + 1}^T D_{j_2} D_{i+\lfloor \eta n \rfloor + 1}^T D_{j_3} D_{i+\lfloor \eta n \rfloor + 1}^T D_{j_4}] \lesssim \|\Gamma\|_F^4.$$

Thus, we have $\sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} E[(\xi_{a,i}^u)^4] \lesssim O(1/n)$.

Proof of Equation (3.4)

Firstly, write $S_{X_n}^u(a, b|\eta)$ as

$$\frac{\sqrt{2}}{n\|\Gamma\|_F} S_{X_n}^u(a, b|\eta) = \frac{\sqrt{2}}{n\|\Gamma\|_F} S_{D_n}^u(a, b|\eta) + R_u,$$

where

$$R_u = \frac{\sqrt{2}}{n\|\Gamma\|_F} \left\{ - \underbrace{\sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor an \rfloor}^i w_{i,j}^u D_{i+\lfloor \eta n \rfloor + 1}^T \varepsilon_j}_{R_{u,1}} - \underbrace{\sum_{j=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{i=j}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} w_{i,j}^u \varepsilon_{i+\lfloor \eta n \rfloor + 1}^T D_j}_{R_{u,2}} + \underbrace{\sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor an \rfloor}^i w_{i,j}^u \varepsilon_{i+\lfloor \eta n \rfloor + 1}^T \varepsilon_j}_{R_{u,3}} \right\}.$$

We then show that each of the terms $R_{u,1}, R_{u,2}, R_{u,3}$ converges weakly to 0 in $l^\infty([0, 1]^2)$. The proof techniques for $R_{u,1}$ and $R_{u,2}$ are very similar, here we only give details to show that $R_{u,2} \rightsquigarrow 0$. Some algebra show that

$$\sum_{i=j}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} w_{i,j}^u \varepsilon_{i+\lfloor \eta n \rfloor + 1} = \begin{cases} \tilde{D}_{\lfloor bn \rfloor} - \tilde{D}_{j+\lfloor \eta n \rfloor}, & u = 1; \\ \frac{j}{n} \left\{ \tilde{D}_{\lfloor bn \rfloor} - \tilde{D}_{j+\lfloor \eta n \rfloor} \right\}, & u = 2; \\ \frac{\lfloor bn \rfloor + 1}{n} \tilde{D}_{\lfloor bn \rfloor} - \frac{1}{n} \left\{ \sum_{i=j}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \tilde{D}_{i+\lfloor \eta n \rfloor + 1} \right\} - \frac{j+\lfloor \eta n \rfloor + 1}{n} \tilde{D}_{j+\lfloor \eta n \rfloor}, & u = 3; \\ \frac{j}{n} \left\{ \frac{\lfloor bn \rfloor + 1}{n} \tilde{D}_{\lfloor bn \rfloor} - \frac{1}{n} \left\{ \sum_{i=j}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \tilde{D}_{i+\lfloor \eta n \rfloor + 1} \right\} - \frac{j+\lfloor \eta n \rfloor + 1}{n} \tilde{D}_{j+\lfloor \eta n \rfloor} \right\}, & u = 4; \end{cases}$$

The above decomposition for $R_{u,2}$ is complex, fortunately it can be simplified with the following lemma.

Lemma 54. Under Assumption 48, it holds in $l^\infty([0, 1]^2)$ that

$$\sup_{a,b} \left| \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right| = o_p(1),$$

where $\{c_i\}$ is a sequence of constants such that $\sup_i |c_i| \leq 1$.

Thus, we can throw away the following terms

$$\begin{aligned} & \sum_{j=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} D_j^T \tilde{D}_{j+\lfloor \eta n \rfloor}, \quad \sum_{j=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \frac{j}{n} D_j^T \tilde{D}_{j+\lfloor \eta n \rfloor}, \\ & \sum_{j=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \frac{j + \lfloor \eta n \rfloor + 1}{n} D_j^T \tilde{D}_{j+\lfloor \eta n \rfloor}, \quad \sum_{j=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \frac{j}{n} \frac{j + \lfloor \eta n \rfloor + 1}{n} D_j^T \tilde{D}_{j+\lfloor \eta n \rfloor}. \end{aligned}$$

Next, we focus on the following term

$$\frac{\sqrt{2}}{n^2 \|\Gamma\|_F} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=i}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} c_i D_i^T \tilde{D}_{j+\lfloor \eta n \rfloor+1} = \frac{\sqrt{2}}{n^2 \|\Gamma\|_F} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor an \rfloor}^i c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j.$$

Then, applying the triangle inequality

$$\left| \frac{\sqrt{2}}{n^2 \|\Gamma\|_F} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor an \rfloor}^i c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right| \quad (3.6)$$

$$\leq \left| \frac{\sqrt{2}}{n^2 \|\Gamma\|_F} \sum_{i=1}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=1}^i c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right| \quad (3.7)$$

$$+ \left| \frac{\sqrt{2}}{n^2 \|\Gamma\|_F} \sum_{i=1}^{\lfloor an \rfloor - 1} \sum_{j=1}^i c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right| \quad (3.8)$$

$$+ \left| \frac{\sqrt{2}}{n^2 \|\Gamma\|_F} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=1}^{\lfloor an \rfloor - 1} c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right|. \quad (3.9)$$

Due to the following lemma, terms (3.7) and (3.8) are both of order $o_p(1)$ in metric space $l^\infty([0, 1]^2)$.

Lemma 55. Under Assumption 48, it holds in $l^\infty([0, 1]^2)$ that

$$\frac{1}{n} \left\| \sup_a \frac{\sqrt{2}}{n \|\Gamma\|_F} \left| \sum_{i=1}^{\lfloor an \rfloor} \sum_{j=1}^i c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right| \right\|_2 = o(1),$$

where $\{c_j\}$ is a sequence of constants such that $\sup_j |c_j| \leq 1$.

Next, denote the L^p -norm of a random variable X as

$$\|X\|_p := (E[|X|^p])^{1/p}.$$

For any two parameter process $W(a, b)$, if $\|W(a, b)\|_6 \lesssim 1/\sqrt{n}$, then its marginals $(W(a_1, b_1), \dots, W(a_k, b_k))$

converges to 0 and from the proof of Equation 8.2 in Wang et al. (2019), it is asymptotically tight. Thus, we have $W(a, b) \rightsquigarrow 0$. With this logic and the following lemma, term (3.9) is also asymptotically negligible.

Lemma 56. *Under Assumption 48, it holds in $l^\infty([0, 1]^2)$ that*

$$\left\| \frac{\sqrt{2}}{n^2 \|\Gamma\|_F} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=1}^{\lfloor an \rfloor - 1} c_j \tilde{D}_{i+\lfloor \eta n \rfloor + 1}^T D_j \right\|_6 \lesssim \frac{1}{\sqrt{n}},$$

where $\{c_i\}$ is a sequence of constants such that $\sup_i |c_i| \leq 1$.

Finally, using the same logic, it can be seen from the lemma below that all the other terms in $R_{u,2}$ converge weakly to 0.

Lemma 57. *Under Assumption 48, it holds in $l^\infty([0, 1]^2)$ that*

$$\left\| \frac{\sqrt{2}}{n \|\Gamma\|_F} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} c_i D_i^T \tilde{D}_{\lfloor bn \rfloor} \right\|_6 \lesssim \frac{1}{\sqrt{n}},$$

where $\{c_i\}$ is a sequence of constants such that $\sup_i |c_i| \leq 1$.

This concludes the proof that $R_{u,2} \rightsquigarrow 0$. For $R_{u,3}$, notice that

$$|R_{u,3}| \leq \left| \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor an \rfloor}^i w_{i,j}^u \tilde{D}_{i+\lfloor \eta n \rfloor + 1}^T \varepsilon_j \right| + \left| \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor an \rfloor}^i w_{i,j}^u \tilde{D}_{i+\lfloor \eta n \rfloor}^T \varepsilon_j \right|.$$

Comparing the above two terms with $R_{u,2}$, we have $\tilde{D}_{i+\lfloor \eta n \rfloor}$ instead of $D_{i+\lfloor \eta n \rfloor}$. Since both Lemma 51 and 52 hold for any combination of \tilde{D}_i and D_j , we can show similarly as in the proof for $R_{u,2}$ that these two terms are of order $o_p(1)$.

