Indian Statistical Institute

Doctoral Thesis

On Tests of Independence among Multiple Random Vectors of Arbitrary Dimensions

Author:
Angshuman Roy

Supervisor:
Prof. Anil K. Ghosh

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the

Applied Statistics Unit
Applied Statistics Division

April 20, 2020

"*If I have seen further than others, it is by standing upon the shoulders of giants.*"

Sir Isaac Newton

Acknowledgements

They say that 'life is a journey'. On the eve of submitting my doctoral thesis, when I recollect my memories, I come to the conclusion that my life as a Ph.D. student is a long eventful journey with numerous ups and downs. I take this opportunity to express my gratitude to those persons to whom I am much indebted.

First and foremost, I would like to thank my late father whom I lost during my Ph.D. years. He provided me constant encouragement and tried his best to ensure a suitable environment for me to pursue my Ph.D. as long as he was with us. Next, I would like to thank my mother for her tireless support. Furthermore, I would like to thank my sister and brother-in-law for their kind help in my troubled days. I am thankful to my adorable niece for bringing smiles on my face at the end of my stressful days.

In my school life, I met with Dr. Mrinal Nandi, who is an amazing teacher. He is one of the reasons I started loving mathematics and statistics. I am grateful to him for his encouragement and guidance that immensely helped me to get admission to the Indian Statistical Institute. In my bachelors and masters courses, I was taught by many talented and distinguished professors. Their teaching inspired me to grow keen interest in the topics of statistics and probability. I would like to thank them, especially Prof. Anil K. Ghosh, Prof. Arnab Chakraborty, Prof. Subir K. Bhandari, Prof. Bimal K. Roy and late Prof. Kamal K. Roy, for their amazing tutoring.

I started my research under the supervision of late Prof. C. A. Murthy. He was a distinguished researcher and a very generous person. I am immensely thankful to him for introducing me to the field of research. I would like to thank Prof. Alok Goswami for providing me worthwhile suggestions and helping me in the mathematical aspects of my research works. Next, I would like to express my gratitude towards Prof. Sourabh Bhattacharya and Prof. Gopal K. Basak for assisting me in many research-related issues. My sincere gratitude goes to dear Prof. Anil K. Ghosh, who started supervising me after Prof. Murthy's tragic demise. His tireless efforts, valuable suggestions, productive subject discussions helped me tremendously to achieve my goals. I am indebted to him forever.

Thanks to Soham, Biswadeep and Kunal for all useful academic discussions. I would also like to thank the editors, associate editors and reviewers of different journals for proving me helpful comments and suggestions. Thanks to my dear colleagues Noirrit, Adhideb,

Joydeep, Jayabrata, Abhishek with whom I spent my spare moments. I am grateful to Research Fellow Advisory Committee of the Applied Statistics Division, Deans and Directors of Indian Statistical Institute, all members of the Applied Statistics Unit and the Theoretical Statistics & Mathematics Unit for providing me nice academic atmosphere. Thanks to Computer & Statistical Service Center for providing me necessary computing facilities. Finally, I would like to thank my nearest and dearest friends Titir and Prithwish, who were always beside me even in my darkest days and kept me motivated with their kind generous emotional support.

Contents

List of Figures

List of Tables

List of Abbreviations

AR(1)	**A**uto **R**egressive model of order **1**
a.s.	**a**lmost **s**urely
CCPP	**C**ombined **C**ycle **P**ower **P**lant
DAG	**D**iscrerte **A**cyclic **G**raph
dCov	**D**istance **Cov**ariance
DCT	**D**ominated **C**onvergence **T**heorem
FDA	**F**unctional **D**ata **A**nalysis
FDR	**F**alse **D**iscovery **R**ate
HHG	Test proposed by **H**eller **H**eller **G**orfine
HSIC	**H**ilbert **S**chmidt **I**ndependence **C**riterion
i.i.d.	**i**ndependent and **i**dentically **d**istributed
JdCov	**J**oint **D**istance **Cov**ariance
MMD	**M**aximum **M**ean **D**iscrepancy
MST	**M**inimum **S**panning **T**ree
SEM	**S**tructural **E**quation **M**odel
SLLN	**S**trong **L**aw of **L**arge **N**umbers

List of Symbols

$\|\cdot\|$	norm.
$\langle\cdot,\cdot\rangle$	Inner product.
$\sim$	Distributed as.
$\overset{\text{a.s.}}{\to}$	Converges almost surely.
$\overset{\mathcal{D}}{\to}$	Converges in distribution.
$\overset{\mathcal{D}}{=}$	Equal in distribution.
$\overset{\text{i.i.d.}}{\sim}$	Independently and identically distributed as.
$\overset{\text{Pr}}{\to}$	Converges in probability.
α	Level of significance.
$\gamma_K(P,Q)$	MMD between P and Q computed using the kernel K.
$\gamma_{K_\sigma}(P,Q)$	MMD between P and Q computed using the Gaussian kernel with bandwidth σ.
ζ_n	Statistics for the test based on pairwise distances.
$\zeta_{\text{sum},n}$	Multi-scale analogue of ζ_n based on summation of the test statistics.
$\zeta_{\text{max},n}$	Multi-scale analogue of ζ_n based on maximum of the test statistics.
$\widetilde{\zeta}_n$	Statistics for the test based on linear projections.
$\widetilde{\zeta}_{\text{sum},n}$	Multi-scale analogue of $\widetilde{\zeta}_n$ based on summation of the test statistics.
$\widetilde{\zeta}_{\text{max},n}$	Multi-scale analogue of $\widetilde{\zeta}_n$ based on maximum of the test statistics.
π	Permutation of $\{1,2,\ldots,n\}$.
Π	Product copula/ uniform copula.
Π_n	Empirical version of product copula/ uniform copula.
σ_n	Bandwidth chosen (using median heuristic) based on n observations.
σ_0	A positive real number such that $2\sigma_0^2$ is the median of $\|\mathbf{Z}-\mathbf{Z}_*\|^2$, where $\mathbf{Z},\mathbf{Z}_*\sim\Pi$.
$\sigma^{(i)}$	i-th bandwidth used for multi-scale methods.
ϕ	Density function of the standard normal distribution.
Φ	Cumulative distribution function of the standard normal distribution.

$\mathbf{a}$	A vector in $\mathbb{R}^d$.
$\mathbf{a}^{(j)}$	j^{th} sub-vector (of dimension d_j) of $\mathbf{a}$.
$\mathcal{B}$	Banach space.
$B(\mathbf{c}, r)$	An open ball of radius $r > 0$ with center $\mathbf{c}$ i.e. $B(\mathbf{c}, r) = \{\mathbf{u} : \|\mathbf{u} - \mathbf{c}\| < r\}$.
C	Copula distribution of $\mathbf{X}$.
C^{Π}	Checkerboard copula distribution of $\mathbf{X}$.
C_n	Empirical copula based on n observations.
C_n^{Π}	Empirical checkerboard copula based on n observations.
c_n^{Π}	Empirical checkerboard copula density based on n observations.
Cov	Covariance.
Cor	Correlation.
$C_{\sigma,p}$	A constant that depends on σ and p such that $C_{\sigma,p} = \gamma^2_{K_\sigma}(\mathsf{M}, \Pi)$.
d	Dimension of $\mathbf{X}$.
d_j	Dimension of $\mathbf{X}^{(j)}$.
E	Expectation.
F	Distribution function of $\mathbf{X}$.
F_j	Distribution function of $\mathbf{X}^{(j)}$.
$F_j(t^-)$	Left limit of F_j at t.
$F_j^{\mathbf{a}}$	Distribution of $X^{(\mathbf{a},j)}$.
$F^{\mathbf{a}}$	Joint distribution function of $\left(X^{(\mathbf{a},1)}, X^{(\mathbf{a},2)}, \ldots, X^{(\mathbf{a},p)}\right)$.
$\mathbb{F}_k^{(q)}$	Distribution of $\mathbf{R}_C^{(q)}(i,k)$.
$\mathbb{H}_0$	Null hypothesis.
$\mathbb{H}_1$	Alternative hypothesis.
$\mathbb{I}[\cdot]$	Indicator function.
$\mathbf{I}_d$	Identity matrix of dimension d.
$I_\sigma(\mathbf{X})$	Copula based measure of dependence among the sub-vectors of $\mathbf{X}$ computed using the bandwidth σ.
$\widehat{I}_{\sigma,n}(\mathbf{X})$	Estimate (empirical version) of $I_\sigma(\mathbf{X})$ based on n observations.
$I_\sigma^{\Pi}(\mathbf{X})$	Checkerboard copula based measure of dependence among the sub-vectors of $\mathbf{X}$ computed using the bandwidth σ.
$\widehat{I}_{\sigma,n}^{\Pi}(\mathbf{X})$	Estimate (empirical version) of $I_\sigma^{\Pi}(\mathbf{X})$ based on n observations.
K	A symmetric bounded positive definite kernel.

K_σ	The Gaussian kernel with bandwidth σ i.e., $K_\sigma(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$.
$\mathcal{L}(\mathbf{X})$	The law or distribution of $\mathbf{X}$.
$\mathcal{L}_2[0,1]$	The space of square integrable functions on $[0,1]$.
M	Maximum copula.
M_n	Empirical version of maximum copula.
m	Number of bandwidths used for multi-scale methods.
n	Sample size.
N	Number of one-dimensional projections.
$N(\mu,\sigma^2)$	Normal distribution with mean μ and variance σ^2.
$N_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$	d-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
p	Number of sub-vectors of $\mathbf{X}$.
p_i	p-value corresponding to the test based on $\sigma^{(i)}$.
$p_{(i)}$	i-th order statistic among $p_1, p_2, \ldots, p_m$.
Pr	Probability.
P, Q	Probability distributions.
$P \otimes Q$	Product measure of P and Q.
$\mathbf{PA}^{(j)}$	The set of all parent nodes of $\mathbf{X}^{(j)}$ in a DAG representing an SEM.
$\mathbf{r}_i$	Multivariate coordinate-wise rank of $\mathbf{x}_i$ among $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.
$\mathbb{R}$	Set of real numbers.
$\mathbb{R}^d$	d-dimensional Euclidean space.
$R^{(s\|q)}(i,k)$	$\mathbf{X}^{(s)}$-rank of k-th $\mathbf{X}^{(q)}$-neighbor of $\mathbf{X}_i$.
$\mathbf{R}_C^{(q)}(i,k)$	Coordinate-wise rank of k-th $\mathbf{X}^{(q)}$-neighbor of $\mathbf{X}_i$.
$\mathbf{R}_S^{(q)}(i,k)$	Spatial rank of k-th $\mathbf{X}^{(q)}$-neighbor of $\mathbf{X}_i$.
$\mathbf{Sign}_d(\cdot)$	d-dimensional spatial sign function.
T_n	Copula based test statistics for testing independence among continuous random variables.
$T_{\text{sum},n}$	Multi-scale analogue of T_n based on summation of the test statistics.
$T_{\text{max},n}$	Multi-scale analogue of T_n based on maximum of the test statistics.
$T_n^{\maltese}$	Checkerboard copula based test statistics for testing independence among random variables.
$T_{\text{sum},n}^{\maltese}$	Multi-scale analogue of $T_n^{\maltese}$ based on summation of the test statistics.
$T_{\text{max},n}^{\maltese}$	Multi-scale analogue of $T_n^{\maltese}$ based on maximum of the test statistics.

$T^U_{\text{sum},n}, T^U_{\text{max},n}$	Test statistics based on univariate ranks of neighbors.
$T^C_{\text{sum},n}, T^C_{\text{max},n}$	Test statistics based on coordinate-wise ranks of neighbors.
$T^S_{\text{sum},n}, T^S_{\text{max},n}$	Test statistics based on spatial ranks of neighbors.
$T^M_{\text{sum},n}, T^M_{\text{max},n}$	Test statistics based on MMD and ranks of neighbors.
$\mathcal{T}$	Functional on the set of probability distributions on $\mathbb{R}^p$.
$\mathcal{T}_n(F)$	An estimator of $\mathcal{T}(F)$ based on n independent observations.
$\mathbb{T}_\sigma(G)$	$\mathbb{T}_\sigma(G) = I_\sigma(\mathbf{Z})$, where $\mathbf{Z} \sim G$.
$\mathbb{T}_{\sigma,n}(G)$	$\mathbb{T}_{\sigma,n}(G) = \widehat{I}_{\sigma,n}(\mathbf{Z})$, where $\mathbf{Z} \sim G$.
t_ν	Student's t distribution with ν degrees of freedom.
$U(a,b)$	Uniform distribution on $[a,b]$.
$\mathbb{U}^d_n$	Uniform probability distribution on $\{1,2,\ldots,n\}^d$.
Var	Variance.
$\mathbf{X}$	Random vector consisting of several sub-vectors.
$\mathbf{X}^{(j)}$	j^{th} sub-vector of $\mathbf{X}$.
$\mathbf{X}^{(-j)}$	Collection of $(p-1)$ sub-vectors of $\mathbf{X}$ excluding $\mathbf{X}^{(j)}$.
$X^{(j)}$	j^{th} sub-vector of $\mathbf{X}$ when the sub-vector is one-dimensional.
$X^{(\mathbf{a},j)}$	Random variable defined as $\|\mathbf{X}^{(j)} - \mathbf{a}^{(j)}\|$.
$\widetilde{X}^{(\mathbf{a},j)}$	Random variable defined as $\langle \mathbf{a}^{(j)}, \mathbf{X}^{(j)} \rangle$.
$\mathbf{x}_i$	i^{th} observation on $\mathbf{X}$.
$\mathbf{x}_i^{(j)}$	i^{th} observation on $\mathbf{X}^{(j)}$.
$x_i^{(j)}$	i^{th} observation on $X^{(j)}$.
$\lceil x \rceil$	The smallest integer greater than or equal to x.
$\lfloor x \rfloor$	The largest integer less than or equal to x.
$\mathbf{y}_i$	Normalized rank vector $\mathbf{r}_i$, i.e., $\mathbf{y}_i = \mathbf{r}_i/n$.
$y_i^{(j)}$	j^{th} coordinate of $\mathbf{y}_i$.
$\mathbf{0}_d$	d-dimensional vector with all elements equal to 0.
$\mathbf{1}_d$	d-dimensional vector with all elements equal to 1.

Dedicated to my parents

Late Jogendranath Roy

and

Tapti Roy

Chapter 1

Introduction

Measures of dependence among several random vectors and associated tests of independence play a major role in different statistical applications. Blind source separation or independent component analysis (see, e.g., Hyvärinen *et al.*, 2001; Shen *et al.*, 2009), feature selection and feature extraction (see, e.g., Li *et al.*, 2012), detection of serial correlation in time series (see, e.g., Ghoudi *et al.*, 2001) and finding the causal relationships among the variables (see, e.g., Chakraborty and Zhang, 2019) are some examples of their wide-spread applications. Tests of independence has vast applications in other areas of sciences as well. For instance, to characterize the genetic mechanisms of a complex disease, a biologist or a medical scientist often needs to carry out some tests of independence to investigate the causal relationship among multiple quantitative traits and test for their association with disease genes (see, e.g., Hsieh *et al.*, 2014). Proper understanding of the structure of dependence among several groups of variables often helps a psychologist or social scientist to construct a meaningful structural equation model (see, e.g., De Jonge *et al.*, 2001) for data analysis. In order to develop a micro-economic model for health care and health insurance, an economist needs to study the dependence (or independence) between several measures of health-care utilization and the insurance status of the house-hold for a variety of socio-economic and health-status variables (see, e.g. Cameron *et al.*, 1988).

In this thesis, we deal with this problem of testing independence among several random vectors. This is a well-known problem in statistics and machine leaning literature, and several methods of are available for it. But most of these existing methods deal with two random vectors (or random variables) only. Moreover, instead of testing for independence, many of them only test for uncorrelatedness between two vectors. Now a days, we often deal with data sets having dimension larger than sample size. Many existing tests cannot be used in such situations. Keeping all these in mind, in this thesis, we propose and investigate

some methods that can be used for testing independence among several random vectors of arbitrary dimensions. Later we shall see that these proposed tests can also be used for testing independence among several random functions or random elements taking values in infinite dimensional Banach or Hilbert spaces.