Proof of Equation (3.5)

Proof. Notice that

$$\sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} E[\xi_{a',i}^u \xi_{a,i}^v | \mathcal{F}_{i-1}] = \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} E[\xi_{a,i}^u \xi_{a,i}^v | \mathcal{F}_{i-1}] + \tilde{R},$$

where

$$\begin{aligned} \tilde{R} &= \frac{2}{n^2 \|\Gamma\|_F^2} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j'=\lfloor a'n \rfloor}^{\lfloor an \rfloor - 1} \sum_{j=\lfloor an \rfloor}^i w_{i,j'}^u w_{i,j}^v E[D_{i+\lfloor \eta n \rfloor + 1}^T D_{j'} D_{i+\lfloor \eta n \rfloor + 1}^T D_j | \mathcal{F}_{i-1}] \\ &= \frac{2}{n^2 \|\Gamma\|_F^2} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j'=\lfloor a'n \rfloor}^{\lfloor an \rfloor - 1} \sum_{j=\lfloor an \rfloor}^i w_{i,j'}^u w_{i,j}^v \text{tr}(D_{j'}^T \Gamma D_j) \end{aligned}$$

We then show that \tilde{R} is negligible.

$$\begin{aligned}
E[\tilde{R}^2] &\leq \frac{4}{n^4 \|\Gamma\|_F^4} \sum_{i_1, i_2 = \lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j'_1, j'_2 = \lfloor a'n \rfloor}^{\lfloor an \rfloor - 1} \sum_{j_1 = \lfloor an \rfloor}^{i_1} \sum_{j_2 = \lfloor an \rfloor}^{i_2} \left| E[D_{j'_1}^T \Gamma D_{j_1} D_{j'_2}^T \Gamma D_{j_2}] \right| \\
&\asymp \frac{4}{n^4 \|\Gamma\|_F^4} \sum_{i_1 < i_2}^{\lfloor an \rfloor - 1} \sum_{j' = \lfloor a'n \rfloor}^{i_1} \sum_{j = \lfloor an \rfloor}^{i_1} |E[D_{j'}^T \Gamma D_j D_{j'}^T \Gamma D_j]| \\
&\asymp \frac{4}{n^4 \|\Gamma\|_F^4} n^4 \text{tr}(\Gamma^4) \\
&= o(1).
\end{aligned}$$

Next, we focus on the first term

$$\begin{aligned}
&\sum_{i = \lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} E[\xi_{a,i}^u \xi_{a,i}^v | \mathcal{F}_{i-1}] \\
&= \frac{2}{n^2 \|\Gamma\|_F^2} \sum_{i = \lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j = \lfloor an \rfloor}^i \sum_{j' = \lfloor an \rfloor}^i w_{i,j}^u w_{i,j'}^v E[D_i^T D_j D_{j'}^T D_i | \mathcal{F}_{i-1}] \\
&= \frac{2}{n^2 \|\Gamma\|_F^2} \sum_{i = \lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j = \lfloor an \rfloor}^i \sum_{j' = \lfloor an \rfloor}^i w_{i,j}^u w_{i,j'}^v D_j^T \Gamma D_{j'}
\end{aligned}$$

whose mean can be calculated as

$$\begin{aligned}
\sum_{i = \lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} E[\xi_{a,i}^u \xi_{a,i}^v] &= \frac{2}{n^2 \|\Gamma\|_F^2} \sum_{i = \lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j = \lfloor an \rfloor}^i w_{i,j}^u w_{i,j}^v \text{tr}(\Gamma^2) \\
&= \frac{2}{n^2} \sum_{i = \lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j = \lfloor an \rfloor}^i w_{i,j}^u w_{i,j}^v \\
&\rightarrow C_{u,v}(a, b) \text{ as } n \rightarrow \infty.
\end{aligned}$$

Next, we show that the variance of the first term is asymptotically 0. Observe that

$$\text{cov}(D_{j_1}^T \Gamma D_{j'_1}, D_{j_2}^T \Gamma D_{j'_2}) = \begin{cases} E[D_1^T \Gamma D_1 D_1^T \Gamma D_1] - \text{tr}(\Gamma^2)^2, & j_1 = j'_1 = j_2 = j'_2; \\ \text{tr}(\Gamma^4), & j_1 = j_2, j'_1 = j'_2, j_1 \neq j'_1; \\ \text{tr}(\Gamma^4), & j_1 = j'_2, j'_1 = j_2, j_1 \neq j'_1; \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$\begin{aligned}
& \text{var} \left(\sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} E[\xi_{a,i}^u \xi_{a,i}^v | \mathcal{F}_{i-1}] \right) \\
&= \sum_{i_1, i_2 = \lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j_1, j'_1 = \lfloor an \rfloor}^{i_1} \sum_{j_2, j'_2 = \lfloor an \rfloor}^{i_2} \text{cov}(D_{j_1}^T \Gamma D_{j'_1}, D_{j_2}^T \Gamma D_{j'_2}) \\
&\lesssim \frac{1}{n^4 \|\Gamma\|_F^4} O(n^4) \text{tr}(\Gamma^4) \\
&= o(1),
\end{aligned}$$

where the above inequality holds true since there are at most $O(n^4)$ non-zero terms. \square

3.6.3 Proof of Theorem 13

We first state a lemma.

Lemma 58. *Under Assumption 48, for any constant vector $\delta_n \in \mathbb{R}^p$,*

$$\sup_{1 \leq k, l \leq n} \left| \sum_{i=l}^k c_i X_i^T \delta_n \right| = O_p(\sqrt{n} \log_2(n) \|\delta_n\|_1).$$

where $\{c_i\}$ is a sequence of constants such that $\sup_i |c_i| \leq 1$.

Notice that under the above lemma, we have

$$\begin{aligned}
S_{Y_n}^u(a, b | \eta) &= S_{X_n}^u(a, b | \eta) + n O_p(\sqrt{n} \log_2(n) \|\delta_n\|_1) \\
&\quad + \begin{cases} \left(\frac{1}{n} \sum_{i=\lfloor \phi n \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=\lfloor \phi n \rfloor}^i w_{i,j}^u \right) n \|\delta_n\|_2^2, & \text{if } a < \phi < b - \eta, \\ 0, & \text{otherwise,} \end{cases}
\end{aligned}$$

where $\sqrt{n} \log_2(n) \|\delta_n\|_1 = o(n \|\delta_n\|_2^2)$. Then, Theorem 13 follows straight-forwardly under the case that $n \|\delta_n\|_2^2 / \|\Gamma\|_F \rightarrow c \in (0, \infty)$ or $n \|\delta_n\|_2^2 / \|\Gamma\|_F \rightarrow 0$. For the case that $n \|\delta_n\|_2^2 / \|\Gamma\|_F \rightarrow \infty$, it follows from Lemma 58 that

$$\begin{aligned}
\frac{\sqrt{2}}{n^3 \|\Gamma\|_F} \tilde{H}_{Y_n}(\lfloor \phi n \rfloor; 1, n | \lfloor \eta n \rfloor) &= \frac{\sqrt{2}}{n^3 \|\Gamma\|_F} \tilde{H}_{X_n}(\lfloor \phi n \rfloor; 1, n | \lfloor \eta n \rfloor) \\
&\quad + O_p \left(\frac{\sqrt{n} \log_2(n) \|\delta_n\|_1}{\|\Gamma\|_F} \right) + \frac{n \|\delta_n\|_2^2}{\|\Gamma\|_F} \frac{\sum_{1 \leq j_1, j_3 \leq \lfloor \phi n \rfloor} \sum_{\lfloor \phi n \rfloor + \lfloor \eta n \rfloor + 1 \leq j_2, j_4 \leq n} \mathbb{1}_{|j_1 - j_3| > \lfloor \eta n \rfloor} \mathbb{1}_{|j_2 - j_4| > \lfloor \eta n \rfloor}}{n^4},
\end{aligned}$$

which implies that $\frac{\sqrt{2}}{n^3 \|\Gamma\|_F} \tilde{H}_{Y_n}(\lfloor \phi n \rfloor; 1, n | \tau)$ goes to infinity in probability. Then, the result follows similarly from Wang et al. (2019).