Consider a d-dimensional random vector $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)})$ with sub-vectors $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ of dimensions $d_1, d_2, \ldots, d_p$, respectively $(d_1 + d_2 + \ldots + d_p = d)$. Suppose that we have n independent observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ on $\mathbf{X}$, and based on these observations, we need to construct a test for independence among the sub-vectors $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$. This is a well studied problem in statistics, especially for $p = 2$ and $d_1 = d_2 = 1$. Pearson's product moment correlation coefficient is arguably the most simplest and popular measure of association between two random variables, and one can easily construct a test of independence based on this measure (see, e.g., Anderson, 2003). But this product moment correlation coefficient only measures the degree of linear relationship between two variables, and it gets severely affected by the presence of outliers in the data. Popular rank-based measures like Spearman's rank correlation coefficient ρ (Spearman, 1904), Kendall's concordance-discordance statistic τ (Kendall, 1938) and Blomqvist's quadrant statistic β (Blomqvist, 1950) are robust against outliers and extreme values. The tests based on these statistics have the distribution-free property as well, but instead of independence, they only test for monotone relationship between two variables. Hoeffding (1948) constructed a statistic, known as the Φ-statistic, based on empirical distribution function to measure the dependence between two random variables. The distribution-free test constructed based on this statistic was probably the first attempt in the literature to actually test for independence (not uncorrelatedness) between two continuous random variables. Rényi (1959) postulated seven properties for an appropriate measure of dependence and showed that there exists a unique dependency measure, namely maximal correlation coefficient, that satisfies all those properties. However, that dependency measure cannot be computed in practice. Lopez-Paz *et al.* (2013) proposed a measure, called randomized dependence coefficient, that can estimate maximal correlation coefficient with a given precision. But, in addition to randomness, it involves several hyper-parameters, which makes it less attractive. Reshef *et al.* (2011) developed a measure of dependence based on mutual information, but Simon and Tibshirani (2014) noted that the power of the test based on that measure falls rapidly as noise increases.

Several attempts have also been made for measuring dependence among several random variables and testing for statistical significance of that measure. Joe (1990), Nelsen (1996) and Schmid and Schmidt (2007) proposed generalizations of Spearman's ρ and Kendall's τ statistics in this context. Úbeda-Flores (2005) generalized the notion of Blomqvist's β statistics. Gaißer *et al.* (2010) generalized Hoeffding's Φ-statistic for more than two variables. These rank-based generalized versions can also be viewed as copula based measures of dependence. However, these measures are not invariant under strictly monotone transformations of the variables; they are invariant only under the same type of transformation (either strictly increasing or strictly decreasing) in all coordinates. Using the idea of copula and kernel embedding of probability distributions, Póczos *et al.* (2012) proposed two dependency measures and associated tests. But, their proposed choices of the cut-offs based on probability inequalities made these tests very conservative. Motivated by the idea of Póczos *et al.* (2012), in Chapter 2 of this thesis, we propose a new measure of dependence, which is invariant under permutations and strictly monotone transformations of the variables. To construct this measure, we use the Gaussian kernel, which helps us to get a nice closed form expression for its empirical version. So, unlike Póczos *et al.* (2012), one does not need to generate observations from a uniform distribution for computing its data-based estimate. We use this measure to construct a distribution-free test. However, for the implementation of the test, we need to choose the bandwidth parameter associated with the Gaussian kernel. The method commonly used for choosing the bandwidth is based on median heuristic (see, e.g., Fukumizu *et al.*, 2009b, Sec. 5). But, this may not always be the best choice, and depending on the data set, sometimes other choices of the bandwidth may lead to better results. In order to take care of this problem, we adopt a multi-scale approach, where we look at the results for various choices of the bandwidth and then aggregate them judiciously to arrive at the final decision. We propose several methods for aggregation and prove the consistency of the resulting tests under appropriate regularity conditions. Several simulated and real data sets are analyzed to demonstrate the utility of these proposed methods. The contents of this chapter are taken from Roy *et al.* (2020b).

It is to be noted that all rank-based or copula based tests mentioned in the previous paragraph including our proposed ones need the underlying variables to be continuous so that ties occur with zero probability, and the ranks can be uniquely defined with probability one. However, in practice, we often encounter data comprise of a mixture of continuous,

discrete, ordinal and binary variables. Even the observations on a continuous variable may have ties due to limited precision. To cope with such situations, recently Genest *et al.* (2019) developed a test, which is applicable to random variables having arbitrary probability distributions. They used checkerboard copula for generalization of Hoeffding's Φ-statistic for more than two variables and developed a test based on it. But just like the Hoeffding's Φ-statistic, neither this measure nor the resulting test is invariant under strictly monotone transformations of the variables. Moreover, this test often fails to perform well in the presence of complex non-monotone relationships among the variables. In order to take care of this issue, in Chapter 3, we propose a new measure of association among several random variables and develop some tests based on it. These proposed measure and the associated tests are invariant under permutations and strictly monotone transformations of the variables, and they can be viewed as the checkerboard copula versions of the same proposed in Chapter 2. We establish consistency of these proposed tests and demonstrate their usefulness using empirical study. This chapter is mainly based on Roy (2020).

The methods discussed so far deal with two or more random variables. But there are several methods in the literature that deal with random vectors of dimensions higher than one. If we assume normality of the underlying distribution, the likelihood ratio test based on Wilks' Λ statistic (see, e.g., Anderson, 2003) can be used for testing independence between two random vectors. One can also use Roy's largest root test, Pillai-Bartlett trace test or Hotelling-Lawley trace test (see, e.g., Anderson, 2003) for this purpose. Other popular tests of independence between two random vectors include the tests based on coordinate wise signs and ranks (see, e.g., Sen and Puri, 1971), spatial signs and ranks (see, e.g., Taskinen *et al.*, 2003, 2005) and inter-directions (see, e.g., Gieser and Randles, 1997). However, these tests are mainly motivated by the elliptic symmetry (see. e.g., Fang *et al.*, 1990) of the underlying distribution, and they actually test for uncorrelatedness between two multivariate sign or rank vectors. So, they are not very useful for detecting complex relationships between two sub-vectors. Moreover, none of these tests can be used if the dimension of one of the sub-vectors exceeds the sample size.

Székely *et al.* (2007) developed a test of independence (known as the dCov test) based on distance covariance (dCov) or distance correlation, which can be used even for vectors with dimensions larger than sample size. Gretton *et al.* (2008) constructed a test based on the Hilbert-Schmidt norm of the covariance kernel. It test is known as the Hilbert Space

independence criterion (HSIC) test. A test based on 2×2 contingency tables of pairwise distances (known as the HHG test) was proposed by Heller *et al.* (2013). Some graph-based tests of independence between two random vectors were proposed by Friedman and Rafsky (1983); Heller *et al.* (2012); Biswas *et al.* (2016); Sarkar and Ghosh (2018).

The method based on Wilk's Λ statistic can be generalized for testing independence between several random vectors. Similarly, the tests based on coordinate-wise signs and ranks or those based on spatial signs and ranks can also be generalized. Um and Randles (2001) generalized Gieser and Randles (1997)'s test for multiple random vectors. But these tests have the same limitations as discussed before. Bilodeau and Lafaye de Micheaux (2005) proposed a test of independence among several normally distributed random vectors, whose joint distribution may not be normal. A test based on half-spaces was proposed by Beran *et al.* (2007), but its computing cost grows up exponentially as the dimension increases. Bilodeau and Nangue (2017) developed some methods for testing mutual and serial independence among several random vectors. Recently, some generalizations of the dCov test (see Fan *et al.*, 2017; Jin and Matteson, 2018; Chakraborty and Zhang, 2019) and the HSIC test (see Pfister *et al.*, 2018) have been proposed in the literature. The distance multivariance measure proposed by Böttcher *et al.* (2019) can also be viewed as a generalization of the distance correlation measure (see Székely *et al.*, 2007), and it can be used to construct a test of independence among several random vectors.

In Chapters 4 and 5 of this thesis, we propose and investigate some methods for testing independence among several random vectors of arbitrary dimensions. We have seen that there are methods for testing independence among several random variables. In Chapter 4, we propose two common recipes, one based on linear projections and the other based on pairwise distances, for their multivariate extensions. In both cases, we transform the observations on sub-vectors into univariate observations, and then use the existing tests on the transformed data. Heller and Heller (2016) also used somewhat similar strategies for constructing multivariate tests of independence, but their proposed tests were restricted to two random vectors only. We use our recipes on the copula based tests proposed in Chapter 2, and investigate the theoretical as well as the empirical performance of the resulting tests. Materials of this chapter are taken from Roy *et al.* (2020c).

Chapter 5 deals with some tests based on nearest neighbors. Researchers have observed that dependence between two random vectors is often manifested by a strong positive or

negative association between their respective pairwise distances (see, e.g., Friedman and Rafsky, 1983; Heller *et al.*, 2012; Biswas *et al.*, 2016; Sarkar and Ghosh, 2018). Based on this idea, Sarkar and Ghosh (2018) proposed some tests of independence between two random vectors ($\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, say), where they suggested to find the neighbors of an observation based on pairwise distances in one space ($\mathbf{X}^{(1)}$-space, say) and compute the ranks of these neighbors based on corresponding pairwise distances in the other space ($\mathbf{X}^{(2)}$-space, say). In Chapter 5, we propose some generalizations these tests so that one can deal with more than two random vectors. Most of these generalizations are based on multivariate rank functions, and some of them use the idea of maximal mean discrepancy (MMD) as well. Empirical performance of these tests are investigated by analyzing several simulated and real data sets. The contents of this chapter are taken from Roy *et al.* (2020a) and Roy and Ghosh (2020).

In Chapters 4 and 5, we also briefly consider the problem of testing independence between two or more random functions. The branch of statistics that deals with function valued data is referred to as Functional Data Analysis (FDA) (see, e.g. Ramsay and Silverman, 2005; Ferraty and Vieu, 2006), and it is gaining momentum over the last couple of decades. Real world applications of FDA is as diverse as hand writing recognition, speech recognition, spectometry and meteorology to name a few. But, the literature on testing of independence between two random functions is almost non-existent. Lyons (2013) generalized the notion of the distance correlation for random functions having distribution on metric spaces of strong negative type (e.g., the Hilbert space of square integrable functions on $[0, 1]$), and hence generalized the dCov test (Székely *et al.*, 2007) for testing independence between two random functions. The tests we proposed in Chapters 4 and 5 can be used for such functions as well. So, we analyze some functional data sets to evaluate their performance. We also consider some generalizations of the joint distance covariance (JdCov) tests (Chakraborty and Zhang, 2019) and the dHSIC test (Pfister *et al.*, 2018) for comparison. A generalization of the HHG test (Heller *et al.*, 2013) based on contingency tables is also considered.

Finally, Chapter 6 contains a brief summary of our contributions and a comparative discussion on the performance of our proposed tests. The thesis ends with some discussions on possible directions for future research.

Chapter 2

Tests of Independence among Continuous Random Variables

Here we consider all sub-vectors to be one dimensional, i.e, $d_1 = d_2 = \cdots = d_p = 1$ and $d = p$. So, instead of using vector notations $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ for these sub-vectors, in this chapter, we denote these sub-vectors by $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$, respectively. As we have mentioned before, for $p = 2$, Pearson's product moment correlation coefficient (see, e.g., Anderson, 2003), Spearman's rank correlation coefficient (Spearman, 1904), Kendall's concordance-discordance statistic (Kendall, 1938) or Blomqvist's quadrant statistic β (Blomqvist, 1950) can be used to measure the dependence between two random variables and to construct a test of independence. But these measures and the resulting tests are mainly useful for detecting linear or monotone relationships between the variables. Hoeffding (1948) also constructed a distribution-free test based on the Φ-statistic. All these measures of dependence and the associates tests have been generalized for more than two random variables as well (see, e.g., Joe, 1990; Nelsen, 1996; Úbeda-Flores, 2005; Gaißer *et al.*, 2010). But none of them are invariant under strictly monotone transformations of the variables. Using the idea of copula and MMD, Póczos *et al.* (2012) developed two dependency measures and related tests. Motivated by their work, here we propose a new measure of dependence, which has this invariance property. The description of the measure is given below.

2.1 The proposed measure of dependence

Our measure of dependence is based on the copula distribution of a p-dimensional random vector. A p-dimensional copula is a probability distribution on the p-dimensional unit

hypercube $[0,1]^p$ such that all of its one-dimensional marginals are uniform on $[0,1]$. Assume that the $X^{(i)}$'s $(i = 1, 2, \ldots, p)$ are continuous random variables, and F is the joint distribution of $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(p)})$. The copula transformation of F or the copula distribution of $\mathbf{X}$ is then given by

$$\mathsf{C}(\mathbf{u}) = F\left(F_1^{-1}(u^{(1)}), F_2^{-1}(u^{(2)}), \ldots, F_p^{-1}(u^{(p)})\right),$$

where $\mathbf{u} = (u^{(1)}, u^{(2)}, \ldots, u^{(p)}) \in [0,1]^p$ and $F_i^{-1}(u^{(i)}) = \inf\{x : F_i(x) > u^{(i)}\}$ for all $i = 1, 2, \ldots, p$ (see, e.g., Nelsen, 2007, for further discussion on copula). If C is the cumulative distribution function of a uniform distribution on $[0,1]^p$, i.e., $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$ are independent, it is called the uniform copula or the product copula, and it is denoted by Π. On the other hand, if $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$ are comonotonic, i.e. there exist strictly increasing functions f_i's and a random variable V such that $\mathbf{X} \overset{\mathcal{D}}{=} (f_1(V), f_2(V), \ldots, f_p(V))$, it is called the maximum copula and denoted by M. So, for every $\mathbf{u} \in [0,1]^p$, we have $\Pi(\mathbf{u}) = \prod_{i=1}^p u^{(i)}$ and $\mathsf{M}(\mathbf{u}) = \min\{u^{(1)}, u^{(2)}, \ldots, u^{(p)}\}$.

Naturally, larger difference between C and Π indicates higher degree of dependence among $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$. To measure the difference between two probability distributions P and Q on $\mathbb{R}^p$, we use

$$\gamma_K(P, Q) = \left[\mathrm{E}K(\mathbf{Y}, \mathbf{Y}_*) + \mathrm{E}K(\mathbf{Z}, \mathbf{Z}_*) - 2\mathrm{E}K(\mathbf{Y}, \mathbf{Z})\right]^{\frac{1}{2}}, \tag{2.1}$$

where $\mathbf{Y}, \mathbf{Y}_* \overset{i.i.d.}{\sim} P$, $\mathbf{Z}, \mathbf{Z}_* \overset{i.i.d.}{\sim} Q$ are independent, and $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a symmetric, bounded, positive definite kernel. This measure is also called the maximum mean discrepancy (MMD) between P and Q (see, e.g., Gretton *et al.*, 2012). It is known that γ_K is a pseudo-metric on the space of all probability distributions on $\mathbb{R}^p$, and it is a metric when K is a characteristic kernel (see, e.g., Fukumizu *et al.*, 2009b). Gaussian kernel $K_\sigma(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$ with a bandwidth parameter $\sigma > 0$ is a popular choice as a characteristic kernel, and we shall use it throughout this thesis.

From the above discussion, it is clear that for any characteristic kernel K on $[0,1]^p \times [0,1]^p$, one can use $\gamma_K(\mathsf{C}, \Pi)$ or $\gamma_K^2(\mathsf{C}, \Pi)$ as a measure of dependence. In this thesis, we use a scaled version of this measure given by

$$I_\sigma(\mathbf{X}) = \gamma_{K_\sigma}(\mathsf{C}, \Pi)/\gamma_{K_\sigma}(\mathsf{M}, \Pi),$$

where $\gamma_{K_\sigma}(P, Q)$ denotes the MMD between two probability distributions P and Q computed using the Gaussian kernel with bandwidth σ. Note that the denominator $\gamma_{K_\sigma}(\mathsf{M}, \Pi)$

is strictly positive. So, $I_\sigma(\mathbf{X})$ is well defined. The use of the Gaussian kernel makes the measure $I_\sigma(\mathbf{X})$ invariant under permutations and strictly monotone transformations of the coordinate variables. This result is stated below.

Proposition 2.1. *$I_\sigma(\mathbf{X})$ is invariant under permutations and strictly monotone transformations of $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$.*

From the definition of $I_\sigma(\mathbf{X})$, it is clear that it takes the value 0 if and only if the coordinates of $\mathbf{X}$ are independent, and its value is supposed to increase as the dependence among $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$ increases. The following proposition shows that in case of extreme dependence (i.e., when for each pair of variables, one is a strictly monotone function of the other), it turns out to be 1.

Proposition 2.2. *For all $i = 2, 3, \ldots, p$, if $X^{(i)}$ is almost surely a strictly monotone function of $X^{(1)}$, then $I_\sigma(\mathbf{X})$ takes the value 1.*

This desirable property of $I_\sigma(\mathbf{X})$ helps us to properly assess the degree of dependence among $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$. Note that many well-known dependency measures like the copula based multivariate extensions of Spearman's ρ, Kendall's τ, Blomqvist's β and Hoeffding's ϕ statistics (see, e.g., Úbeda-Flores, 2005; Nelsen, 1996; Gaißer *et al.*, 2010) do not have this property unless all monotone functions considered in Proposition 2.2 are strictly increasing. We know that the distance correlation measure proposed by Székely *et al.* (2007) can be expressed as a weighted squared distance between the characteristic functions of two distributions. The following theorem shows that $I_\sigma(\mathbf{X})$ also has a similar property.

Theorem 2.1. *Let φ_{C} and φ_Π be the characteristic functions of C and Π, respectively. Define $C_{\sigma,p} = \kappa\left(\frac{\sigma}{\sqrt{p}}\right) + \kappa^p(\sigma) - 2\int_0^1 \lambda^p(u,\sigma)\,du$, where $\kappa(\sigma) = \sqrt{2\pi}\sigma\left[2\Phi\left(\frac{1}{\sigma}\right) - 1\right] - 2\sigma^2\left[1 - \exp\left(-\frac{1}{2\sigma^2}\right)\right]$, $\lambda(x,\sigma) = \sqrt{2\pi}\sigma\left[\Phi\left(\frac{x}{\sigma}\right) + \Phi\left(\frac{1-x}{\sigma}\right) - 1\right]$ and $\Phi(\cdot)$ is the cumulative distribution function of the N(0,1) distribution. Then $I^2_\sigma(\mathbf{X})$ can be expressed as*

$$I^2_\sigma(\mathbf{X}) = C^{-1}_{\sigma,p} \cdot \left(\frac{\sigma}{\sqrt{2\pi}}\right)^p \int_{\mathbb{R}^p} |\varphi_{\mathsf{C}}(\mathbf{w}) - \varphi_\Pi(\mathbf{w})|^2 \exp\left(-\frac{\sigma^2}{2}\mathbf{w}^\intercal\mathbf{w}\right) d\mathbf{w}.$$

Another interesting property of $I_\sigma(\mathbf{X})$ is its continuity. Note that if $\{\mathbf{X}_n; n \geq 1\}$ is a sequence of random vectors converging in distribution to $\mathbf{X}$, then $\mathsf{C}_{\mathbf{X}_n}$ (the copula

distribution of $\mathbf{X}_n$) converges to C weakly. So, using the dominated convergence theorem (DCT), from Theorem 2.1 it follows that $I_\sigma(\mathbf{X}_n)$ converges to $I_\sigma(\mathbf{X})$ as n increases. This result is stated below.

Proposition 2.3. *Let $\{\mathbf{X}_n : n \geq 1\}$ be a sequence of p-dimensional random vectors with continuous one-dimensional marginals. If $\mathbf{X}_n$ converges to $\mathbf{X}$ weakly, then we have $\lim_{n\to\infty} I_\sigma(\mathbf{X}_n) = I_\sigma(\mathbf{X})$.*

In the case of $p = 2$, $I_\sigma(\mathbf{X})$ enjoys some additional properties. For instance, $I^2_\sigma(\mathbf{X})$ can be viewed as a product moment correlation coefficient between two random quantities. If $\mathbf{X}$ follows a bivariate normal distribution with correlation coefficient r, $I_\sigma(\mathbf{X})$ turns out to be a strictly increasing function of $|r|$. These results are asserted by the following theorem.

Theorem 2.2. *Let $\mathbf{X} = (X^{(1)}, X^{(2)})$ be a bivariate random vector with continuous one-dimensional marginals.*

(*a*) *Suppose that $\mathbf{T}_1 = (T_1^{(1)}, T_1^{(2)})$ and $\mathbf{T}_2 = (T_2^{(1)}, T_2^{(2)})$ are independent, and they follow the distribution C, the copula distribution of $\mathbf{X}$. For $i = 1, 2$, define*

$$V^{(i)} = K_\sigma(T_1^{(i)}, T_2^{(i)}) - \mathrm{E}\left[K_\sigma(T_1^{(i)}, T_2^{(i)}) \middle| T_1^{(i)}\right] - \mathrm{E}\left[K_\sigma(T_1^{(i)}, T_2^{(i)}) \middle| T_2^{(i)}\right] + \mathrm{E}K_\sigma(T_1^{(i)}, T_2^{(i)}).$$

Then, we have $I^2_\sigma(\mathbf{X}) = \mathrm{Cor}(V^{(1)}, V^{(2)})$, which takes the value 1 if only if $X^{(1)}$ is almost surely a strictly monotone function of $X^{(2)}$.

(*b*) *If $\mathbf{X}$ follows a bivariate normal distribution with correlation coefficient r, then $I_\sigma(\mathbf{X})$ is a strictly increasing function of $|r|$ with $I_\sigma(\mathbf{X}) \leq |r|$.*

Another interesting property of $I_\sigma(\mathbf{X})$ is its irreducibility. Following Schmid *et al.* (2010), we call a dependency measure I irreducible if, for all $p > 2$, $I(X^{(1)}, X^{(2)}, \ldots, X^{(p)})$ is not a function of the quantities $\{I(\mathbf{X}^{(i_1)}, \mathbf{X}^{(i_2)}, \ldots, \mathbf{X}^{(i_k)}) : \{i_1, i_2, \ldots, i_k\} \subsetneqq \{1, 2, \ldots, p\}\}$. Naturally, any reasonable multivariate measure of dependence is expected to be irreducible. Note that if $I(X^{(1)}, X^{(2)}, X^{(3)})$ gets completely determined by $I(X^{(1)}, X^{(2)})$, $I(X^{(2)}, X^{(3)})$ and $I(X^{(3)}, X^{(1)})$, instead of mutual dependence among $X^{(1)}$, $X^{(2)}$ and $X^{(3)}$, it can only detect pairwise dependence. The following theorem shows that any copula based multivariate dependency measure, which takes the value zero only for the uniform copula, is irreducible.

Theorem 2.3. *Let C be the copula distribution of $\mathbf{X}$ and $I(\mathbf{X}) = \mathcal{M}(\mathsf{C})$ be a copula based multivariate dependency measure. If $\mathcal{M}(\mathsf{C}) = 0$ implies $\mathsf{C} = \Pi$, then I is irreducible.*

For any fixed bandwidth parameter σ, the irreducibility of our proposed measure $I_\sigma(\mathbf{X})$ follows from Theorem 2.3 as a corollary. However, this property vanishes when σ diverges to infinity. In such a situation, the limiting value of $I_\sigma(\mathbf{X})$ turns out to be the average of squared Spearman's rank correlation coefficients between $\binom{p}{2}$ pairs of random variables as stated in the following theorem.

Proposition 2.4. *As σ diverges to infinity, $I^2_\sigma(\mathbf{X})$ converges to $\frac{1}{\binom{p}{2}} \sum_{1\le i<j\le p} \mathrm{Cor}^2(T^{(i)}, T^{(j)})$, where $\mathbf{T} = (T^{(1)}, T^{(2)}, \ldots, T^{(p)}) \sim \mathsf{C}$*

2.2 Estimation of the proposed measure

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be n independent observations on the random vector $\mathbf{X}$. For any fixed $j = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$, we define $r_i^{(j)}$ as the rank of $x_i^{(j)}$ (the j-th component of $\mathbf{x}_i$) in the set $\{x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}\}$ to get $\mathbf{r}_i = (r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(p)})$, the coordinate-wise rank of $\mathbf{x}_i$. For $i = 1, 2, \ldots, n$, we use the normalized rank vectors $\mathbf{y}_i = \mathbf{r}_i/n$ to define the empirical version of the copula distribution C, which is given by

$$\mathsf{C}_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} \mathbb{I}[y_i^{(j)} \le u^{(j)}],$$

where $\mathbb{I}$ is the indicator function. Clearly, C_n is the empirical distribution function based on $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$. Similarly, we define empirical versions of the maximum copula and the uniform copula as

$$\mathsf{M}_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} \mathbb{I}[u^{(j)} \ge i/n] \quad \text{and} \quad \Pi_n(\mathbf{u}) = \prod_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[u^{(j)} \ge i/n],$$

respectively. While M_n puts the equal mass $1/n$ on each of the n points $\{(i/n, i/n, \ldots, i/n) : 1 \le i \le n\}$, Π_n assigns equal mass to n^p points of the form $(i_1/n, i_2/n, \ldots, i_p/n)$ for $i_1, i_2, \cdots, i_p \in \{1, 2, \ldots, n\}$. We estimate $I_\sigma(\mathbf{X})$ by its empirical analog

$$\widehat{I}_{\sigma,n}(\mathbf{X}) = \gamma_{K_\sigma}(\mathsf{C}_n, \Pi_n)/\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n).$$

Note that $\widehat{I}_{\sigma,n}(\mathbf{X})$ is well-defined since $\mathsf{M}_n \ne \Pi_n$ for every $n > 1$. Unlike $\gamma_{K_\sigma}(\mathsf{M}, \Pi)$, $\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n)$ has a closed form expression, and this leads to a closed form expression for $\widehat{I}_{\sigma,n}(\mathbf{X})$ as well (see Equation (2.2)). Here one does not need to use the numerical integration method or the statistical simulation technique for its computation. One can also check that

from equation (2.1), it is easy to get the following expression for the proposed estimator

$$\widehat{I}_{\sigma,n}(\mathbf{X}) = \sqrt{\frac{s_1 - 2s_2 + v_3}{v_1 - 2v_2 + v_3}}, \tag{2.2}$$

where $s_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j>i}^{n} K_\sigma(\mathbf{y}_i, \mathbf{y}_j) + \frac{1}{n}$, $s_2 = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}\frac{1}{n}\sum_{l=1}^{n} e^{-\frac{1}{2}\left\{(l-ny_i^{(j)})/n\sigma\right\}^2}$,
$v_1 = \frac{2}{n^2}\sum_{i=1}^{n-1}(n-i)e^{-\frac{p}{2}(i/n\sigma)^2} + \frac{1}{n}$, $v_2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n} e^{-\frac{1}{2}\{(i-j)/n\sigma\}^2}\right]^p$ and
$v_3 = \left[\frac{2}{n^2}\sum_{i=1}^{n-1}(n-i)e^{-\frac{1}{2}(i/n\sigma)^2} + \frac{1}{n}\right]^p$.

The above formula shows that the computing cost of $\widehat{I}_{\sigma,n}(\mathbf{X})$ is of the order $\mathcal{O}(pn^2)$. This estimate enjoys some nice theoretical properties similar to those of $I_\sigma(\mathbf{X})$. Some of these properties are mentioned below.

Proposition 2.5. *Let $\widehat{I}_{\sigma,n}(\mathbf{X})$ be the empirical version of $I_\sigma(\mathbf{X})$ based on n independent observations from the joint distribution of $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$.*

(a) $\widehat{I}_{\sigma,n}(\mathbf{X})$ is invariant under permutation and strictly monotone transformations of the coordinate variables $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$.

(b) For all $i = 2, 3, \ldots, p$, if $X^{(i)}$ is almost surely a strictly monotone function of $X^{(1)}$, then $\widehat{I}_{\sigma,n}(\mathbf{X})$ takes the value 1.

As we have mentioned before, other existing copula based dependency measures do not have the property mentioned in part (b) of the above proposition. For instance, multivariate extensions of Spearman's ρ, Kendall's τ, Blomqvist's β and Hoeffding's Φ statistics (see, e.g., Nelsen, 1996, 2002; Úbeda-Flores, 2005; Gaißer *et al.*, 2010) may not take the value 1 even when the measurement variables have monotone relationships among them. To demonstrate this, we considered a simple example. We generated 10000 independent observations on $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(p)})$, where $X^{(i)} = V$ or $X^{(i)} = -V$ depending on $i = 1, 2, \ldots, p$ and $V \sim U(0, 1)$. Hence each pair of variables were monotonically related. We considered three choices of p ($p = 3, 4, 5$), and for each value of p, results are reported in Table 2.1 for different types of relationships shown in the orientation column. For example, the $(\uparrow, \uparrow, \downarrow)$ sign in the orientation column indicates that $(X^{(1)}, X^{(2)}, X^{(3)}) = (V, V, -V)$. Table 2.1 clearly shows that all dependency measures considered here take the value 1 when the relationships among the variables are strictly increasing. But, barring $\widehat{I}_{\sigma,n}(\mathbf{X})$, all other measures fail to have this property for other monotone relationships among the variables.

TABLE 2.1: Different measures of dependence for monotonically related variables.

Dimension	Orientation	$\widehat{I}_{\sigma,n}(\mathbf{X})$	Spearman	Kendall	Blomqvist	Hoeffding
3	$(\uparrow,\uparrow,\uparrow)$	1.000	1.000	1.000	1.000	1.000
3	$(\uparrow,\uparrow,\downarrow)$	1.000	-0.333	-0.333	-0.333	0.517
4	$(\uparrow,\uparrow,\uparrow,\uparrow)$	1.000	1.000	1.000	1.000	1.000
4	$(\uparrow,\uparrow,\uparrow,\downarrow)$	1.000	-0.091	-0.143	-0.143	0.382
4	$(\uparrow,\uparrow,\downarrow,\downarrow)$	1.000	-0.212	-0.143	-0.143	0.327
5	$(\uparrow,\uparrow,\uparrow,\uparrow,\uparrow)$	1.000	1.000	1.000	1.000	1.000
5	$(\uparrow,\uparrow,\uparrow,\uparrow,\downarrow)$	1.000	0.016	-0.067	-0.067	0.347
5	$(\uparrow,\uparrow,\uparrow,\downarrow,\downarrow)$	1.000	-0.108	-0.067	-0.067	0.236

Since $\widehat{I}_{\sigma,n}(\mathbf{X})$ is based on coordinate-wise ranks of the observations, it is robust against contamination and outliers generated from heavy-tailed distributions. Following the results in Póczos *et al.* (2012), one can show that addition of a new observation can change its value by at most $\mathcal{O}(n^{-1})$. Just like $I_\sigma(\mathbf{X})$, its empirical analog $\widehat{I}_{\sigma,n}(\mathbf{X})$ also enjoys some additional properties for $p = 2$. The following theorem shows that a result analogous to Theorem 2.2 holds for $\widehat{I}_{\sigma,n}(\mathbf{X})$ as well.

Theorem 2.4. *Suppose that* $\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_n$ *are normalized coordinate-wise ranks (as defined in the beginning of Subsection 2.2) corresponding to bivariate observations* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. *Define*

$$v_{i,j}^{(l)} = K_\sigma(y_i^{(l)}, y_j^{(l)}) - \frac{1}{n}\sum_{i=1}^{n} K_\sigma(y_i^{(l)}, y_j^{(l)}) - \frac{1}{n}\sum_{j=1}^{n} K_\sigma(y_i^{(l)}, y_j^{(l)}) + \frac{1}{n^2}\sum_{i,j=1}^{n} K_\sigma(y_i^{(l)}, y_j^{(l)}).$$

for $i, j = 1, 2, \ldots, n$ *and* $l = 1, 2$. *Then* $\widehat{I}_{\sigma,n}(\mathbf{X})$ *can be expressed as*

$$\widehat{I}_{\sigma,n}(\mathbf{X}) = \frac{\sum_{i,j=1}^{n} v_{i,j}^{(1)} v_{i,j}^{(2)}}{\sqrt{\sum_{i,j=1}^{n} \left(v_{i,j}^{(1)}\right)^2 \sum_{i,j=1}^{n} \left(v_{i,j}^{(2)}\right)^2}}.$$

As a consequence, we have $0 \leq \widehat{I}_{\sigma,n}(\mathbf{X}) \leq 1$, *where* $\widehat{I}_{\sigma,n}(\mathbf{X}) = 1$ *holds if and only if one coordinate variable is a strictly monotone function of the other.*

2.3 Test of independence based on $\widehat{I}_{\sigma,n}(\mathbf{X})$

We have seen that $I_\sigma(\mathbf{X})$ serves as a measure of dependence among the coordinates of $\mathbf{X}$. It is non-negative, and takes the value 0 if and only if $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$ are independent. So, we can use $\widehat{I}_{\sigma,n}(\mathbf{X})$ as the test statistic and reject $\mathbb{H}_0$, the null hypothesis of mutual independence, for large values of $\widehat{I}_{\sigma,n}(\mathbf{X})$. The large sample distribution of our test statistic is given by the following theorem.