3.6.4 Proof of Auxiliary Lemmas

Proof of Lemma 51

Proof. Firstly, for the case $j = 0$, let $i_{\min} = \min\{i_1, i_2, \dots, i_h\}$, $\tilde{c}_{i, (l, \cdot)}$ be the l -th row of \tilde{c}_i and $\tilde{c}_{i, (l, k)}$ be the (l, k) -th entry of \tilde{c}_i , the absolute value of cumulant can be bounded as

$$\begin{aligned}
& \sum_{l_1, \dots, l_h=1}^p |\text{cum}(\tilde{D}_{i_1, l_1}, \dots, \tilde{D}_{i_h, l_h})| \\
& \leq \sum_{l_1, \dots, l_h=1}^p \sum_{j=0}^{\infty} \sum_{k_1, \dots, k_h=1}^p \prod_{g=1}^h |\tilde{c}_{i_g - i_{\min} + j, (l_g, k_g)} \text{cum}(\epsilon_{i_{\min} - j, k_1}, \dots, \epsilon_{i_{\min} - j, k_h})| \\
& = \sum_{k_1, \dots, k_h=1}^p \left(\sum_{l_1, \dots, l_h=1}^p \sum_{j=0}^{\infty} \prod_{g=1}^h |\tilde{c}_{i_g - i_{\min} + j, (l_g, k_g)}| \right) |\text{cum}(\epsilon_{0, k_1}, \dots, \epsilon_{0, k_h})| \\
& \leq \left(\sum_{j=0}^{\infty} \prod_{g=1}^h \|\tilde{c}_{i_g - i_{\min} + j}\|_1 \right) \sum_{k_1, \dots, k_h=1}^p |\text{cum}(\epsilon_{0, k_1}, \dots, \epsilon_{0, k_h})| \\
& \lesssim \left(\sum_{j=0}^{\infty} \left(\sum_{u=j+1}^{\infty} \|c_u\|_1 \right)^h \right) \|\Gamma\|_F^h,
\end{aligned}$$

which proves the case $j = 0$. For other cases, the results can be shown similarly. To bound the square cumulant, notice that under assumption 48 (A.1), it can be easily shown that there exists a constant C such that

$$\max_{1 \leq k_1, \dots, k_h \leq p} |\text{cum}(\epsilon_{i_{\min} - j, k_1}, \dots, \epsilon_{i_{\min} - j, k_h})| \leq C.$$

Next, for any (l_1, \dots, l_h) , we can bound $|\text{cum}(\tilde{D}_{i_1, l_1}, \dots, \tilde{D}_{i_h, l_h})|$ as follows

$$\begin{aligned}
& |\text{cum}(\tilde{D}_{i_1, l_1}, \dots, \tilde{D}_{i_h, l_h})| \\
& \leq \sum_{j=0}^{\infty} \sum_{k_1, \dots, k_h=1}^p \prod_{g=1}^h |\tilde{c}_{i_g - i_{\min} + j, (l_g, k_g)} \text{cum}(\epsilon_{i_{\min} - j, k_1}, \dots, \epsilon_{i_{\min} - j, k_h})| \\
& \leq C \sum_{j=0}^{\infty} \sum_{k_1, \dots, k_h=1}^p \prod_{g=1}^h |\tilde{c}_{i_g - i_{\min} + j, (l_g, k_g)}| \\
& \leq C \sum_{j=0}^{\infty} \left(\sum_{u=j+1}^{\infty} \|c_u\|_{\infty} \right)^h.
\end{aligned}$$

The proof is similar for other cases, thus there exists a constant C' such that for any (l_1, \dots, l_h) , we have $|cum(D_{i_1, l_1}, \dots, D_{i_j, l_j}, \tilde{D}_{i_{j+1}, l_{j+1}}, \dots, \tilde{D}_{i_h, l_h})| \leq C'$. As a consequence, it holds that

$$\begin{aligned} \sum_{l_1, l_2, \dots, l_h=1}^p cum^2(D_{i_1, l_1}, \dots, D_{i_j, l_j}, \tilde{D}_{i_{j+1}, l_{j+1}}, \dots, \tilde{D}_{i_h, l_h})/C'^2 &\leq \\ \sum_{l_1, l_2, \dots, l_h=1}^p |cum(D_{i_1, l_1}, \dots, D_{i_j, l_j}, \tilde{D}_{i_{j+1}, l_{j+1}}, \dots, \tilde{D}_{i_h, l_h})|/C' &\lesssim \|\Gamma\|_F^h. \end{aligned}$$

□

Proof of Lemma 52

Proof. We show that $\{X_i\}$ is UGMC(8), then the result follows from Remark 9.6 Wang and Shao (2019). Firstly, we have

$$\begin{aligned} &E[(X_{i,l} - X'_{i,l})^8] \\ &= E\left[\left(\sum_{j=i}^{\infty} c_{j,(l,\cdot)}^T(\epsilon_{i-j} - \epsilon'_{i-j})\right)^8\right] \\ &= \sum_{j_1, \dots, j_8=i}^{\infty} E\left[\left(c_{j_1,(l,\cdot)}^T(\epsilon_{i-j_1} - \epsilon'_{i-j_1})\right) \cdots \left(c_{j_8,(l,\cdot)}^T(\epsilon_{i-j_8} - \epsilon'_{i-j_8})\right)\right] \\ &\leq \sum_{j_1, \dots, j_8=i}^{\infty} \left(E\left[\left(c_{j_1,(l,\cdot)}^T(\epsilon_{i-j_1} - \epsilon'_{i-j_1})\right)^8\right]\right)^{1/8} \cdots \left(E\left[\left(c_{j_8,(l,\cdot)}^T(\epsilon_{i-j_8} - \epsilon'_{i-j_8})\right)^8\right]\right)^{1/8} \end{aligned}$$

Next, we bound the term $E[(c_{j_1,(l,\cdot)}^T(\epsilon_{i-j_1} - \epsilon'_{i-j_1}))^8]$ as an example.

$$\begin{aligned} &E\left[\left(c_{j_1,(l,\cdot)}^T(\epsilon_{i-j_1} - \epsilon'_{i-j_1})\right)^8\right] \\ &= E\left[\left(\sum_{k=1}^p c_{j_1,(l,k)}(\epsilon_{i-j_1,k} - \epsilon'_{i-j_1,k})\right)^8\right] \\ &= \sum_{k_1, \dots, k_8} \left(\prod_{g=1}^8 c_{j_1,(l,k_g)}\right) E\left[\prod_{g=1}^8 (\epsilon_{i-j_1,k_g} - \epsilon'_{i-j_1,k_g})\right] \\ &\lesssim \left(\sum_{k=1}^p c_{j_1,(l,k)}\right)^8 \leq \|c_{j_1}\|_{\infty}^8 \end{aligned}$$

Thus, we have $E[(X_{i,l} - X'_{i,l})^8] \leq (\sum_{j=i}^{\infty} \|c_j\|_{\infty})^8$. Finally, it can be shown similarly that $\sup_l E[|X_{0,l}|^8] \leq C^8$ for some constant C . □

Proof of Lemma 53

Proof. Let π be any disjoint partition over the set $\{l_1, l_2, l_3, l_4, l_5, l_6\}$, the number of partitions π such that for any $B \in \pi$, $|B| \neq 1$ is

$$\left\{1 + \binom{6}{2} + \binom{6}{3}\right\}.$$

Under assumption 48,

$$\begin{aligned}
& E[(D_i^T D_j)^6] \\
&= \sum_{l_1, l_2, l_3, l_4} E[D_{i, l_1} D_{i, l_2} D_{i, l_3} D_{i, l_4} D_{i, l_5} D_{i, l_6}] E[D_{j, l_1} D_{j, l_2} D_{j, l_3} D_{j, l_4} D_{j, l_5} D_{j, l_6}] \\
&= \sum_{l_1, l_2, l_3, l_4} (E[D_{i, l_1} D_{i, l_2} D_{i, l_3} D_{i, l_4} D_{i, l_5} D_{i, l_6}])^2 \\
&= \sum_{l_1, l_2, l_3, l_4, l_5, l_6} \left(\sum_{\pi} \prod_{B \in \pi} \text{cum}(D_{i, l_k} : l_k \in B) \right)^2 \\
&\leq \left\{ 1 + \binom{6}{2} + \binom{6}{3} \right\} \sum_{l_1, l_2, l_3, l_4, l_5, l_6} \sum_{\pi} \prod_{B \in \pi} \text{cum}^2(D_{i, l_k} : l_k \in B) \text{ by Cauchy's inequality} \\
&\leq \left\{ 1 + \binom{6}{2} + \binom{6}{3} \right\} \sum_{\pi} \prod_{B \in \pi} \left\{ \sum_{l_k \in B} \sum_{l_k=1}^p \text{cum}^2(D_{i, l_k} : l_k \in B) \right\} \\
&\leq \left\{ 1 + \binom{6}{2} + \binom{6}{3} \right\} \sum_{\pi} \prod_{B \in \pi} C \|\Gamma\|_F^{|B|} \quad \text{by Lemma 51} \\
&= \left\{ 1 + \binom{6}{2} + \binom{6}{3} \right\} C \|\Gamma\|_F^6,
\end{aligned}$$

where C is the constant in Lemma 51. □

Proof of Lemma 54

Proof. The triangle inequality implies that

$$\begin{aligned}
& \sup_{a, b} \left| \frac{\sqrt{2}}{n \|\Gamma\|_F} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right| \leq \\
& \sup_b \left| \frac{\sqrt{2}}{n \|\Gamma\|_F} \sum_{i=1}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right| + \sup_a \left| \frac{\sqrt{2}}{n \|\Gamma\|_F} \sum_{i=1}^{\lfloor an \rfloor - 1} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right|.
\end{aligned}$$