Theorem 2.5. *Assume that the copula distribution* C *has continuous partial derivatives.*

(a) If $\mathsf{C} = \Pi$, *then* $n\widehat{I}^2_{\sigma,n}(\mathbf{X}) \xrightarrow{\mathcal{D}} \sum_{i=1}^{\infty} \lambda_i Z_i^2$, *where the* Z_i*'s are i.i.d.* $N(0,1)$ *and the* λ_i*'s are some positive constants (see the proof in Section 2.6 for details on the* λ_i*'s).*

(b) If $\mathsf{C} \neq \Pi$, *then* $\sqrt{n}(\widehat{I}_{\sigma,n}(\mathbf{X}) - I_\sigma(\mathbf{X})) \xrightarrow{\mathcal{D}} N(0, \delta^2)$*; where*

$$\delta^2 = \gamma^{-2}_{K_\sigma}(\mathsf{C}, \Pi) \int_{[0,1]^p} \int_{[0,1]^p} g(\mathbf{u}) g(\mathbf{v})\, \mathrm{E}[\, d\mathbb{G}_{\mathsf{C}}(\mathbf{u})\, d\mathbb{G}_{\mathsf{C}}(\mathbf{v})], \;\; g(\mathbf{u}) = \int_{[0,1]^p} K_\sigma(\mathbf{u}, \mathbf{v})\, d(\mathsf{C} - \Pi)(\mathbf{v})$$

and $\mathbb{G}_{\mathsf{C}}$ *is a 0 mean Gaussian process as defined in Theorem 2.9 in Section 2.6.*

The histograms in Figures 2.1(a) and 2.1(b) show the empirical distributions of $\widehat{I}_{\sigma,n}(\mathbf{X})$ computed based on 5000 independent samples, each of size 200, generated from bivariate normal distributions with correlation coefficient $\rho_0 = 0$ and 0.5, respectively. For $\rho_0 = 0.5$ (i.e., $\mathsf{C} \neq \Pi$), while the empirical distribution looks like a normal distribution, for $\rho_0 = 0$ (i.e., $\mathsf{C} = \Pi$), it turns out to be positively skewed. This is consistent with the result stated in Theorem 2.5.

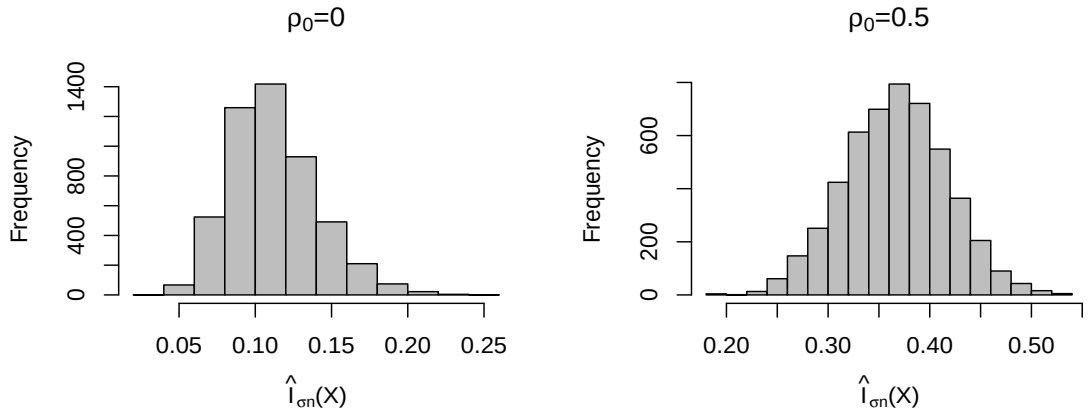


FIGURE 2.1: Empirical distribution of $\widehat{I}_{\sigma,n}(\mathbf{X})$ with $\sigma = 0.2$ for standard bivariate normal distribution with correlation coefficient 0 and 0.5 respectively.

One can notice that, the probability convergence of $\widehat{I}_{\sigma,n}(\mathbf{X})$ follows from Theorem 2.5. But, we also have a stronger result in this context, which is stated below.

Theorem 2.6. $\widehat{I}_{\sigma,n}(\mathbf{X})$ *converges to* $I_\sigma(\mathbf{X})$ *almost surely as* n *tends to infinity.*

From Theorem 2.6, it is clear that under the null hypothesis of independence, $\widehat{I}_{\sigma,n}(\mathbf{X})$ converges to 0 almost surely, while under the alternative, it converges to a positive constant. For any fixed choice of σ, the large sample consistency of the test follows from it. However, for practical implementation of the test, one needs to determine the cut-off. It is difficult to find this cut-off based on the asymptotic null distribution of the test statistic mentioned

in Theorem 2.5. Gretton *et al.* (2008) proposed to approximate an infinite weighted sum of independent chi-square random variables by a two parameter gamma distribution. In such cases, one determines the exact mean and the exact variance of the null distribution of $n\widehat{I}^2_{\sigma,n}(\mathbf{X})$ and then approximates the null distribution by a gamma distribution with the same mean and the same variance. Exact mean and variance of $n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)$ under $\mathbb{H}_0$ is given in Appendix A, from which the exact mean and the exact variance of the null distribution of $n\widehat{I}^2_{\sigma,n}(\mathbf{X})$ can be obtained by using a proper scaling. Instead of exact mean and variance, sometimes asymptotic mean and variance of $n\widehat{I}^2_{\sigma,n}(\mathbf{X})$ are also considered. The asymptotic mean and the asymptotic variance of $n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)$ are also given in Appendix A. However, instead of using gamma approximation, here we use the distribution-free property of $\widehat{I}_{\sigma,n}(\mathbf{X})$ to determine the cut-off. Note that under $\mathbb{H}_0$, for each $j = 1, 2, \ldots, p$, we have $\Pr[r_1^{(j)} = i_1, r_2^{(j)} = i_2, \ldots, r_n^{(j)} = i_n] = 1/n!$ for any permutation $(i_1, i_2, \ldots, i_n)$ of $\{1, 2, \ldots, n\}$, and for different values of j, they are independent. So, we can easily generate normalized coordinate-wise ranks under $\mathbb{H}_0$ and use them to compute the test statistic. We repeat this procedure 10,000 times to approximate the $(1-\alpha)$-th quantile of the null distribution of $\widehat{I}_{\sigma,n}(\mathbf{X})$, which is then used as the cut-off. Note that this whole calculation can be done off-line, and we can prepare a table of critical values for different choices n and σ before handling the actual observations.

Though any fixed choice of σ leads to a consistent test (follows from Theorem 2.6), its finite sample power may depend on this choice. The method commonly used for choosing the bandwidth is based on "median heuristic" (see, e.g., Fukumizu *et al.*, 2009a, Sec 5), where one computes all pairwise distances among the observations and then the median of those distances is used to select the bandwidth. Since we are using the kernel on the normalized rank vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ having the null distribution Π_n, following the idea of median heuristic, we can choose $2\sigma^2$ to be the median of $\|\mathbf{Z} - \mathbf{Z}_*\|^2$, where $\mathbf{Z}, \mathbf{Z}_* \overset{i.i.d.}{\sim} \Pi_n$. Note that the bandwidth chosen in this way is non-random function of n. We denote it by σ_n. As n increases, since Π_n converges to Π, $2\sigma_n^2$ converges to the median of $\|\mathbf{Z} - \mathbf{Z}_*\|^2$, where $\mathbf{Z}, \mathbf{Z}_* \overset{i.i.d}{\sim} \Pi$. The following theorem shows that our test remains consistent for a such choice of the bandwidth.

Theorem 2.7. *Consider a sequence of bandwidths* $\{\sigma_n : n \geq 1\}$ *converging to some* $\sigma_0 > 0$. *Then, under alternative hypothesis, power of the proposed test based on* $\widehat{I}_{\sigma_n,n}(X)$ *converges to* 1 *as the sample size* n *diverges to infinity.*

To evaluate the performance of our test based on this choice of bandwidth, we analyzed some simulated data sets. For each example, we repeated our experiment 10000 times, and the power of the test was estimated by the proportion of times it rejected $\mathbb{H}_0$. Powers were also computed for the generalized versions of HSIC and dCov tests, known as the dHSIC test (Pfister *et al.*, 2018) and the JdCov test (Chakraborty and Zhang, 2019), respectively. Since our proposed test is based on ranks, a rank version of the JdCov test (referred to as the rank-JdCov test) was also used for comparison. Results are also reported for the tests based on generalized versions of Hoeffding's ϕ statistic (Gaißer *et al.*, 2010) and Spearman's ρ statistic (Nelsen, 1996) (henceforth referred to as the Hoeffding test and the Spearman test, respectively). For $p = 2$, we also used the HHG test proposed by Heller *et al.* (2013). Brief descriptions of these tests are given in Appendix B. For the implementation of the dHSIC test, we used the R package "dHSIC" (Pfister and Peters, 2019), where we used the Gaussian kernel with the default bandwidth chosen using median heuristic. For JdCov and rank-JdCov tests, we used the R codes provided by the authors. We considered the scale invariant version of JdCov test and the U-statistics version of rank-JdCov test. Following the suggestion of the authors, the value of the tuning parameter C was taken as 1. The HHG test was implemented using the R package "HHG" (Brill and Kaufman, 2019). For the Hoeffding test, the Spearman test and our proposed test, we used our own codes. For our proposed tests, we created an R package 'CGK' containing all necessary codes. This package is available at `https://github.com/angshumanroycode/CGK`. Throughout this thesis, unless mentioned otherwise, all tests are considered to have 5% nominal level. In all cases, cut-offs were computed using the permutation principle. For the permutation method, keeping the first coordinate (variable) fixed, we randomly permuted the values of the other coordinates (variables) to get a new set of observations $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_n^*$, where $\mathbf{x}_i^* = \left(x_i^{(1)}, x_{\pi_2(i)}^{(2)}, \ldots, x_{\pi_p(i)}^{(p)}\right)$ for $i = 1, 2, \ldots, n$. Here $\pi_2, \ldots, \pi_p$ are $p-1$ independent random permutations of $\{1, 2, \ldots, n\}$. We computed the test statistic based on this new set of observations and repeated this procedure several times (here we used 1000 repetitions) to get an empirical distribution of the test statistic under $\mathbb{H}_0$. The upper α-th quantile (we used $\alpha = 0.05$) of this distribution was used as the cut-off. We computed powers of different tests for different sample sizes and they are reported in Figure 2.2. In all examples, these sample sizes were chosen in such a way that most of the tests had powers appreciably different from the nominal level of 0.05 and unity.

We began with two examples involving bivariate data. In the 'Correlated Normal' example, we generated observations from the bivariate normal distribution with correlation coefficient 0.4 to get the two coordinate variables $X^{(1)}$ and $X^{(2)}$, respectively. In the 'Hyperplane' example, we took $X^{(1)} = U$ and $X^{(2)} = U + V$, where $U, V \overset{i.i.d.}{\sim} \mathrm{U}(-1, 1)$. In these two examples, dHSIC and HHG tests had somewhat inferior performance than their competitors (see Figures 2.2(a) and 2.2(b)). All other tests, including our proposed test based on $T_n = \widehat{I}_{\sigma_n,n}(\mathbf{X})$ had almost similar powers.

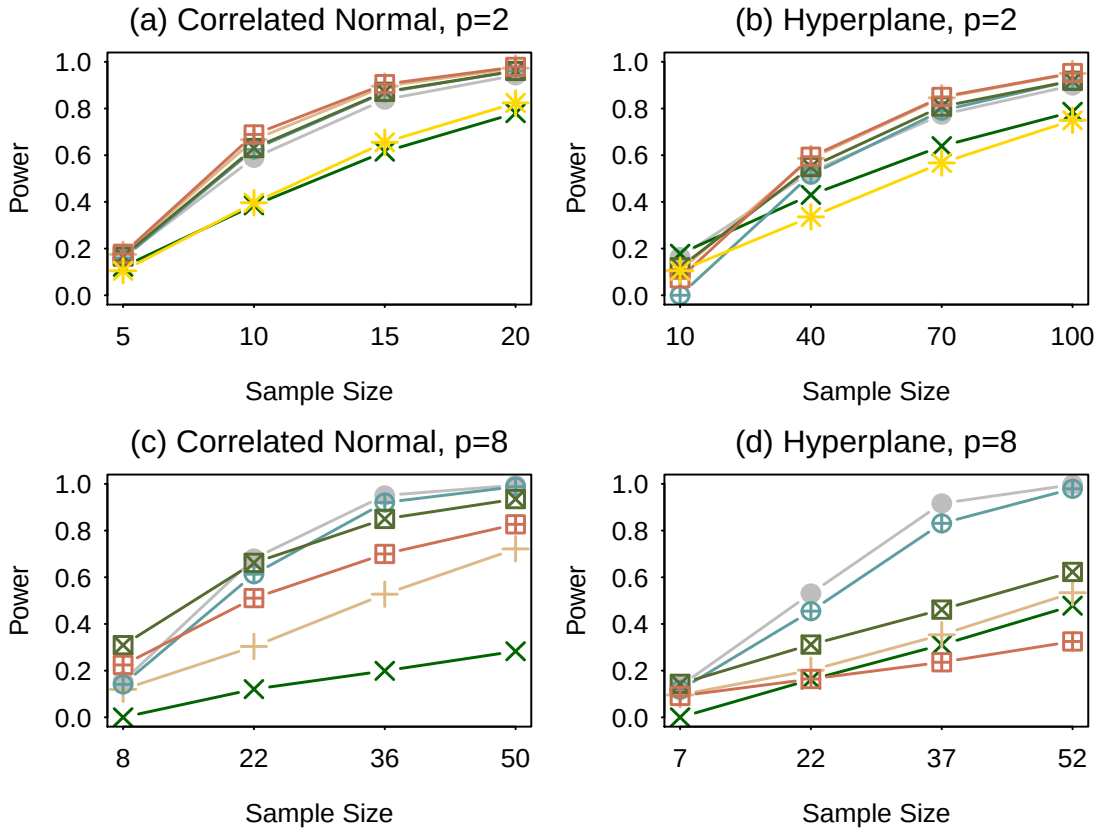


FIGURE 2.2: Powers of dHSIC (×), JdCov (+), rank-JdCov (⊕), Hoeffding (⊠), Spearman (⊞), HHG (✳) tests and the proposed test based on T_n (●), in 'Correlated Normal' and 'Hyperplane' examples.

Next we carried out our experiments with eight-dimensional versions of these data sets. In the 'Correlated Normal' example, observations on $\mathbf{X}$ were generated from a 8-dimensional normal distribution with the mean vector $\mathbf{0}$ and the dispersion matrix $\mathbf{\Sigma} = ((a_{i,j}))$, where $a_{i,j} = 0.4^{|i-j|}\ \forall\ i, j = 1, 2, \ldots, 8$. In the 'Hyperplane' example, $X^{(2)}, X^{(3)}, \ldots, X^{(8)}$ and ε were independently generated from the $U(-1, 1)$ distribution, and $X^{(1)}$ was defined as $X^{(1)} = X^{(2)} + X^{(3)} + \cdots + X^{(8)} + \varepsilon$. In these examples involving more than two random vectors, the HHG test could not be used. Figures 2.2(c) and 2.2(d) show that in these two

examples, our proposed test had the best performance, which was closely followed by the rank-JdCov test. Unlike the two-dimensional examples, here the Spearman test could not perform well. The dHSIC test also performed poorly in both examples. In the 'Hyperplane' example, the Hoeffding test and the JdCov test had somewhat inferior performance as well. Along with these tests, we also considered the tests proposed by Póczos *et al.* (2012), where the cut-offs were computed based on their suggested probability inequalities. But those tests had much lower powers compared to all other tests considered here, and we decided not to report those results in this thesis.

2.4 Multi-scale approach and aggregation of results

Though in the examples considered in Section 2.3, the bandwidth chosen using median heuristic yielded good results, this may not always be the case. Our empirical experience suggests that median heuristic performs well when the relationships among the variables are nearly monotone (i.e., the conditional expectation of one variable given others is a non-constant monotone functions of those variables). But in cases of complex non-monotone relationships, the use of smaller bandwidths often yields better results. In such cases, instead of median, one can use lower quantiles of pairwise distances.

To demonstrate this, we considered two simple examples involving bivariate data sets. In one example, observations were generated from the 'Two parabolas'-type distribution mentioned in Newton (2009) (see Figure 2.5(e)) and in the other example, they were generated from a bivariate normal distribution with correlation coefficient 0.5. In each case, we generated 25 observations and repeated the experiment 10000 times to estimate the powers of the tests based on $\widehat{I}_{\sigma,n}(\mathbf{X})$ for different choices of σ based on different quantiles $(0.01, 0.02, 0.05, 0.1, 0.2$ and $0.5)$ of pairwise distances. Though the bandwidth based on median of pairwise distances worked well in the second example, those based on smaller quantiles had better results in the first (see Figure 2.3). Figure 2.3 clearly shows that depending on the underlying distribution of $\mathbf{X}$, sometimes we need to use larger bandwidth, whereas sometimes smaller bandwidths may perform better. While larger bandwidths successfully detect global linear or monotone relationships among the variables, smaller bandwidths are useful for detecting non-monotone or local patterns. In order to capture both types of dependence, borrowing the idea from computer vision (see, e.g. Lindeberg, 2013) and machine learning literature (see, e.g. Ghosh *et al.*, 2006; Dutta *et al.*, 2016),

here we adopt a multi-scale approach, where we look at the results for several choices of bandwidth and then aggregate them judiciously to come up with the final decision.

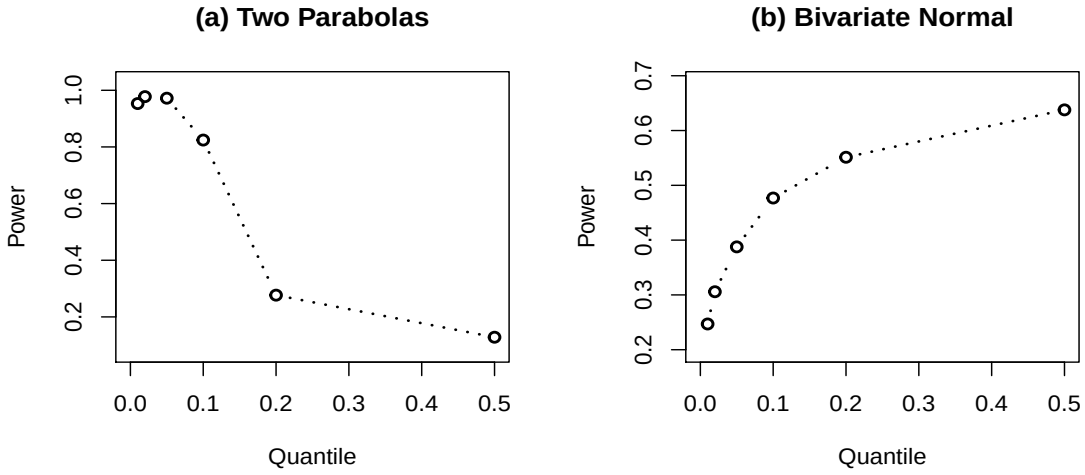


FIGURE 2.3: Powers of the test based on $\widehat{I}_{\sigma,n}(\mathbf{X})$ for bandwidths corresponding to different quantiles of pairwise distances.

Figures 2.4(a) and 2.4(b) show the observed p-values for different choices of the bandwidth (based on quantiles of pairwise distances) when a sample of size 25 was generated from bivariate normal distributions with correlation coefficient 0 and 0.5, respectively. Clearly, these plots of p-values carry more information than just the final result. In the first case, higher p-values for all choices of the bandwidth give a visual evidence in favor $\mathbb{H}_0$, while smaller p-values for a long range of bandwidths in the second case indicates dependence between the two coordinate variables. Also the pattern of p-values can reveal the structure of dependence among the variables. For instance, smaller p-values for larger bandwidths indicate that the relationship between the two variables is nearly monotone (like the case here), while those for smaller bandwidths indicate complex, non-monotone relations.

One way of aggregating the results corresponding to m bandwidths $\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(m)}$ is to use $T_{\text{sum},n} = \sum_{i=1}^{m} \widehat{I}_{\sigma^{(i)},n}(\mathbf{X})$ or $T_{\max,n} = \max_{1 \le i \le m} \widehat{I}_{\sigma^{(i)},n}(\mathbf{X})$ as the test statistic. Following Sarkar and Ghosh (2018), one can also use another method based on false discovery rate (FDR). Let p_i be the p-value of the test based on $\sigma^{(i)}$ (for $i = 1, 2, \ldots, m$) and $p_{(1)} \le p_{(2)} \le \ldots \le p_{(m)}$ be the corresponding order statistics. We reject $\mathbb{H}_0$ at level α if and only if the set $\{i : p_{(i)} < i\ \alpha/m\}$ is non-empty. Benjamini and Hochberg (1995) proposed this method for controlling FDR for a set m independent tests. Later, Benjamini and Yekutieli (2001) showed that it also controls FDR when the tests statistics are positively regression dependent. Since we are testing the same hypothesis for different choices of the

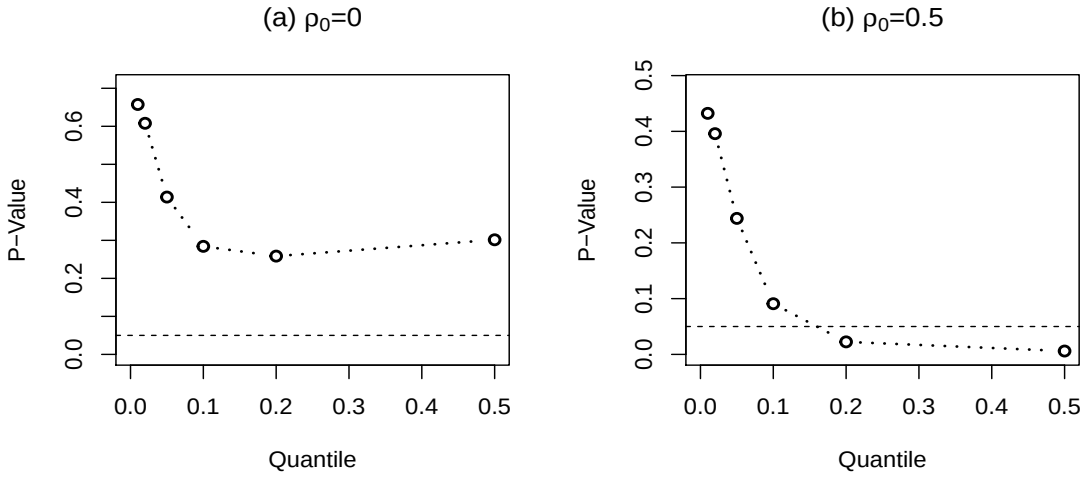


FIGURE 2.4: p-values of the test based on $\widehat{I}_{\sigma,n}(\mathbf{X})$ for bandwidths corresponding to different quantiles of pairwise distances.

bandwidth, this method controls the level of the test as well (see, e.g., Cuesta-Albertos and Febrero-Bande, 2010). It is difficult to prove positive regression dependence among the test statistics corresponding to different choices of bandwidth. However, all pairwise correlations (computed over 10000 simulations) among these test statistics were found to be positive in all of our numerical experiments. This gives an indication of positive regression dependence among the test statistics and thereby provides an empirical justification for using the above method. The following theorem shows the large sample consistency of the multi-scale versions of our tests based on $T_{\text{sum},n}$, $T_{\max,n}$ and FDR.

Theorem 2.8. *Under the alternative hypothesis, powers of the proposed tests based on* $T_{\text{sum},n}$, $T_{\max,n}$ *and FDR converge to* 1 *as the sample size tends to infinity.*

2.5 Results from the analysis of simulated and real data sets

We analyzed several simulated and real data sets to compare the performance of our proposed tests based on $T_{\text{sum},n}, T_{\max,n}$ and FDR with some popular tests available in the literature. In particular, we considered the tests used in Section 2.3 for comparison. For the multi-scale versions of the proposed test, we started with the bandwidth based on median heuristic (σ_n, say) and considered other bandwidths of the form $(0.5)^i \times \sigma_n$ for $i = 1, 2, \ldots$. However, we did not consider any bandwidth smaller than one-third of the fifth percentile of pairwise distances. The choice of this fraction 'one-third' was motivated by the use of the Gaussian kernel, and following the idea of Ghosh *et al.* (2006), we chose

the bandwidths at equal intervals in the logarithmic scale. Note that like the test based on T_n, these multi-scale tests also have the distribution-free property. So, the cut-offs of all these tests were computed based on 1000 random permutations as before.

2.5.1 Analysis of simulated data sets

We began with six examples involving six unusual bivariate distributions considered by Newton (2009). Scatter plots of these data sets are displayed in Figure 2.5. For each of these examples, we considered samples of different sizes, and for each sample size, the experiment was repeated 1000 times to compute the powers of different tests. These powers are reported in Figure 2.6.

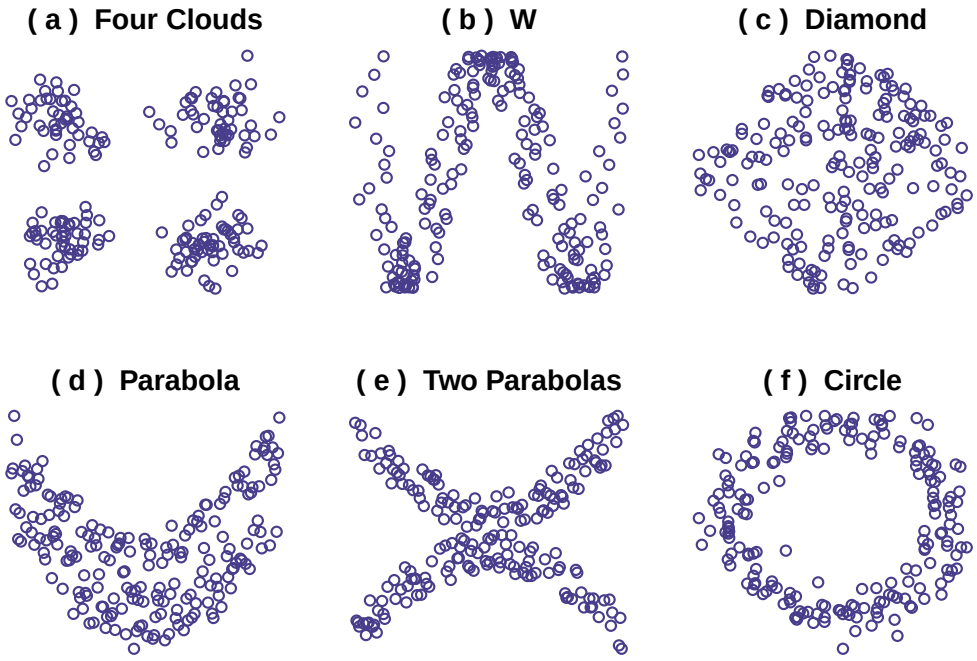


FIGURE 2.5: Observations from Newton (2009)'s six unusual bivariate distributions.

Note that in these six examples, $X^{(1)}$ and $X^{(2)}$ are uncorrelated. In 'Four Clouds' data, they are independent as well. In this data set, almost all tests had powers close to the nominal level of 0.05 (see Figure 2.6(a)). Only the test based on FDR had slightly low powers, which is quite expected in view of the conservative nature of such tests.

In the next five examples, $X^{(1)}$ and $X^{(2)}$ are not independent. In the 'W' example, our proposed test based on $T_{\max,n}$ had the best overall performance followed by the test based on FDR, the HHG test and the dHSIC test (see Figure 2.6(b)). Powers of all other tests were much lower. JdCov, rank-JdCov, Spearman and Hoeffding tests failed to reject $\mathbb{H}_0$

on almost all occasions. These four tests had zero power in the 'Circle' example as well (see Figure 2.6(f)). In that example, the dHSIC test did not have satisfactory performance either. But our proposed test based on $T_{\max,n}$ and FDR had excellent performance. In cases of 'Parabola' and 'Two Parabolas' examples, though the HHG test had the highest power, our proposed tests also had competitive performance (see Figures 2.6(d) and 2.6(e)). Once again, Spearman, Hoeffding, JdCov and rank-JdCov tests had much lower powers than all other tests considered here. In the case of 'Diamond' example, the HHG test and the dHSIC test outperformed all other competing methods (see Figure 2.6(c)). However, in this example also, our proposed tests performed well. They had much higher powers than the rest of the competitors.

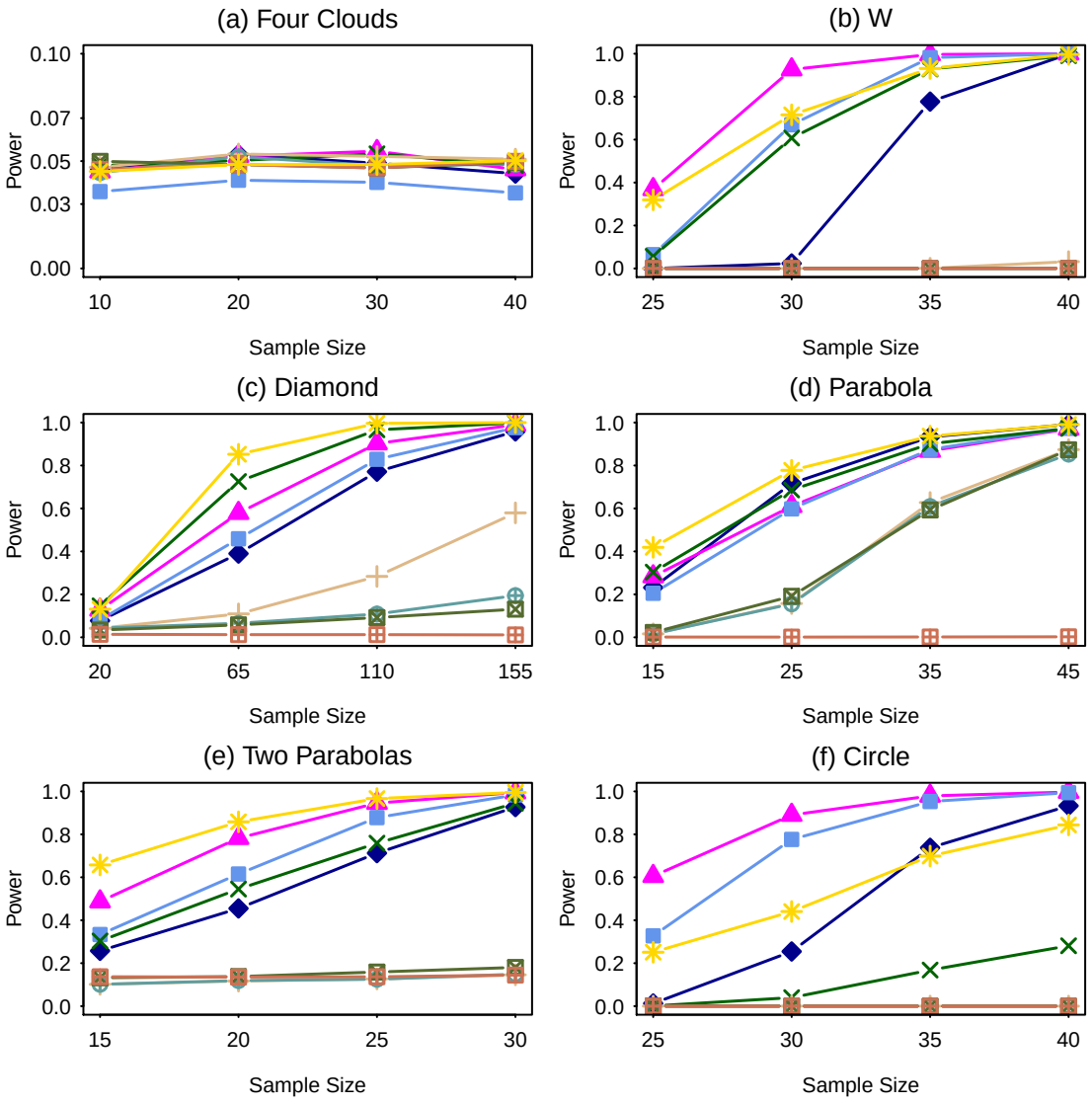


FIGURE 2.6: Powers of $T_{\text{sum},n}$ (◆), $T_{\max,n}$ (▲), FDR (■), dHSIC (×), JdCov (+), rank-JdCov (⊕), Hoeffding (⊠), Spearman (⊞) and HHG (✳) tests in Newton (2009)'s bivariate data sets.