It is sufficient to show that

$$\left\| \sup_{a \in (0, 1-\eta)} \left| \frac{\sqrt{2}}{n \|\Gamma\|_F} \sum_{i=1}^{\lfloor an \rfloor} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right| \right\|_2 = o(1).$$

To prove this, the idea is to use Proposition 1 in Wu (2007), for any $n = 2^d$

$$\left\| \sup_{a \in (0, 1-\eta)} \left| \frac{\sqrt{2}}{n \|\Gamma\|_F} \sum_{i=1}^{\lfloor an \rfloor} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right| \right\|_2 \leq \sum_{v=0}^d \left[\sum_{u=1}^{2^{d-v}} \left\| \frac{\sqrt{2}}{2^d \|\Gamma\|_F} \sum_{i=2^{v(u-1)+1}}^{2^v u} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right\|_2^2 \right]^{1/2}.$$

Applying Lemma 52 and 51,

$$\begin{aligned}
& \left\| \frac{\sqrt{2}}{2^d \|\Gamma\|_F} \sum_{i=2^v(u-1)+1}^{2^v u} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right\|_2^2 \\
& \leq \frac{2}{2^{2d} \|\Gamma\|_F^2} \sum_{i_1, i_2=2^v(u-1)+1}^{2^v u} \left| E \left[D_{i_1}^T \tilde{D}_{i_1+\lfloor \eta n \rfloor} D_{i_2}^T \tilde{D}_{i_2+\lfloor \eta n \rfloor} \right] \right| \\
& \leq \frac{2}{2^{2d} \|\Gamma\|_F^2} \sum_{i_1, i_2=2^v(u-1)+1}^{2^v u} \sum_{l_1, l_2=1}^p \left\{ |cum(D_{i_1, l_1}, \tilde{D}_{i_1+\lfloor \eta n \rfloor, l_1}, D_{i_2, l_2}, \tilde{D}_{i_2+\lfloor \eta n \rfloor, l_2})| \right. \\
& \quad + |cum(D_{i_1, l_1}, \tilde{D}_{i_1+\lfloor \eta n \rfloor, l_1}) cum(D_{i_2, l_2}, \tilde{D}_{i_2+\lfloor \eta n \rfloor, l_2})| \\
& \quad + |cum(D_{i_1, l_1}, D_{i_2, l_2}) cum(\tilde{D}_{i_1+\lfloor \eta n \rfloor, l_1}, \tilde{D}_{i_2+\lfloor \eta n \rfloor, l_2})| \\
& \quad \left. + |cum(D_{i_1, l_1}, \tilde{D}_{i_2+\lfloor \eta n \rfloor, l_2}) cum(D_{i_2, l_2}, \tilde{D}_{i_1+\lfloor \eta n \rfloor, l_1})| \right\} \\
& \lesssim \frac{2}{2^{2d} \|\Gamma\|_F^2} \sum_{i_1, i_2=2^v(u-1)+1}^{2^v u} \left\{ \rho^{|i_1-i_2|} p^2 \rho^{\lfloor \eta n \rfloor} + p^2 \rho^{2\lfloor \eta n \rfloor} \right. \\
& \quad \left. + \rho^{|i_1-i_2|} \|\Gamma\|_2^2 + \rho^{|i_1-i_2-\lfloor \eta n \rfloor|} \|\Gamma\|_2^2 \right\} \\
& \lesssim 2^{v-2d}.
\end{aligned}$$

To continue the calculation,

$$\left\| \sup_{a \in (0, 1-\eta)} \left| \frac{\sqrt{2}}{n \|\Gamma\|_F} \sum_{i=1}^{\lfloor an \rfloor} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right| \right\|_2 \lesssim \sum_{v=0}^d \left[\sum_{u=1}^{2^{d-v}} 2^{v-2d} \right]^{1/2} = O\left(\frac{d}{2^{d/2}}\right),$$

which concludes the case when $n = 2^d$. For arbitrary integer n , the statement follows from the fact that there exists d such that $2^{d-1} \leq n < 2^d$ and

$$\left\| \sup_{a \in (0, 1-\eta)} \left| \frac{\sqrt{2}}{n \|\Gamma\|_F} \sum_{i=1}^{\lfloor an \rfloor} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right| \right\|_2 \leq \left\| \sup_{a \in (0, 1-\eta)} \left| \frac{\sqrt{2}}{2^{d-1} \|\Gamma\|_F} \sum_{i=1}^{\lfloor a 2^d \rfloor} c_i D_i^T \tilde{D}_{i+\lfloor \eta n \rfloor} \right| \right\|_2,$$

where the trimming is set as $\lfloor \eta n \rfloor$ on the left hand side. This does not change the proof and the left hand side is still of order $o(1)$. \square

3.6.5 Proof of Lemma 55

Using Proposition 1 in Wu (2007), for any $n = 2^d$

$$\left\| \sup_a \frac{\sqrt{2}}{n \|\Gamma\|_F} \left| \sum_{i=1}^{\lfloor an \rfloor} \sum_{j=1}^i c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right| \right\|_2 \leq \sum_{v=0}^d \left[\sum_{u=1}^{2^{d-v}} \left\| \frac{\sqrt{2}}{2^d \|\Gamma\|_F} \sum_{i=2^v(u-1)+1}^{2^v u} \sum_{j=1}^i c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right\|_2^2 \right]^{1/2}. \quad (3.10)$$

For the summands inside the bracket,

$$\begin{aligned} & \left\| \frac{\sqrt{2}}{2^d \|\Gamma\|_F} \sum_{i=2^v(u-1)+1}^{2^v u} \sum_{j=1}^i c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right\|_2^2 \\ & \leq \frac{2}{2^{2d} \|\Gamma\|_F^2} \sum_{i_1, i_2=2^v(u-1)+1}^{2^v u} \sum_{j_1=1}^{i_1} \sum_{j_2=1}^{i_2} \sum_{l_1, l_2=1}^p \left| E \left[\tilde{D}_{i_1+\lfloor \eta n \rfloor+1, l_1} D_{j_1, l_1} \tilde{D}_{i_2+\lfloor \eta n \rfloor+1, l_2} D_{j_2, l_2} \right] \right|. \end{aligned}$$

Express the expectation using cumulants

$$\begin{aligned} & \left| E \left[\tilde{D}_{i_1+\lfloor \eta n \rfloor+1, l_1} D_{j_1, l_1} \tilde{D}_{i_2+\lfloor \eta n \rfloor+1, l_2} D_{j_2, l_2} \right] \right| \\ & = |cum(\tilde{D}_{i_1+\lfloor \eta n \rfloor+1, l_1}, D_{j_1, l_1}, \tilde{D}_{i_2+\lfloor \eta n \rfloor+1, l_2}, D_{j_2, l_2})| \\ & + |cum(\tilde{D}_{i_1+\lfloor \eta n \rfloor+1, l_1}, D_{j_1, l_1}) cum(\tilde{D}_{i_2+\lfloor \eta n \rfloor+1, l_2}, D_{j_2, l_2})| \\ & + |cum(\tilde{D}_{i_1+\lfloor \eta n \rfloor+1, l_1} \tilde{D}_{i_2+\lfloor \eta n \rfloor+1, l_2}) cum(D_{j_1, l_1}, D_{j_2, l_2})| \\ & + |cum(\tilde{D}_{i_1+\lfloor \eta n \rfloor+1, l_1}, D_{j_2, l_2}) cum(D_{j_1, l_1}, \tilde{D}_{i_2+\lfloor \eta n \rfloor+1, l_2})|. \end{aligned}$$