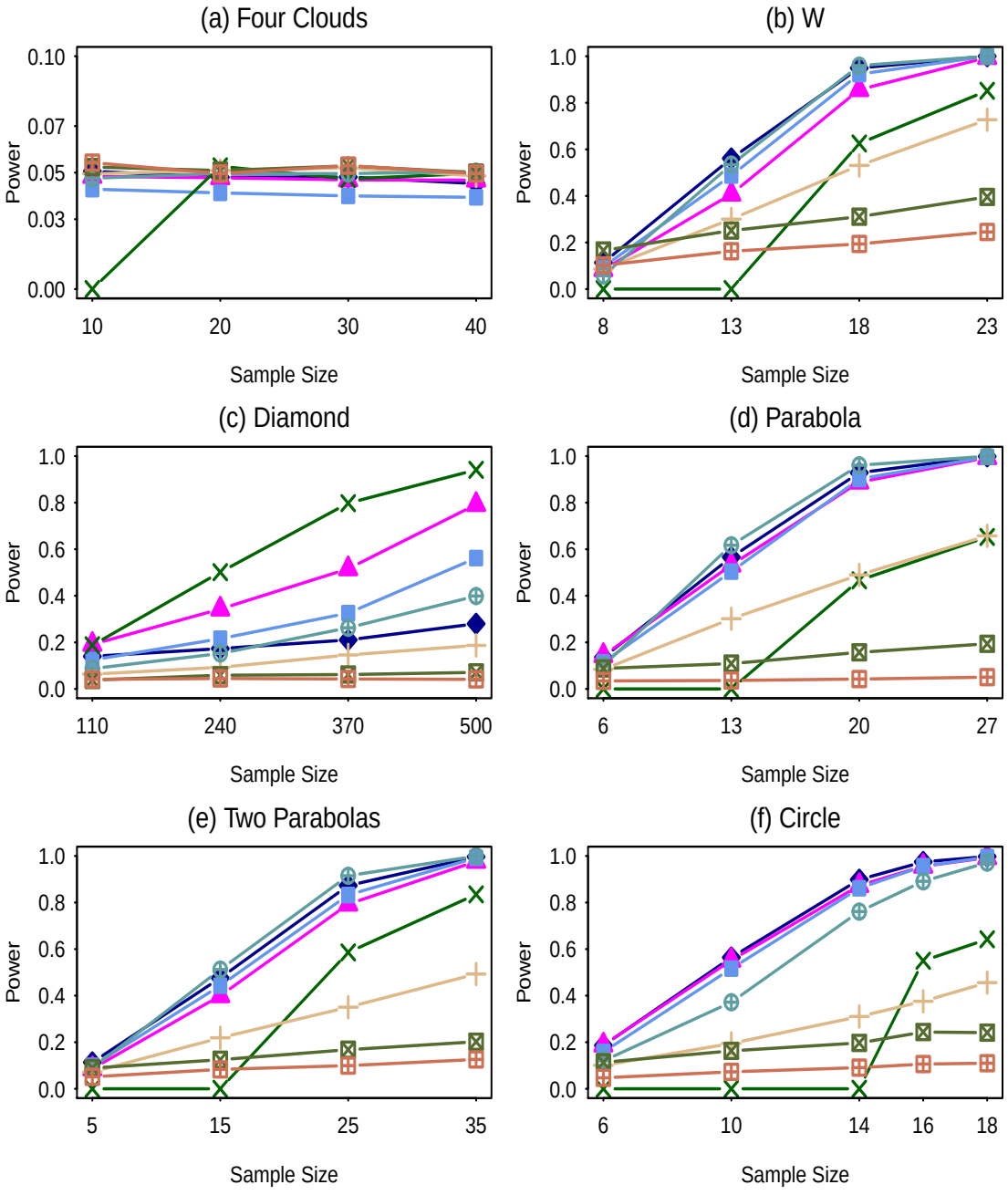


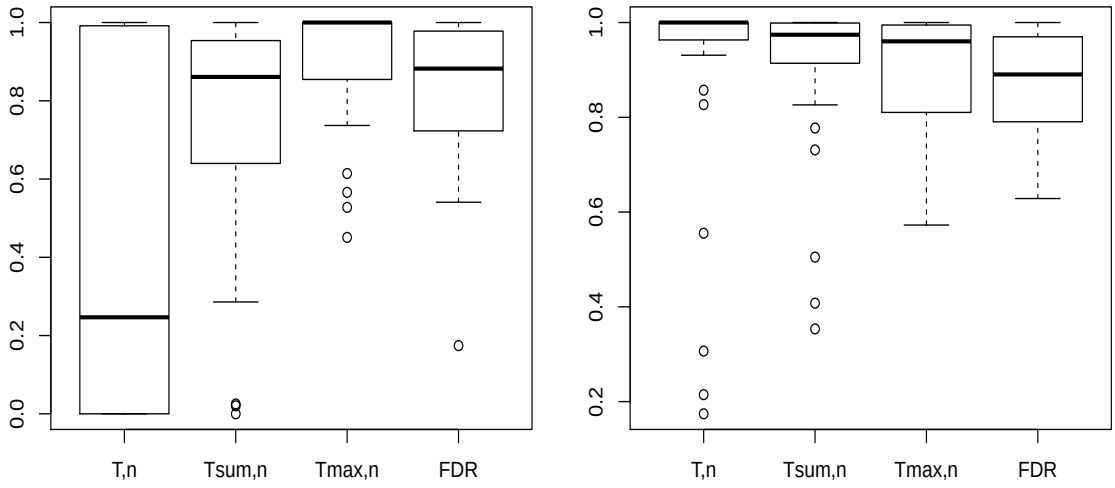
FIGURE 2.7: Powers of $T_{\text{sum},n}$ (◆), $T_{\text{max},n}$ (▲), FDR (■), dHSIC (×), JdCov (+), rank-JdCov (⊕), Hoeffding (⊠), and Spearman (⊞) tests in eight-dimensional simulated data sets.

Next we carried out our experiments with some eight dimensional data sets, which can be viewed as noisy multivariate extensions of the six bivariate data sets considered above. For each of the six examples, we generated two independent observations from the bivariate distribution, and then four independent $N(0, 1)$ variables were added to them to get a vector of dimension eight. In the case of 'Four Clouds' data set, again the test best on FDR had powers slightly lower than 0.05, but those of all other tests were close to the nominal level (see Figure 2.7(a)). Figure 2.7 clearly shows that in cases of 'W', 'Parabola'

and 'Two Parabolas' examples, our proposed tests and the rank-JdCov test had the best overall performance among the tests considered here. In the 'Circle' example also, our proposed tests outperformed all of their competitors. Only in the case of 'Diamond' data, the dHSIC test performed better than them. Note that the dHSIC test needs the sample size to be at least twice the dimension of the data (i.e., twice the number of coordinate variables) for its implementation. So, it could not be used in some experiments. In such cases, we considered its power to be zero.

In 'Correlated Normal' and 'Hyperplane' examples considered in Section 2.3, all multi-scale methods had powers similar to that of test based on T_n. That is why they are not reported separately. To have an overall comparison among the performance of the test based on T_n and its multi-scale versions in these simulated data sets, we followed the idea of Sarkar and Ghosh (2018). For a given data set and a given sample size, we defined the efficiency score of a test as the observed power of the test divided by the power of the best (among these four tests) test. So, the efficiency score of a test lies between 0 and 1, where value closer to 1 indicates that the test is more efficient. These efficiency scores were computed for the seven data sets (barring the 'Four Clouds' example, where the random variables were independent) and sample sizes considered in Sections 2.3 and 2.4, and they are presented using boxplots in Figure 2.8. This figure clearly shows us the necessity of the multi-scale approach. Overall performance of all multi-scale methods was better than the test based on T_n, especially in the case of $p = 2$. Among the multi-scale methods, the tests based on $T_{\max,n}$ had the best overall performance, followed by that based on FDR. Except for a few cases, the test based on $T_{\text{sum},n}$ also had competitive performance, particularly in cases of eight-dimensional data sets.

Next, we considered two interesting examples, where none of the lower dimensional marginals have dependence among the coordinate variables. In Example E1, we generated four independent $U(-1, 1)$ variables U_1, U_2, U_3, U_4, and defined $X^{(i)} = U_i$ for $i = 1, 2, \ldots, 4$ when $\prod_{i=1}^{4} U_i \geq 0$. In Example E2, we generated U_1, U_2, U_3, U_4 independently from $N(0, 1)$ to define $X_i = U_i \text{ sign}(U_{i+1})$ for $i = 1, 2, 3$ and $X_4 = U_4 \text{ sign}(U_1)$. Note that tests based on any dependency measure, which is not irreducible, will fail to detect the dependence among the coordinate variables in these examples. In both of these examples, the JdCov test had the highest power, but the rank-JdCov test performed miserably (see Figure 2.9). Spearman and Hoeffding tests also had poor performance. The dHSIC performed well only



(A) Bivariate data sets. (B) Eight-dimensional data sets.

FIGURE 2.8: Comparison between single-scale (based on T_n) and multi-scale (based on $T_{\text{sum},n}, T_{\max,n}$ and FDR) tests using boxplots of efficiency scores.

in Example E1. But our proposed methods based on $T_{\max,n}$ and FDR, particularly the former one had good performance in these two examples.

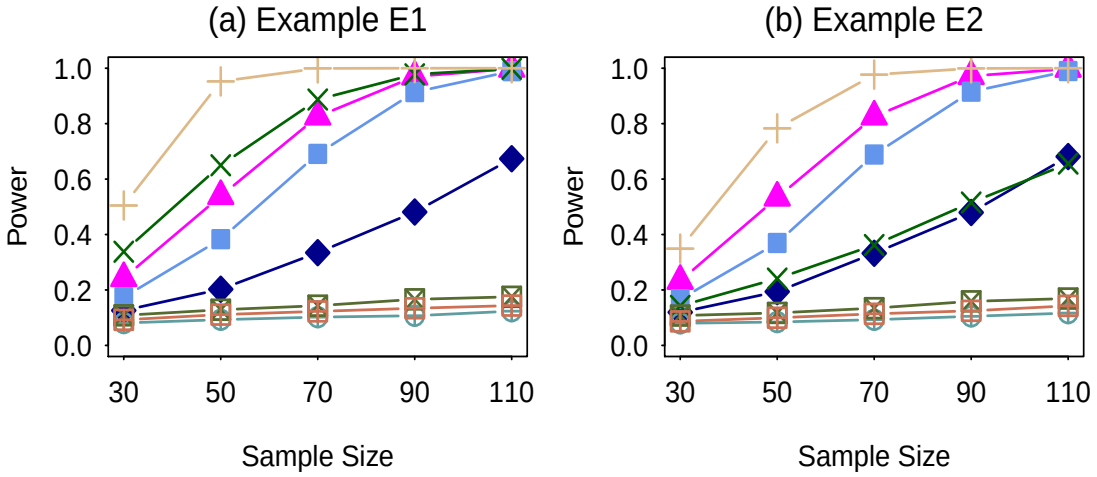


FIGURE 2.9: Powers of $T_{\text{sum},n}$ (◆), $T_{\max,n}$ (▲), FDR (■), dHSIC (×), JdCov (+), rank-JdCov (⊕), Hoeffding (⊠), and Spearman (⊞) tests in Examples E1 and E2.

2.5.2 Analysis of real data sets

We also analyzed two real data sets for further evaluation of our proposed methods. These data sets, viz. the Combined Cycle Power Plant (CCPP) data and the Airfoil Self-noise data, are available at the UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/. Brief description of these data sets is given below.

CCPP data contains 9568 observations from a combined cycle power plant over a period of six years (2006-2011), when the plant was set to work with full load. Each observation consists of hourly average values of ambient temperature, ambient pressure, relative humidity, exhaust vacuum and electric energy output. The idea was to predict electric energy output, which is dependent on other variables. When we used different methods to test for the independence among these five variables, all of them rejected the null hypothesis even for very small sample size. So, we removed the variable 'electric energy output' from our analysis and carried out our experiment with the remaining four variables.

Airfoil self-noise data set comprises different size NACA 0012 airfoils at various wind tunnel speeds and angles of attack. The span of the airfoil and the observer position were the same in all of the experiments. Brooks *et al.* (1989) analyzed this data set to develop a model for the scaled sound pressure level based on five input variables, viz. frequency, angle of attack, chord length, free-stream velocity and suction side displacement thickness. Here we want to know whether our tests can find dependence among these variables.

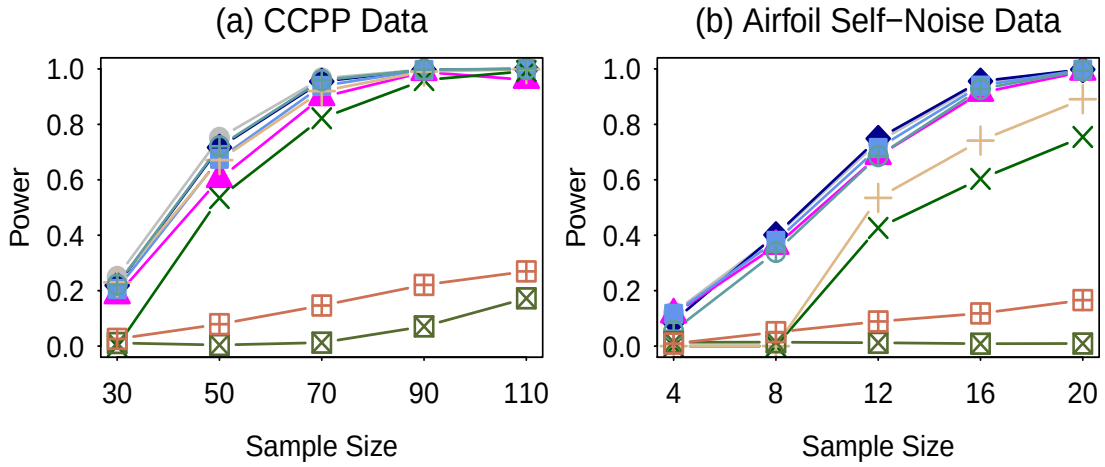


FIGURE 2.10: Powers of T_n (●), $T_{\text{sum},n}$ (◆), $T_{\text{max},n}$ (▲), FDR (■), dHSIC (×), JdCov (+), rank-JdCov (⊕), Hoeffding (⊠), and Spearman (⊞) tests in real data sets.

In both of these examples, when we used the full data set for testing, all tests rejected the null hypothesis. Based on that single experiment, it was not possible to compare among different test procedure. So, following the idea of Sarkar and Ghosh (2018), for each data set, we carried out our experiment with subsets of different sizes, and different tests were compared based on their powers. These subsets were chosen randomly, and for each subset size, the experiment was repeated 10000 times to compute the powers of different tests. These powers are shown in Figure 2.10, which clearly shows that in both of these examples,

our test based on T_n and its all three multi-scale analogs had excellent performance. The rank-JdCov test also performed well in these examples. The JdCov test had comparable power in CCPP data set, but not in Airfol Self-noise data set. Spearman and Hoeffding tests did not have satisfactory powers in either of these data sets. The dHSIC test performed better than them, but in both examples, it had lower power than our proposed tests.

2.6 Proofs and mathematical details

Proof of Proposition 2.1. For any permutation ξ on $\mathbb{R}^p$, we have $K_\sigma(\xi(\mathbf{x}), \xi(\mathbf{y})) = K_\sigma(\mathbf{x}, \mathbf{y})$ and also, $\mathbf{T} \sim \Pi$ implies $\xi(\mathbf{T}) \sim \Pi$. Using these, one gets

$$\mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim \mathsf{C}_{\xi(\mathbf{X})}\otimes \mathsf{C}_{\xi(\mathbf{X})}}[K_\sigma(\mathbf{S},\mathbf{S}_*)]=\mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim \mathsf{C}_{\mathbf{X}}\otimes \mathsf{C}_{\mathbf{X}}}[K_\sigma(\xi(\mathbf{S}),\xi(\mathbf{S}_*))]=\mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim \mathsf{C}_{\mathbf{X}}\otimes \mathsf{C}_{\mathbf{X}}}[K_\sigma(\mathbf{S},\mathbf{S}_*)]$$

and $\mathrm{E}_{(\mathbf{S},\mathbf{T})\sim \mathsf{C}_{\xi(\mathbf{X})}\otimes \Pi}[K_\sigma(\mathbf{S},\mathbf{T})] = \mathrm{E}_{(\mathbf{S},\mathbf{T})\sim \mathsf{C}_{\mathbf{X}}\otimes \Pi}[K_\sigma(\xi(\mathbf{S}),\xi(\mathbf{T}))] = \mathrm{E}_{(\mathbf{S},\mathbf{T})\sim \mathsf{C}_{\mathbf{X}}\otimes \Pi}[K_\sigma(\mathbf{S},\mathbf{T})]$.

From these, it follows that $\gamma_{K_\sigma}(\mathsf{C}_{\xi(\mathbf{X})}, \Pi) = \gamma_{K_\sigma}(\mathsf{C}, \Pi)$ and hence $I_\sigma(\xi(\mathbf{X})) = I_\sigma(\mathbf{X})$.