By using Lemma 52 and 51, we have

$$\begin{aligned} & \sum_{l_1, l_2=1}^p |cum(\tilde{D}_{i_1+\lfloor \eta n \rfloor+1, l_1}, D_{j_1, l_1}, \tilde{D}_{i_2+\lfloor \eta n \rfloor+1, l_2}, D_{j_2, l_2})| \lesssim \rho^{i_1 \vee i_2 - j_1 \wedge j_2} p^2 \rho^{\lfloor \eta n \rfloor}, \\ & \sum_{l_1, l_2=1}^p |cum(\tilde{D}_{i_1+\lfloor \eta n \rfloor+1, l_1}, D_{j_1, l_1}) cum(\tilde{D}_{i_2+\lfloor \eta n \rfloor+1, l_2}, D_{j_2, l_2})| \lesssim \rho^{i_1 - j_1} \rho^{i_2 - j_2} p^2 \rho^{\lfloor \eta n \rfloor}, \\ & \sum_{l_1, l_2=1}^p |cum(\tilde{D}_{i_1+\lfloor \eta n \rfloor+1, l_1} \tilde{D}_{i_2+\lfloor \eta n \rfloor+1, l_2}) cum(D_{j_1, l_1}, D_{j_2, l_2})| \lesssim \rho^{|j_1 - j_2|} \|\Gamma\|_F^2, \\ & \sum_{l_1, l_2=1}^p |cum(\tilde{D}_{i_1+\lfloor \eta n \rfloor+1, l_1}, D_{j_2, l_2}) cum(D_{j_1, l_1}, \tilde{D}_{i_2+\lfloor \eta n \rfloor+1, l_2})| \lesssim \rho^{|i_1 + \lfloor \eta n \rfloor - j_2|} \|\Gamma\|_F^2. \end{aligned}$$

Since $p^2 \rho^{\lfloor \eta n \rfloor} = O(\|\Gamma\|_F^2)$, some straightforward calculation shows that

$$\begin{aligned}
& \left\| \frac{\sqrt{2}}{2^d \|\Gamma\|_F} \sum_{i=2^v(u-1)+1}^{2^v u} \sum_{j=1}^i \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right\|_2^2 \\
& \lesssim \frac{1}{2^{2d}} \sum_{i_1, i_2=2^v(u-1)+1}^{2^v u} \sum_{j_1=1}^{i_1} \sum_{j_2=1}^{i_2} \left\{ \rho^{i_1 \vee i_2 - j_1 \wedge j_2} + \rho^{i_1 - j_1} \rho^{i_2 - j_2} + \rho^{|j_1 - j_2|} + \rho^{|i_1 + \lfloor \eta n \rfloor - j_2|} \right\} \\
& \lesssim \frac{1}{2^{2d}} 2^{3v} u.
\end{aligned}$$

Plugging the above bound into Equation (3.10) results

$$\frac{1}{n} \left\| \sup_a \frac{\sqrt{2}}{n \|\Gamma\|_F} \left| \sum_{i=1}^{\lfloor an \rfloor} \sum_{j=1}^i c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right| \right\|_2 \lesssim \frac{1}{2^d} \sum_{v=0}^d \left[\frac{1}{2^{2d}} 2^{3v} \sum_{u=1}^{2^{d-v}} u \right]^{1/2} \lesssim 2^{-d/2},$$

which concludes Equation (55) when $n = 2^d$. It can be shown similarly as in the proof of Lemma 54 for arbitrary n .

Proof of Lemma 56

For term (3.9), we prove that

$$\left\| \frac{\sqrt{2}}{n^2 \|\Gamma\|_F} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=1}^{\lfloor an \rfloor - 1} c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right\|_6 \lesssim \frac{1}{\sqrt{n}}.$$

Firstly, notice that

$$\begin{aligned}
& E \left[\left(\frac{\sqrt{2}}{n^2 \|\Gamma\|_F} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=1}^{\lfloor an \rfloor - 1} c_j \tilde{D}_{i+\lfloor \eta n \rfloor+1}^T D_j \right)^6 \right] \\
& \lesssim \frac{1}{n^{12} \|\Gamma\|_F^6} \sum_{i_1, \dots, i_6=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j_1=1}^{\lfloor an \rfloor - 1} \dots \sum_{j_6=1}^{\lfloor an \rfloor - 1} \left| E \left[\prod_{k=1}^6 \tilde{D}_{i_k+\lfloor \eta n \rfloor+1}^T D_{j_k} \right] \right|.
\end{aligned}$$

Then, let π be any partition of the index set $\{(i_k + \lfloor \eta n \rfloor + 1, l_k), (j_k, l_k)\}_{k=1}^6$ such that $|B| > 1$ for any $B \in \pi$.

$$\begin{aligned}
& \left| E \left[\prod_{k=1}^6 \tilde{D}_{i_k+\lfloor \eta n \rfloor+1}^T D_{j_k} \right] \right| = \left| \sum_{l_1, \dots, l_6=1}^p E \left[\prod_{k=1}^6 \tilde{D}_{i_k+\lfloor \eta n \rfloor+1, l_k} D_{j_k, l_k} \right] \right| \\
& \leq \sum_{l_1, \dots, l_6=1}^p \sum_{\pi} \prod_{B \in \pi} |\text{cum}(Z_{i,l} : (i, l) \in B)|,
\end{aligned}$$

where $Z_{i,l} = D_{i,l}$ if $(i,l) \in \{(j_k, l_k)\}_{k=1}^6$; otherwise $Z_{i,l} = \tilde{D}_{i,l}$. Set $\mathbb{I}_1 = \{(i_k + \lfloor \eta n \rfloor + 1, l_k)\}_{k=1}^6$ and $\mathbb{I}_2 = \{(j_k, l_k)\}_{k=1}^6$, we then apply Lemma 52 and 51,

$$\begin{cases} \text{cum}(Z_{i,l} : (i,l) \in B) \lesssim \rho^{i_{max}^B - i_{min}^B}, & \text{if } B \in \pi_1, \\ \text{cum}(Z_{i,l} : (i,l) \in B) \lesssim \rho^{i_{max}^B + \lfloor \eta n \rfloor - j_{min}^B}, & \text{if } B \in \pi_2, \\ \sum_{(i,l) \in B, l=1}^p |\text{cum}(Z_{i,l} : (i,l) \in B)| \lesssim \|\Gamma\|_F^{|B|}, & \text{if } B \in \pi_3. \end{cases}$$

where $\pi_1 := \{A | A \in \pi, A \subseteq \mathbb{I}_1\}$, $\pi_2 := \{A | A \in \pi, A \not\subseteq \mathbb{I}_1, A \not\subseteq \mathbb{I}_2\}$, $\pi_3 := \{A | A \in \pi, A \subseteq \mathbb{I}_2\}$; $i_{max}^B = \max\{i | (i + \lfloor \eta n \rfloor + 1, l) \in B \cap \mathbb{I}_1\}$ and $j_{max}^B = \max\{j | (j, l) \in B \cap \mathbb{I}_2\}$; i_{min}^B and j_{min}^B are defined similarly.

$$\begin{aligned} & \sum_{i_1, \dots, i_6 = \lfloor \eta n \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j_1=1}^{\lfloor an \rfloor - 1} \dots \sum_{j_6=1}^{\lfloor an \rfloor - 1} \sum_{l_1, \dots, l_6=1}^p \sum_{\pi} \prod_{B \in \pi} |\text{cum}(Z_{i,l} : (i,l) \in B)| \\ & \lesssim \sum_{\pi} \prod_{B \in \pi_1} \left(\sum_{\substack{\lfloor an \rfloor \leq i \leq \lfloor bn \rfloor - \lfloor \eta n \rfloor - 1 \\ (i + \lfloor \eta n \rfloor + 1, l) \in B}} \rho^{i_{max}^B - i_{min}^B} \right) \prod_{B \in \pi_3} \left(\sum_{\substack{1 \leq j \leq \lfloor an \rfloor - 1 \\ (j, l) \in B \cap \mathbb{I}_2}} p^{6 - |B|} \|\Gamma\|_F^{|B|} \right) \\ & \quad \times \prod_{B \in \pi_2} \left(\sum_{\substack{\lfloor an \rfloor \leq i \leq \lfloor bn \rfloor - \lfloor \eta n \rfloor - 1 \\ (i + \lfloor \eta n \rfloor + 1, l) \in B \cap \mathbb{I}_1}} \sum_{\substack{1 \leq j \leq \lfloor an \rfloor - 1 \\ (j, l) \in B \cap \mathbb{I}_2}} \rho^{i_{max}^B - j_{min}^B} \rho^{\lfloor \eta n \rfloor} \right). \end{aligned}$$