Now, consider any fixed set $A \subseteq \{1, 2, \ldots, p\}$ and a function $\mathbf{f}(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = (f_1(x^{(1)}), f_2(x^{(2)}), \ldots, f_p(x^{(p)}))$ such that for each $i \in A$, $f_i : \mathbb{R} \mapsto \mathbb{R}$ is strictly increasing and for each $i \notin A$, $f_i : \mathbb{R} \mapsto \mathbb{R}$ is strictly decreasing. Also define a function $\mathbf{g}(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = (g_1(x^{(1)}), g_2(x^{(2)}), \ldots, g_p(x^{(p)}))$ with $g_i(x) = x \ \forall i \in A$ and $g_i(x) = 1 - x \ \forall i \notin A$. It can be easily verified that if $\mathbf{S} \sim \mathsf{C}_{\mathbf{f}(\mathbf{X})}$ then $\mathbf{g}(\mathbf{S}) \sim \mathsf{C}_{\mathbf{X}}$. Applying this and the fact that $K_\sigma(\mathbf{S}, \mathbf{S}_*) = K_\sigma(\mathbf{g}(\mathbf{S}), \mathbf{g}(\mathbf{S}_*))$, we get

$$\begin{aligned}\mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim \mathsf{C}_{\mathbf{f}(\mathbf{X})}\otimes \mathsf{C}_{\mathbf{f}(\mathbf{X})}}[K_\sigma(\mathbf{S},\mathbf{S}_*)] &= \mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim \mathsf{C}_{\mathbf{f}(\mathbf{X})}\otimes \mathsf{C}_{\mathbf{f}(\mathbf{X})}}[K_\sigma(\mathbf{g}(\mathbf{S}),\mathbf{g}(\mathbf{S}_*))] \\ &= \mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim \mathsf{C}_{\mathbf{X}}\otimes \mathsf{C}_{\mathbf{X}}}[K_\sigma(\mathbf{S},\mathbf{S}_*)].\end{aligned}$$

By similar argument and using the fact that $\mathbf{T} \sim \Pi$ implies $\mathbf{g}(\mathbf{T}) \sim \Pi$, one gets

$$\mathrm{E}_{(\mathbf{S},\mathbf{T})\sim \mathsf{C}_{\mathbf{f}(\mathbf{X})}\otimes \Pi}[K_\sigma(\mathbf{S},\mathbf{T})] = \mathrm{E}_{(\mathbf{S},\mathbf{T})\sim \mathsf{C}_{\mathbf{f}(\mathbf{X})}\otimes \Pi}[K_\sigma(\mathbf{g}(\mathbf{S}),\mathbf{g}(\mathbf{T}))] = \mathrm{E}_{(\mathbf{S},\mathbf{T})\sim \mathsf{C}_{\mathbf{X}}\otimes \Pi}[K_\sigma(\mathbf{S},\mathbf{T})].$$

Thus $\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{f}(\mathbf{X})}, \Pi) = \gamma_{K_\sigma}(\mathsf{C}_{\mathbf{X}}, \Pi)$, whence, $I_\sigma(\mathbf{f}(\mathbf{X})) = I_\sigma(\mathbf{X})$, proving the invariance of $I_\sigma(\mathbf{X})$ under strictly monotonic transformations of $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$. □

Proof of Proposition 2.2. Let $\mathbf{X}$ be a random vector with continuous marginals, such that for any j, each $X^{(i)}$, $(i \neq j)$ is a strictly monotonic function of $X^{(j)}$. Then, by Proposition 2.1, we have $I_\sigma(\mathbf{X}) = I_\sigma(\mathbf{Y})$ where $\mathbf{Y} = (X^{(j)}, X^{(j)}, \ldots, X^{(j)})$. But then $\mathsf{C}_{\mathbf{Y}}$ is the maximum copula M. So, by definition, $I_\sigma(\mathbf{Y}) = 1$. □

Proof of Theorem 2.1. It has two steps. At the first step, we prove $C_{\sigma,p} = \gamma^2_{K_\sigma}(\mathsf{M}, \Pi)$. At the second step, we prove $\gamma^2_{K_\sigma}(\mathsf{C}, \Pi) = \left(\frac{\sigma}{\sqrt{2\pi}}\right)^p \int\limits_{\mathbb{R}^p} |\varphi_{\mathsf{C}}(\mathbf{w}) - \varphi_{\Pi}(\mathbf{w})|^2 \exp\left(-\frac{\sigma^2}{2}\mathbf{w}^\intercal\mathbf{w}\right) d\mathbf{w}$. Clearly, proving these two steps will complete the proof.

First step: Note that for $(\mathbf{S}, \mathbf{S}_*, \mathbf{T}, \mathbf{T}_*) \sim \mathsf{M} \otimes \mathsf{M} \otimes \Pi \otimes \Pi$, we have

$$\begin{aligned}
&\gamma^2_{K_\sigma}(\mathsf{M}, \Pi) = \mathrm{E}[K_\sigma(\mathbf{S}, \mathbf{S}_*)] - 2\mathrm{E}[K_\sigma(\mathbf{S}, \mathbf{T})] + \mathrm{E}[K_\sigma(\mathbf{T}, \mathbf{T}_*)] \\
&= \int_0^1 \int_0^1 e^{-\frac{p(u-v)^2}{2\sigma^2}}\, du\, dv - 2\int_0^1 \left[\int_0^1 e^{-\frac{(u-v)^2}{2\sigma^2}}\, du\right]^p dv + \left[\int_0^1 \int_0^1 e^{-\frac{(u-v)^2}{2\sigma^2}}\, du\, dv\right]^p \\
&= \kappa\left(\frac{\sigma}{\sqrt{p}}\right) - 2\int_0^1 \lambda^p(u, \sigma)\, du + \kappa^p(\sigma) = C_{\sigma,p}.
\end{aligned}$$

Second step: We use the well-known formula for Fourier transform of the d-dimensional Gaussian density:

$$\exp\left(-\frac{1}{2\sigma^2}\mathbf{x}^\intercal\mathbf{x}\right) = \int\limits_{\mathbb{R}^p} e^{-\sqrt{-1}\mathbf{x}^\intercal\mathbf{w}} \cdot \left(\frac{\sigma}{\sqrt{2\pi}}\right)^p \exp\left(-\frac{\sigma^2}{2}\mathbf{w}^\intercal\mathbf{w}\right) d\mathbf{w},\ \mathbf{x} \in \mathbb{R}^p.$$

This gives us

$$K_\sigma(\mathbf{x}, \mathbf{y}) = \left(\frac{\sigma}{\sqrt{2\pi}}\right)^p \int\limits_{\mathbb{R}^p} e^{-\sqrt{-1}\mathbf{x}^\intercal\mathbf{w}} \cdot e^{\sqrt{-1}\mathbf{y}^\intercal\mathbf{w}} \exp\left(-\frac{\sigma^2}{2}\mathbf{w}^\intercal\mathbf{w}\right) d\mathbf{w},\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^p.$$

Using the representation of γ^2_K from equation (2.1) and Fubini's theorem, one gets

$$\begin{aligned}
\gamma^2_{K_\sigma}(\mathsf{C}, \Pi) = \left(\frac{\sigma}{\sqrt{2\pi}}\right)^p \int\limits_{\mathbb{R}^p} \Big[&\varphi_{\mathsf{C}}(\mathbf{w})\overline{\varphi_{\mathsf{C}}}(\mathbf{w}) + \varphi_{\Pi}(\mathbf{w})\overline{\varphi_{\Pi}}(\mathbf{w}) \\
&- 2\varphi_{\mathsf{C}}(\mathbf{w})\overline{\varphi_{\Pi}}(\mathbf{w}) \Big] \exp\left(-\frac{\sigma^2}{2}\mathbf{w}^\intercal\mathbf{w}\right) d\mathbf{w},
\end{aligned}$$

from which the second part follows. □

Lemma 2.1. *Let $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}_*, \mathbf{Y}_*)$ be independent and identically distributed random vectors taking values in $\mathcal{X} \times \mathcal{Y}$. Given symmetric measurable functions $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\overline{k} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, define*

$$\begin{aligned}
V &= k(\mathbf{X}, \mathbf{X}_*) - \mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|\mathbf{X}\right] - \mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|\mathbf{X}_*\right] + \mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\right] \\
W &= \overline{k}(\mathbf{Y}, \mathbf{Y}_*) - \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|\mathbf{Y}\right] - \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|\mathbf{Y}_*\right] + \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\right].
\end{aligned}$$

Then, we have $\mathrm{E}[VW] = \mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\,\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\right] - 2\,\mathrm{E}\left[\mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|\mathbf{X}\right]\mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|\mathbf{Y}\right]\right]$ $+ \mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\right]\mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\right]$.

Proof. The proof is based on expanding the product VW and then taking term-by-term expectations. One and only one term gives $\mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\,\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\right]$. The seven terms,

where at least one of $\mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\right]$ or $\mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\right]$ appear as a factor, and the two terms $\mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|\mathbf{X}\right] \cdot \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|\mathbf{Y}_*\right]$ and $\mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|\mathbf{X}_*\right] \cdot \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|\mathbf{Y}\right]$, will all give the same expectation $\mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\right] \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\right]$ (the last two because of independence of $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}_*, \mathbf{Y}_*)$). Taking into account the signs of these nine terms with the same expectation, we would be left with just one with a positive sign. Next, the remaining six terms will all have the same expectation, namely, $\mathrm{E}\left[\mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|\mathbf{X}\right] \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|\mathbf{Y}\right]\right]$. For two of the terms, this is straightforward. But the other four terms need judicious use of properties of conditional expectation. For example, by independence of $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}_*, \mathbf{Y}_*)$, we have $\mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|\mathbf{Y}\right] = \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|(\mathbf{X}, \mathbf{Y})\right]$ and similarly $\mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|(\mathbf{X}, \mathbf{Y})\right] = \mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|\mathbf{X}\right]$. Using these, we get

$$\begin{aligned}
\mathrm{E}\left[\, k(\mathbf{X}, \mathbf{X}_*)\mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|\mathbf{Y}\right]\,\right] &= \mathrm{E}\left[\, k(\mathbf{X}, \mathbf{X}_*)\mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|(\mathbf{X}, \mathbf{Y})\right]\,\right] \\
&= \mathrm{E}\left[\, \mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|(\mathbf{X}, \mathbf{Y})\right] \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|(\mathbf{X}, \mathbf{Y})\right]\,\right] \\
&= \mathrm{E}\left[\, \mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|\mathbf{X}\right] \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|\mathbf{Y}\right]\,\right].
\end{aligned}$$

One can similarly handle other three terms. Considering the signs of these six terms with the same expectation, one is left with $-2\mathrm{E}\left[\mathrm{E}\left[k(\mathbf{X}, \mathbf{X}_*)\middle|\mathbf{X}\right] \mathrm{E}\left[\overline{k}(\mathbf{Y}, \mathbf{Y}_*)\middle|\mathbf{Y}\right]\right]$. This completes the proof. □

Proof of Theorem 2.2. (a) By definition, $V^{(1)}$ and $V^{(2)}$ have zero means. Using Lemma 2.1 with kernels $k = \overline{k} = K_\sigma$ on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^p$, we get

$$\begin{aligned}
&\mathrm{Cov}[V^{(1)}, V^{(2)}] = \mathrm{E}[V^{(1)}V^{(2)}] \\
&= \mathrm{E}\left[K_\sigma(T_1^{(1)}, T_2^{(1)})K_\sigma(T_1^{(2)}, T_2^{(2)})\right] + \mathrm{E}\left[K_\sigma(T_1^{(1)}, T_2^{(1)})\right] \mathrm{E}\left[K_\sigma(T_1^{(2)}, T_2^{(2)})\right] \\
&\qquad - 2\,\mathrm{E}\left[\mathrm{E}\left[K_\sigma(T_1^{(1)}, T_2^{(1)})\middle|T_1^{(1)}\right] \mathrm{E}\left[K_\sigma(T_1^{(2)}, T_2^{(2)})\middle|T_1^{(2)}\right]\right] \\
&= \gamma^2_{K_\sigma}(\mathsf{C}_{(X^{(1)}, X^{(2)})}, \Pi).
\end{aligned}$$

Similarly, one can show that $\mathrm{Var}[V^{(1)}] = \mathrm{Var}[V^{(2)}] = \gamma^2_{K_\sigma}(\mathsf{M}, \Pi)$ and hence $I^2_\sigma(\mathbf{X}) = \mathrm{Cor}[V^{(1)}, V^{(2)}]$. The inequality $I_\sigma(X^{(1)}, X^{(2)}) \leq 1$ follows from it. Now, from the condition for equality in Cauchy-Schwartz inequality and the fact that $V^{(1)}$ and $V^{(2)}$ are identically distributed, it follows that $I_\sigma(X^{(1)}, X^{(2)}) = 1$ if and only if $V^{(1)} = V^{(2)}$ almost surely.

Since $T_1^{(1)}$ and $T_2^{(1)}$ are independent and uniformly distributed random variables on $[0, 1]$ and so also are $T_1^{(2)}$ and $T_2^{(2)}$, it follows that $V^{(1)} = V^{(2)}$ (a.s.) if and only if $g(T_1^{(1)}, T_2^{(1)}) =$

$g(T_1^{(2)}, T_2^{(2)})$ (a.s.), where $g(x,y) = K_\sigma(x,y) - \lambda(x,\sigma) - \lambda(y,\sigma)$ with $\lambda(\cdot,\sigma)$ as defined in Theorem 2.1.

Now, using the facts that $(T_1^{(1)}, T_1^{(2)})$ and $(T_2^{(1)}, T_2^{(2)})$ are independent and identically distributed with values in $[0,1]^2$ and that the function g is uniformly continuous on the compact set $[0,1]^2$, one can easily deduce that $g(T_1^{(1)}, T_2^{(1)}) = g(T_1^{(2)}, T_2^{(2)})$ a.s. implies $g(T_1^{(1)}, T_1^{(1)}) = g(T_1^{(2)}, T_1^{(2)})$ a.s. But, this, in turn, implies that

$$\Phi\left(\frac{T_1^{(1)}}{\sigma}\right) + \Phi\left(\frac{1-T_1^{(1)}}{\sigma}\right) = \Phi\left(\frac{T_1^{(2)}}{\sigma}\right) + \Phi\left(\frac{1-T_1^{(2)}}{\sigma}\right) \quad \text{almost surely.}$$

From this, we get $\Pr\left[T_1^{(2)} = T_1^{(1)} \text{ or } T_1^{(2)} = 1 - T_1^{(1)}\right] = 1$ and $\Pr\left[\lambda(T_1^{(1)},\sigma) = \lambda(T_1^{(2)},\sigma)\right] = 1$. Of course, the same would be true for the pair $(T_2^{(1)}, T_2^{(2)})$, which is moreover independent of the pair $(T_1^{(1)}, T_1^{(2)})$.

Using these in the equality $g(T_1^{(1)}, T_2^{(1)}) = g(T_1^{(2)}, T_2^{(2)})$ a.s., one obtains $K_\sigma(T_1^{(1)}, T_2^{(1)}) = K_\sigma(T_1^{(2)}, T_2^{(2)})$ a.s., which implies that $|T_1^{(1)} - T_2^{(1)}| = |T_1^{(2)} - T_2^{(2)}|$ a.s. So, we conclude that either $T_1^{(2)} = T_1^{(1)}$ a.s. or $T_1^{(2)} = 1 - T_1^{(1)}$ a.s. Thus the copula distribution of $(X^{(1)}, X^{(2)})$ is either the distribution of $(T_1^{(1)}, T_1^{(1)})$ or that of $(T_1^{(1)}, 1 - T_1^{(1)})$, where $T_1^{(1)}$ is uniformly distributed on $[0,1]$. Therefore, $X^{(1)}$ and $X^{(2)}$ are almost surely strictly monotone functions of each other.