We know from Equation (3.11) that for $B \in \pi_1$,

$$\sum_{\substack{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1 \\ (i + \lfloor \eta n \rfloor + 1, l) \in B, i = \lfloor an \rfloor}} \rho^{i_{max}^B - i_{min}^B} \lesssim O(n).$$

Also, it can be easily seen that

$$\prod_{B \in \pi_3} \left(\sum_{\substack{1 \leq j \leq \lfloor an \rfloor - 1 \\ (j, l) \in B \cap \mathbb{I}_2}} p^{6 - |B|} \|\Gamma\|_F^{|B|} \right) \lesssim n^{\sum_{B \in \pi_3} |B|} p^{6 - \sum_{B \in \pi_3} |B|} \|\Gamma\|_F^{\sum_{B \in \pi_3} |B|}.$$

In addition, using Equation (3.12), for any $B \in \pi_2$

$$\begin{aligned} & \sum_{\substack{\lfloor an \rfloor \leq i \leq \lfloor bn \rfloor - \lfloor \eta n \rfloor - 1 \\ (i, l) \in B \cap \mathbb{I}_1}} \sum_{\substack{1 \leq j \leq \lfloor an \rfloor - 1 \\ (j, l) \in B \cap \mathbb{I}_2}} \rho^{i_{max}^B - j_{min}^B} = \\ & \left\{ \sum_{\substack{\lfloor an \rfloor \leq i \leq \lfloor bn \rfloor - \lfloor \eta n \rfloor - 1 \\ (i + \lfloor \eta n \rfloor + 1, l) \in B \cap \mathbb{I}_1}} \rho^{i_{max}^B - \lfloor an \rfloor} \right\} \left\{ \sum_{\substack{1 \leq j \leq \lfloor an \rfloor - 1 \\ (j, l) \in B \cap \mathbb{I}_2}} \rho^{\lfloor an \rfloor - j_{min}^B} \right\} = O(1). \end{aligned}$$

In conclusion, since $\|\Gamma\|_F^h / p^h = O(1)$, we have

$$\begin{aligned} & E \left[\left(\frac{\sqrt{2}}{n^2 \|\Gamma\|_F} \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{j=1}^{\lfloor an \rfloor - 1} c_j \tilde{D}_{i + \lfloor \eta n \rfloor + 1}^T D_j \right)^6 \right] \\ & \lesssim \frac{1}{n^{12} \|\Gamma\|_F^6} \max \left\{ n^9, n^3 n^{\sum_{B \in \pi_3} |B|} \rho^{\lfloor \eta n \rfloor} p^{6 - \sum_{B \in \pi_3} |B|} \|\Gamma\|_F^{\sum_{B \in \pi_3} |B|} \right\} \lesssim \frac{1}{n^3}. \end{aligned}$$

Proof of Lemma 57

Proof. For notational convenience, denote

$$\hat{R}(a, b|\eta) = \sum_{i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} c_i D_i^T \tilde{D}_{\lfloor bn \rfloor}.$$

Let π be any disjoint partition over the set \mathbb{I} such that $|B| > 1$ for any $B \in \pi$, where \mathbb{I} is defined as

$$\begin{aligned} \mathbb{I} := & \{(i_1, l_1), (i_2, l_2), (i_3, l_3), (i_4, l_4), (i_5, l_5), (i_6, l_6), \\ & (\lfloor bn \rfloor, l_1), (\lfloor bn \rfloor, l_2), (\lfloor bn \rfloor, l_3), (\lfloor bn \rfloor, l_4), (\lfloor bn \rfloor, l_5), (\lfloor bn \rfloor, l_6)\}, \end{aligned}$$

Any such π is a disjoint union of 3 sets as $\pi = \pi_1 \cup \pi_2 \cup \pi_3$, where $\pi_1 := \{A | A \in \pi, A \subseteq \mathbb{I}_1\}$, $\pi_2 := \{A | A \in \pi, A \not\subseteq \mathbb{I}_1, A \not\subseteq \mathbb{I}_2\}$, $\pi_3 := \{A | A \in \pi, A \subseteq \mathbb{I}_2\}$ and $\mathbb{I}_1, \mathbb{I}_2$ are defined as

$$\begin{aligned} \mathbb{I}_1 &:= \{(i_1, l_1), (i_2, l_2), (i_3, l_3), (i_4, l_4), (i_5, l_5), (i_6, l_6)\}, \\ \mathbb{I}_2 &:= \{(\lfloor bn \rfloor, l_1), (\lfloor bn \rfloor, l_2), (\lfloor bn \rfloor, l_3), (\lfloor bn \rfloor, l_4), (\lfloor bn \rfloor, l_5), (\lfloor bn \rfloor, l_6)\}. \end{aligned}$$

For notational convenience, denote

$$Z_{i,l} = \begin{cases} D_{i,l} & (i, l) \in \mathbb{I}_1, \\ \tilde{D}_{i,l} & (i, l) \in \mathbb{I}_2. \end{cases}$$

Then, we have

$$\begin{aligned} E \left[\hat{R}(a, b|\eta)^6 \right] &\leq \sum_{i_1, i_2, i_3, i_4, i_5, i_6 = \lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \sum_{l_1, l_2, l_3, l_4, l_5, l_6 = 1}^p \left| E \left[D_{i_1, l_1}^T \tilde{D}_{\lfloor bn \rfloor, l_1} \cdots D_{i_6, l_6}^T \tilde{D}_{\lfloor bn \rfloor, l_6} \right] \right| \\ &\leq \sum_{i_1, i_2, i_3, i_4, i_5, i_6} \sum_{l_1, l_2, l_3, l_4, l_5, l_6} \sum_{\pi} \prod_{B \in \pi} |cum(Z_{i,l} : (i, l) \in B)|. \end{aligned}$$

Accordingly, we can decompose the product of cumulants as

$$\begin{aligned} \prod_{B \in \pi} cum(Z_{i,l} : (i, l) \in B) &= \prod_{B \in \pi_1} cum(Z_{i,l} : (i, l) \in B) \\ &\quad \times \prod_{B \in \pi_2} cum(Z_{i,l} : (i, l) \in B) \times \prod_{B \in \pi_3} cum(Z_{i,l} : (i, l) \in B). \end{aligned}$$

Denote $i_{max}^B = \max\{i | (i, l) \in B\}$ and $i_{min}^B = \min\{i | (i, l) \in B\}$. The following bounds on the cumulants are from Lemma 52 and 51

$$\begin{cases} cum(Z_{i,l} : (i, l) \in B) \lesssim \rho^{i_{max}^B - i_{min}^B} & \text{if } B \in \pi_1, \\ cum(Z_{i,l} : (i, l) \in B) \lesssim \rho^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - i_{min}^B} \rho^{\lfloor \eta n \rfloor} & \text{if } B \in \pi_2, \\ \sum_{(i,l) \in B, l=1}^p |cum(Z_{i,l} : (i, l) \in B)| \lesssim \|\Gamma\|_F^{|B|} & \text{if } B \in \pi_3. \end{cases}$$

where we use an example to explain the notation $\sum_{(i,l) \in B, l=1}^p$, i.e., $\sum_{(i,l) \in B, l=1}^p = \sum_{l_1, l_2=1}^p$ if $B = \{(\lfloor bn \rfloor, l_1), (\lfloor bn \rfloor, l_2)\}$. Thus, we have

$$E \left[\hat{R}(a, b|\eta)^6 \right] \lesssim \sum_{\pi} \prod_{B \in \pi_1} \left(\sum_{(i,l) \in B, i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \rho^{i_{max}^B - i_{min}^B} \right) \times \\ \prod_{B \in \pi_2} \left(\sum_{(i,l) \in B \cap \mathbb{I}_1, i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \rho^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - i_{min}^B} \rho^{\lfloor \eta n \rfloor} \right) p^{6 - \sum_{B \in \pi_3} |B|} \|\Gamma\|_F^{\sum_{B \in \pi_3} |B|}.$$

To continue the proof, notice that there exists a constant N_ρ such that $m^6 \rho^{m/2} < 1$ if $m > N_\rho$. Thus, for any $B \in \pi_1$, we have

$$\sum_{(i,l) \in B, i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \rho^{i_{max}^B - i_{min}^B} \lesssim \sum_{\lfloor an \rfloor \leq i_1 < i_2 \leq \lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} (i_2 - i_1)^{|B|-2} \rho^{i_2 - i_1} \\ \lesssim N_\rho^{|B|-2} \sum_{|i_1 - i_2| \leq N_\rho} \rho^{i_2 - i_1} + \sum_{|i_1 - i_2| > N_\rho} (\sqrt{\rho})^{i_2 - i_1} \quad (3.11) \\ = O(n).$$

and for any $B \in \pi_2$,

$$\sum_{(i,l) \in B \cap \mathbb{I}_1, i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \rho^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - i_{min}^B} \\ \lesssim \sum_{i_1=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} (\lfloor bn \rfloor - \lfloor \eta n \rfloor - i_1)^{|B \cap \mathbb{I}_1| - 1} \rho^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - i_1} \quad (3.12) \\ \lesssim \sum_{\lfloor bn \rfloor - \lfloor \eta n \rfloor - i_1 > N_\rho} (\sqrt{\rho})^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - i_1} \\ = O(1).$$

As a consequence, for any $\pi = \pi_1 \cup \pi_2 \cup \pi_3$, we have

$$\prod_{B \in \pi_1} \left(\sum_{(i,l) \in B, i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \rho^{i_{max}^B - i_{min}^B} \right) = O(n^3), \\ \prod_{B \in \pi_2} \left(\sum_{(i,l) \in B \cap \mathbb{I}_1, i=\lfloor an \rfloor}^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - 1} \rho^{\lfloor bn \rfloor - \lfloor \eta n \rfloor - i_{min}^B} \right) = O(1).$$

Since for any $0 \leq h \leq 6$, $\|\Gamma\|_F^h / p^h = O(1)$, we have under Assumption 48 (A.4)

$$\frac{1}{n^6 \|\Gamma\|_F^6} E \left[\hat{R}(a, b|\eta)^6 \right] \lesssim \frac{p^6 \|\Gamma\|_F^{-6} \rho^{\lfloor \eta n \rfloor}}{n^3} \lesssim \frac{1}{n^3}.$$

□

Proof of Lemma 58

Proof. Notice that for any $\delta_n \in \mathbb{R}^p$

$$\sup_{1 \leq k, l \leq n} \left| c_i \sum_{i=l}^k X_i^T \delta_n \right| \leq 2 \sup_{1 \leq k \leq n} \left| \sum_{i=1}^k c_i X_i^T \delta_n \right|.$$

Then, According to Proposition 1 in Wu (2007), for any $n = 2^d$ with some positive integer d ,

$$\left[E \left(\sup_{1 \leq k \leq n} \left| \sum_{i=1}^k c_i X_i^T \delta_n \right| \right)^2 \right]^{1/2} \leq \sum_{v=1}^d \left[\sum_{u=1}^{2^{d-v}} E \left(\sum_{i=2^{v(u-1)+1}}^{2^{vu}} c_i X_i^T \delta_n \right)^2 \right]^{1/2}.$$

Next, we look at each term in the square bracket

$$\begin{aligned} E \left(\sum_{i=2^{v(u-1)+1}}^{2^{vu}} X_i^T \delta_n \right)^2 &= \left| \sum_{2^{v(u-1)+1} \leq i_1, i_2 \leq 2^{vu}} \sum_{j_1, j_2=1}^p c_{i_1} c_{i_2} \delta_{n, j_1} \delta_{n, j_2} E[X_{i_1, j_1} X_{i_2, j_2}] \right| \\ &\lesssim \sum_{2^{v(u-1)+1} \leq i_1, i_2 \leq 2^{vu}} \sum_{j_1, j_2=1}^p |\delta_{n, j_1} \delta_{n, j_2}| \rho^{|i_1 - i_2|} \\ &\lesssim 2^v \|\delta_n\|_1^2. \end{aligned}$$

Thus, for any $n = 2^d$, we have

$$\left[E \left(\sup_{1 \leq k \leq 2^d} \left| \sum_{i=1}^k X_i^T \delta_n \right| \right)^2 \right]^{1/2} \lesssim d 2^{d/2} \|\delta_n\|_1.$$

For general n , there exists d such that $2^d \leq n < 2^{d+1}$,

$$\begin{aligned} \left[E \left(\sup_{1 \leq k \leq n} \left| \sum_{i=1}^k X_i^T \delta_n \right| \right)^2 \right]^{1/2} &\leq \left[E \left(\sup_{1 \leq k \leq 2^{d+1}} \left| \sum_{i=1}^k X_i^T \delta_n \right| \right)^2 \right]^{1/2} \\ &\lesssim (d+1) \sqrt{2^{d+1}} \|\delta_n\|_1 \leq \sqrt{2n} (\log_2(n) + 1) \|\delta_n\|_1. \end{aligned}$$

□

References

- Ahn, J., Marron, J., Muller, K. M., and Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):760–766.
- Alvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. A., and Matran, C. (2008). Trimmed comparison of distributions. *Journal of the American Statistical Association*, 103(482):697–704.
- Alvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., et al. (2012). Similarity of samples and trimming. *Bernoulli*, 18(2):606–634.
- Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. *The Annals of Mathematical Statistics*, pages 193–212.
- Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H., and Marron, J. (2018). A survey of high dimension low sample size asymptotics. *Australian & New Zealand Journal of Statistics*, 60(1):4–19.
- Aue, A., Hörmann, S., Horváth, L., Reimherr, M., et al. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B):4046–4087.
- Bedo, J. (2008). Microarray design using the hilbert–schmidt independence criterion. In *IAPR International Conference on Pattern Recognition in Bioinformatics*, pages 288–298. Springer.
- Bergsma, W. and Dassios, A. (2014). A consistent test of independence based on a sign covariance related to kendalls tau. *Bernoulli*, 20(2):1006–1028.
- Berrett, T. B. and Samworth, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika*, 106:547–566.
- 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.
- Bickel, P. J. and Breiman, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *The Annals of Probability*, pages 185–214.
- Biswas, M. and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, 123:160–171.
- Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, pages 485–498.
- Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions, Volume 1*. Kendrick Press, Heber City, Utah.
- Brodsky, E. and Darkhovsky, B. S. (2013). *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media.
- Carlstein, E. G., Müller, H.-G., and Siegmund, D. (1994). Change-point problems. IMS.
- Chakraborty, S. and Zhang, X. (2018). Distance metrics for measuring joint dependence with application to causal inference. *Arxiv: <https://arxiv.org/abs/1711.09179>*.

- Chakraborty, S. and Zhang, X. (2019a). A new framework for distance and kernel-based metrics in high dimensions. <https://arxiv.org/abs/1909.13469>.
- Chakraborty, S. and Zhang, X. (2019b). A new framework for distance and kernel-based metrics in high dimensions. *arXiv preprint arXiv:1909.13469*.
- Chen, S. X., Qin, Y.-L., et al. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835.
- Cho, H. et al. (2016). Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics*, 10(2):2000–2038.
- Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.
- Csörgő, S. (1985). Testing for independence by the empirical characteristic function. *Journal of Multivariate Analysis*, 16(3):290–299.
- DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer New York.
- Dau, H. A., Keogh, E., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., and Batista, G. (2018). The ucr time series classification archive. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.
- Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press.
- De Wet, T. (1980). Cramér-von mises tests for independence. *Journal of Multivariate Analysis*, 10(1):38–50.
- Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis*, 11(1):102–113.
- Dette, H. and Gösmann, J. (2019). A likelihood ratio approach to sequential change point detection for a general class of parameters. *Journal of the American Statistical Association*, 0(0):1–17.
- 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.
- Dueck, J., Edelmann, D., Gneiting, T., and Richards, D. (2014). The affinely invariant distance correlation. *Bernoulli*, 20(4):2305–2330.
- Edelmann, D., Richards, D., and Vogel, D. (2017). The distance standard deviation. *arXiv preprint arXiv:1705.05777*.
- Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.
- Freitag, G., Czado, C., and Munk, A. (2007). A nonparametric test for similarity of marginals With applications to the assessment of population bioequivalence. *Journal of Statistical Planning and Inference*, 137(3):697–711.
- Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, pages 697–717.
- Fryzlewicz, P. et al. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281.
- Gallant, A. R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Basil Blackwell.
- Gieser, P. W. and Randles, R. H. (1997). A nonparametric test of independence between two vectors. *Journal of the American Statistical Association*, 92(438):561–567.

- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012b). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.
- Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pages 585–592.
- Hall, P. and Heyde, C. C. (1981). Rates of convergence in the martingale central limit theorem. *The Annals of Probability*, 9(3):395–404.
- Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press.
- Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444.
- Han, F., Chen, S., and Liu, H. (2017). Distribution-free tests of independence in high dimensions. *Biometrika*, 104(4):813–828.
- Heller, R., Heller, Y., and Gorfine, M. (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510.
- Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, pages 772–783.
- Hettmansperger, T. P. and Oja, H. (1994). Affine invariant multivariate multisample sign tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 235–249.
- Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*, pages 546–557.
- Hsing, T., Wu, W. B., et al. (2004). On weighted u-statistics for stationary processes. *The Annals of Probability*, 32(2):1600–1631.
- Hua, W.-Y. and Ghosh, D. (2015). Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics*, 71(3):812–820.
- Jirak, M. et al. (2015). Uniform change point tests in high dimension. *The Annals of Statistics*, 43(6):2451–2483.
- Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130.
- Kiefer, N. M. and Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, pages 1130–1164.
- Kirch, C., Muhsal, B., and Ombao, H. (2015). Detection of changes in multivariate time series with application to eeg data. *Journal of the American Statistical Association*, 110(511):1197–1216.
- Klebanov, L. B., Beneš, V., and Saxl, I. (2005). *N-distances and Their Applications*. Charles University in Prague, the Karolinum Press.
- Kolmogorov, A. N. (1933). *Sulla determinazione empirica di una legge di distribuzione*. NA.
- Kong, J., Klein, B. E., Klein, R., Lee, K. E., and Wahba, G. (2012). Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proceedings of the National Academy of Sciences*, 109(50):20352–20357.
- Kroupi, E., Yazdani, A., Vesin, J.-M., and Ebrahimi, T. (2012). Multivariate spectral analysis for identifying the brain activations during olfactory perception. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6172–6175. IEEE.

- Kroupi, E., Yazdani, A., Vesin, J.-M., and Ebrahimi, T. (2014). Eeg correlates of pleasant and unpleasant odor perception. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(1s):13.
- Laforgia, A. and Natalini, P. (2012). On the asymptotic expansion of a ratio of gamma functions. *Journal of Mathematical Analysis and Applications*, 389(2):833–837.
- Lahiri, S. N., Chatterjee, A., and Maiti, T. (2006). A sub-gaussian berry-esseen theorem for the hypergeometric distribution. *arXiv preprint math/0602276*.
- Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media.
- Leung, D. and Drton, M. (2018). Testing independence in high dimensions with sums of rank correlations. *The Annals of Statistics*, 46(1):280–307.
- Li, J. (2018). Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika*, 105:529–546.
- Li, J., Xu, M., Zhong, P.-S., and Li, L. (2019). Change point detection in the mean of high-dimensional time series data under dependence. *arXiv preprint arXiv:1903.07006*.
- Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.
- Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305.
- 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.
- Matteson, D. S. and Tsay, R. S. (2017). Independent component analysis via distance covariance. *Journal of the American Statistical Association*, 112:623–637.
- Mikalsen, K. Ø., Soguero-Ruiz, C., Bianchi, F. M., and Jenssen, R. (2019). Noisy multi-label semi-supervised dimensionality reduction. *Pattern Recognition*, 90:257–270.
- Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):223–241.
- Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.
- Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3-4):523–527.
- Pan, G., Gao, J., and Yang, Y. (2014). Testing independence among a large number of high-dimensional random vectors. *Journal of the American Statistical Association*, 109(506):600–612.
- Park, T., Shao, X., and Yao, S. (2015). Partial martingale difference correlation. *Electronic Journal of Statistics*, 9:1492–1517.
- Phillips, P. C. and Solo, V. (1992). Asymptotics for linear processes. *The Annals of Statistics*, pages 971–1001.
- Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Sarkar, S., Biswas, R., and Ghosh, A. K. (2018). On high-dimensional modifications of some graph-based two-sample tests. *arXiv preprint arXiv:1806.02138*.
- Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806.

- Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291.
- 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.
- Shao, X. and Zhang, X. (2010). Testing for change points in time series. *Journal of the American Statistical Association*, 105(491):1228–1240.
- Shen, C., Priebe, C. E., and Vogelstein, J. T. (2018). From distance correlation to multiscale graph correlation. *Journal of the American Statistical Association*, (just-accepted):1–39.
- Sinha, B. K. and Wieand, H. (1977). Multivariate nonparametric tests for independence. *Journal of Multivariate Analysis*, 7(4):572–583.
- Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281.
- Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5:1–6.
- Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265.
- 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.
- Székely, G. J. and Rizzo, M. L. (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412.
- Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.
- Taskinen, S., Kankainen, A., and Oja, H. (2003). Sign test of independence between two random vectors. *Statistics & Probability Letters*, 62(1):9–21.
- Tricomi, F. and Erdélyi, A. (1951). The asymptotic expansion of a ratio of gamma functions. *Pacific Journal of Mathematics*, 1(1):133–142.
- Von Mises, R. (1928). Statistik und wahrheit. *Julius Springer*.
- Walck, C. (1996). Hand-book on statistical distributions for experimentalists. Technical report.
- Wang, R. and Shao, X. (2019). Hypothesis testing for high-dimensional time series via self-normalization. *Annals of Statistics*, to appear.
- Wang, R., Volgushev, S., and Shao, X. (2019). Inference for change points in high dimensional data. *arXiv preprint arXiv:1905.08446*.
- Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83.
- Wei, S., Lee, C., Wichers, L., and Marron, J. S. (2016). Direction-projection-permutation for high-dimensional hypothesis tests. *Journal of Computational and Graphical Statistics*, 25(2):549–569.
- Wu, W.-B. (2005a). Nonlinear system theory: another look at dependence. *Proceedings of the National Academy of Sciences USA*, 102:14150–14154.
- Wu, W. B. (2005b). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154.
- Wu, W. B. (2007). Strong invariance principles for dependent random variables. *The Annals of Probability*, 35(6):2294–2320.

- Wu, W. B. and Shao, X. (2004). Limit theorems for iterated random functions. *Journal of Applied Probability*, 41(2):425–436.
- Xie, Y., Huang, J., and Willett, R. (2012). Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27.
- Xu, J., Liu, J., Yin, J., and Sun, C. (2016). A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously. *Knowledge-Based Systems*, 98:172–184.
- Yang, Y. (2017). *Source-Space Analyses in MEG/EEG and Applications to Explore Spatio-temporal Neural Dynamics in Human Vision*. PhD thesis.
- Yang, Y. and Pan, G. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *Annals of Statistics*, 43(2):467–500.
- 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 (Statistical Methodology)*, 80(3):455–480.
- Yata, K. and Aoshima, M. (2010). Effective pca for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of multivariate analysis*, 101(9):2060–2077.
- Zhang, T. and Lavitas, L. (2018). Unsupervised self-normalized change-point testing for time series. *Journal of the American Statistical Association*, 113(522):637–648.
- Zhang, X. and Cheng, G. (2018). Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli*, 24(4A):2640–2675.
- Zhang, X., Yao, S., Shao, X., et al. (2018). Conditional mean and quantile dependence testing in high dimension. *The Annals of Statistics*, 46(1):219–246.
- Zhang, Y. and Zhou, Z.-H. (2010). Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):14.
- Zhou, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, 33(3):438–457.
- Zhu, C. and Shao, X. (2019). Interpoint distance based two sample tests in high dimension. <https://arxiv.org/pdf/1902.07279.pdf>.
- Zhu, C., Yao, S., Zhang, X., and Shao, X. (2019). Distance-based and rkhs-based dependence metrics in high dimension. *arXiv preprint arXiv:1902.03291*.
- Zhu, L., Xu, K., Li, R., and Zhong, W. (2017). Projection correlation between two random vectors. *Biometrika*, 104(4):829–843.