(b) For $|r| < 1$, let ϕ_r denote the density of the standard bivariate normal distribution with correlation coefficient r. Also, let Φ and ϕ denote respectively the cumulative distribution function and the density function of the standard univariate normal distribution. It is well-known that the copula distribution of any bivariate normal distribution with correlation coefficient r is the same as that of the standard bivariate normal distribution with the same correlation coefficient. Using the well-known Mehler's representation (see Kibble (1945), Page 1) of standard bivariate normal density with correlation r, one then gets that for $|r| < 1$, the copula distribution $\mathcal{C}(r)$ of any bivariate normal distribution with correlation coefficient r has density the given by

$$\frac{\phi_r(\Phi^{-1}(u), \Phi^{-1}(v))}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))} = \sum_{i=0}^{\infty} \frac{r^i}{i!} H_i((\Phi^{-1}(u))H_i(\Phi^{-1}(v)), \quad (u,v) \in [0,1]^2,$$

where $\{H_i(x), i \geq 0\}$ are the well-known Hermite polynomials. Using this, we get that if $(\mathbf{S}, \mathbf{T}) \sim \mathcal{C}(r_1) \otimes \mathcal{C}(r_2)$ with $|r_1| < 1, |r_2| < 1$, then

$$\mathrm{E}[K_\sigma(\mathbf{S},\mathbf{T})] = \int\limits_{[0,1]^4} e^{-\frac{(s_1-t_1)^2+(s_2-t_2)^2}{2\sigma^2}} \sum_{i=0}^{\infty} \frac{r_1^i}{i!} H_i(\Phi^{-1}(s_1)) H_i(\Phi^{-1}(s_2)) \\ \times \sum_{j=0}^{\infty} \frac{r_2^j}{j!} H_j(\Phi^{-1}(t_1)) H_j(\Phi^{-1}(t_2))\, ds_1\, ds_2\, dt_1\, dt_2 \tag{2.3}$$

We now claim that in the above expression, the double summation and integration can be interchanged. To justify this, we recall that the Hermite polynomials $\{H_i(\cdot), i \geq 0\}$ form a complete orthonormal basis for $L_2(\mathbb{R}, \phi(x)dx)$ and, in particular, for any $i \geq 0$,

$$\int\limits_0^1 \left|H_i(\Phi^{-1}(s))\right| \, ds = \int\limits_{\mathbb{R}} |H_i(x)|\, \phi(x)\, dx \leq \left[\int\limits_{\mathbb{R}} H_i^2(x)\phi(x)\, dx\right]^{\frac{1}{2}} = 1.$$

. As a consequence, we have

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \int\limits_{[0,1]^4} \left|e^{-\frac{(s_1-t_1)^2+(s_2-t_2)^2}{2\sigma^2}}\right| \left|\frac{r_1^i r_2^j}{i!j!}\right| \left|H_i(\Phi^{-1}(s_1))\right| \left|H_i(\Phi^{-1}(s_2))\right| \\ \times \left|H_j(\Phi^{-1}(t_1))\right| \left|H_j(\Phi^{-1}(t_2))\right| \, ds_1\, ds_2\, dt_1\, dt_2 \\ \leq \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{1}{i!j!} \left[\int_0^1 \left|H_i(\Phi^{-1}(s))\right| ds\right]^2 \left[\int_0^1 \left|H_j(\Phi^{-1}(t))\right| dt\right]^2 \leq \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{1}{i!j!} < \infty.$$

Therefore we can interchange the double summation and integration on the right-hand-side of the equation (2.3) above to obtain that, for any r_1, r_2 with $|r_1| < 1, |r_2| < 1$,

$$\mathrm{E}_{(\mathbf{S},\mathbf{T})\sim\mathcal{C}(r_1)\otimes\mathcal{C}(r_2)}[K_\sigma(\mathbf{S},\mathbf{T})] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_{i,j} r_1^i r_2^j,$$

$$\text{where} \quad a_{i,j} := \frac{1}{i!j!} \left[\int\limits_{[0,1]^2} e^{-\frac{(u-v)^2}{2\sigma^2}} H_i(\Phi^{-1}(s)) H_j(\Phi^{-1}(t))\, ds\, dt\right]^2 \\ = \frac{1}{i!j!} \left[\int\limits_{\mathbb{R}^2} e^{-\frac{(\Phi(x)-\Phi(y))^2}{2\sigma^2}} H_i(x) H_j(y) \phi(x)\phi(y)\, dx\, dy\right]^2. \tag{2.4}$$

Observe that $a_{i,j} \geq 0, a_{i,j} = a_{j,i}$ and also, $(i!j!)a_{i,j} \leq \left[\int\limits_{\mathbb{R}} |H_i(x)|\, \phi(x)\, dx\right]^2 \leq 1$. Note that for any bivariate normal random vector $(X^{(1)}, X^{(2)})$ with correlation coefficient r (where $|r| < 1$), we have $\gamma^2_{K_\sigma}(\mathcal{C}_{(X^{(1)},X^{(2)})}, \Pi) = \gamma^2_{K_\sigma}(\mathcal{C}(r), \mathcal{C}(0))$, which equals

$$\mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim\mathcal{C}(r)\otimes\mathcal{C}(r)}[K_\sigma(\mathbf{S},\mathbf{S}_*)] - 2\mathrm{E}_{(\mathbf{S},\mathbf{T})\sim\mathcal{C}(r)\otimes\mathcal{C}(0)}[K_\sigma(\mathbf{S},\mathbf{T})] + \mathrm{E}_{(\mathbf{T},\mathbf{T}_*)\sim\mathcal{C}(0)\otimes\mathcal{C}(0)}[K_\sigma(\mathbf{T},\mathbf{T}_*)] \\ = \sum_{k=0}^{\infty} \sum_{\substack{i,j\geq 0 \\ i+j=k}} a_{i,j} r^k - 2\sum_{k=0}^{\infty} a_{k,0}\, r^k + a_{0,0} = \sum_{k=1}^{\infty} \sum_{\substack{i,j\geq 1 \\ i+j=k}} a_{i,j}\, r^k \underset{(a)}{=} \sum_{k=1}^{\infty} \sum_{\substack{i,j\geq 1 \\ i+j=2k}} a_{i,j}\, r^{2k}.$$

Equality (a) is due to the fact that the ith Hermite polynomial H_i is an even or an odd

function according as i is even or odd, so that if exactly one of i and j is odd, then $a_{i,j} = 0$, as can easily be seen from equation (2.4). Therefore, we have

$$I_\sigma^2(X^{(1)}, X^{(2)}) = \gamma_{K_\sigma}^{-2}(\mathsf{M}, \Pi)\gamma_{K_\sigma}^2(\mathcal{C}_{(X^{(1)},X^{(2)})}, \Pi) = \gamma_{K_\sigma}^{-2}(\mathsf{M}, \Pi) \sum_{k=1}^{\infty} \sum_{i,j\geq 1, i+j=2k} a_{i,j}\, r^{2k}.$$

So, $I_\sigma^2(X^{(1)}, X^{(2)}) = r^2 g(r)$, where $g(r) = \gamma_{K_\sigma}^{-2}(\mathsf{M}, \Pi) \sum_{k=1}^{\infty} \sum_{i,j\geq 1: i+j=2k} a_{i,j}\, r^{2(k-1)}$ is a power series in r^2 with positive coefficients and hence increasing in $|r|$. So, $I_\sigma^2(X^{(1)}, X^{(2)}) = r^2 \cdot g(r)$ is an increasing function of $|r|$. □

Proof of Theorem 2.3. It is enough to show that for every dimension p (≥ 3), there exist two p dimensional copulas C_1 and C_2 with $\mathcal{M}(C_1) \neq \mathcal{M}(C_2)$, such that for any choice of coordinates $\{i_1, i_2, \ldots, i_k\} \subsetneqq \{1, 2, \ldots, p\}$, if C_1' and C_2' are the associated marginal copulas arising out of of C_1 and C_2, then $\mathcal{M}(C_1') = \mathcal{M}(C_2')$.

Take C_1 to be the p-dimensional uniform copula Π. Then $\mathcal{M}(C_1) = 0$, and also for any lower dimensional marginal copula C_1' of C_1, $\mathcal{M}(C_1') = 0$. We now exhibit a p-dimensional copula $C_2 \neq \Pi$ such that any lower dimensional marginal copula C_2' of C_2 is uniform copula. We would then have $\mathcal{M}(C_1) = 0 \neq \mathcal{M}(C_2)$ but $\mathcal{M}(C_1') = \mathcal{M}(C_2') = 0$, which will complete the proof.

We take C_2 to be the copula given by the copula density $\mathcal{C}_2$ defined as

$$\mathcal{C}_2(u^{(1)}, u^{(2)}, \ldots, u^{(p)}) = 2\,\mathbb{I}\left[\left(u^{(1)} - \frac{1}{2}\right)\left(u^{(2)} - \frac{1}{2}\right)\cdots\left(u^{(p)} - \frac{1}{2}\right) \geq 0\right],$$

where $\mathbb{I}$ denotes the indicator function. To show that all lower dimensional marginal copulas of C_2 are uniform, it is enough to show that the marginal copula C_2' that we get from C_2 discarding the p-th coordinate, is uniform. Now, note that the density of C_2' is given by

$$\begin{aligned}
\mathcal{C}_2'(u^{(1)}, u^{(2)}, \cdots, u^{(p-1)}) &= \int_0^1 2\mathbb{I}\left[\left(u^{(1)} - \frac{1}{2}\right)\left(u^{(2)} - \frac{1}{2}\right)\cdots\left(u^{(p)} - \frac{1}{2}\right) \geq 0\right] du^{(p)} \\
&= \int_0^{\frac{1}{2}} 2\mathbb{I}\left[\left(u^{(1)} - \frac{1}{2}\right)\left(u^{(2)} - \frac{1}{2}\right)\cdots\left(u^{(p-1)} - \frac{1}{2}\right) \leq 0\right] du^{(p)} \\
&\quad + \int_{\frac{1}{2}}^1 2\mathbb{I}\left[\left(u^{(1)} - \frac{1}{2}\right)\left(u^{(2)} - \frac{1}{2}\right)\cdots\left(u^{(p-1)} - \frac{1}{2}\right) \geq 0\right] du^{(p)} \\
&= \mathbb{I}\left[\left(u^{(1)} - \frac{1}{2}\right)\left(u^{(2)} - \frac{1}{2}\right)\cdots\left(u^{(p-1)} - \frac{1}{2}\right) \leq 0\right] \\
&\quad + \mathbb{I}\left[\left(u^{(1)} - \frac{1}{2}\right)\left(u^{(2)} - \frac{1}{2}\right)\cdots\left(u^{(p-1)} - \frac{1}{2}\right) \geq 0\right] = 1.
\end{aligned}$$

□

Proof of Proposition 2.4. We shall prove that $\sigma^4\gamma_{K_\sigma}^2(\mathsf{C},\Pi) \to \sum_{1\le i<j\le p}\mathrm{Cov}^2(T^{(i)},T^{(j)})$ as $\sigma\to\infty$. It will imply that $\sigma^4\gamma_{K_\sigma}^2(\mathsf{M},\Pi)\to\binom{p}{2}\mathrm{Var}^2(T^{(1)})$ as $\sigma\to\infty$, which in turn will lead to our desired result

$$I_\sigma^2(\mathbf{X}) = \frac{\sigma^4\gamma_{K_\sigma}^2(\mathsf{C},\Pi)}{\sigma^4\gamma_{K_\sigma}^2(\mathsf{M},\Pi)} \to \frac{1}{\binom{p}{2}}\sum_{1\le i<j\le p}\frac{\mathrm{Cov}^2(T^{(i)},T^{(j)})}{\mathrm{Var}^2(T^{(1)})} = \frac{1}{\binom{p}{2}}\sum_{1\le i<j\le p}\frac{\mathrm{Cov}^2(T^{(i)},T^{(j)})}{\mathrm{Var}(T^{(i)})\mathrm{Var}(T^{(j)})}$$
$$= \frac{1}{\binom{p}{2}}\sum_{1\le i<j\le p}\mathrm{Cor}^2(T^{(i)},T^{(j)}) \ \text{ as } \ \sigma\to\infty.$$

Observe that $\mathrm{E}K_\sigma(\mathbf{T},\mathbf{S}) = 1 - \frac{1}{2\sigma^2}\mathrm{E}\|\mathbf{T}-\mathbf{S}\|_2^2 + \frac{1}{8\sigma^4}\mathrm{E}\|\mathbf{T}-\mathbf{S}\|_2^4 + \mathcal{O}\left(\frac{1}{\sigma^6}\right)$. Assume that $\mathbf{T},\mathbf{T}_*,\mathbf{S}$ and $\mathbf{S}_*$ are four random vectors such that $(\mathbf{T},\mathbf{T}_*,\mathbf{S},\mathbf{S}_*)\sim\mathsf{C}\otimes\mathsf{C}\otimes\Pi\otimes\Pi$. Then,

$$\begin{aligned}\gamma_{K_\sigma}^2(\mathsf{C},\Pi) &= \mathrm{E}K_\sigma(\mathbf{T},\mathbf{T}_*) - 2\mathrm{E}K_\sigma(\mathbf{T},\mathbf{S}) + \mathrm{E}K_\sigma(\mathbf{S},\mathbf{S}_*)\\ &= -\frac{1}{2\sigma^2}\mathrm{E}\left[\|\mathbf{T}-\mathbf{T}_*\|^2 + \|\mathbf{S}-\mathbf{S}_*\|^2 - 2\|\mathbf{T}-\mathbf{S}\|^2\right]\\ &\quad + \frac{1}{8\sigma^4}\mathrm{E}\left[\|\mathbf{T}-\mathbf{T}_*\|^4 + \|\mathbf{S}-\mathbf{S}_*\|^4 - 2\|\mathbf{T}-\mathbf{S}\|^4\right] + \mathcal{O}\left(\frac{1}{\sigma^6}\right).\end{aligned}$$

Now, $\mathrm{E}\left[\|\mathbf{T}-\mathbf{T}_*\|^2 + \|\mathbf{S}-\mathbf{S}_*\|^2 - 2\|\mathbf{T}-\mathbf{S}\|^2\right]$

$$= \mathrm{E}\sum_{i=1}^p\left[(T^{(i)}-T_*^{(i)})^2 + (S^{(i)}-S_*^{(i)})^2 - 2(T^{(i)}-S^{(i)})^2\right] = 0 \ \text{ and}$$

$$\begin{aligned}&\mathrm{E}\left[\|\mathbf{T}-\mathbf{T}_*\|^4 + \|\mathbf{S}-\mathbf{S}_*\|^4 - 2\|\mathbf{T}-\mathbf{S}\|^4\right]\\ &= \mathrm{E}\sum_{i=1}^p\left[(T^{(i)}-T_*^{(i)})^4 + (S^{(i)}-S_*^{(i)})^4 - 2(T^{(i)}-S^{(i)})^4\right]\\ &\quad + 2\mathrm{E}\sum_{1\le i<j\le p}\Big[(T^{(i)}-T_*^{(i)})^2(T^{(j)}-T_*^{(j)})^2 + (S^{(i)}-S_*^{(i)})^2(S^{(j)}-S_*^{(j)})^2\\ &\qquad\qquad - 2(T^{(i)}-S^{(i)})^2(T^{(j)}-S^{(j)})^2\Big]\\ &= 2\mathrm{E}\sum_{1\le i<j\le p}\Big[(T^{(i)}-T_*^{(i)})^2(T^{(j)}-T_*^{(j)})^2 + (S^{(i)}-S_*^{(i)})^2(S^{(j)}-S_*^{(j)})^2\\ &\qquad\qquad - 2(T^{(i)}-S^{(i)})^2(T^{(j)}-S^{(j)})^2\Big].\end{aligned}$$

Hence, we have

$$\begin{aligned}\gamma_{K_\sigma}^2(\mathsf{C},\Pi) =& \frac{1}{4\sigma^4}\mathrm{E}\sum_{1\le i<j\le p}\Big[(T^{(i)}-T_*^{(i)})^2(T^{(j)}-T_*^{(j)})^2 + (S^{(i)}-S_*^{(i)})^2(S^{(j)}-S_*^{(j)})^2\\ &\qquad - 2(T^{(i)}-S^{(i)})^2(T^{(j)}-S^{(j)})^2\Big] + \mathcal{O}\left(\frac{1}{\sigma^6}\right).\end{aligned}$$

Now, using some straight-forward but tedious calculations, it can be shown that $\mathrm{E}\Big[(T^{(i)}-T_*^{(i)})^2(T^{(j)}-T_*^{(j)})^2 + (S^{(i)}-S_*^{(i)})^2(S^{(j)}-S_*^{(j)})^2 - 2(T^{(i)}-S^{(i)})^2(T^{(j)}-S^{(j)})^2\Big] = 4\mathrm{Cov}^2(T^{(i)},T^{(j)})$. This implies that $\sigma^4\gamma_{K_\sigma}^2(\mathsf{C},\Pi)\to\sum_{1\le i<j\le p}\mathrm{Cov}^2(T^{(i)},T^{(j)})$ as $\sigma\to\infty$. □

Proof of Proposition 2.5. (a) Clearly, applying a permutation to the coordinates of the observation vectors $\mathbf{x}_i, i = 1, 2, \ldots, n$, changes the coordinates of the normalized rank vectors $\mathbf{y}_i$'s by the same permutation. Since s_1 and s_2 from Equation (2.2) are both invariant under permutation of coordinates of the $\mathbf{y}_i$'s, the proof is complete.

For proving invariance under monotonic transformation, it is enough to consider the case when only one of the coordinates in the observation vectors is changed by a strictly monotonic non-constant transformation. Therefore, assume that only the s-th coordinate of the $\mathbf{x}_i$'s is changed by a strictly monotonic transformation, while the other coordinates are kept the same. This will affect only the s-th coordinate of the $\mathbf{y}_i$'s. Denoting the changed $\mathbf{y}_i$'s as $\mathbf{y}_i^*$'s, it is clear that $y_i^{*(s)}$ will equal $y_i^{(s)}$ or $1 + \frac{1}{n} - y_i^{(s)}$ for all $i = 1, 2, \ldots, n$, according as the transformation is strictly increasing or strictly decreasing. In either case, $K_\sigma(\mathbf{y}_i^*, \mathbf{y}_j^*) = K_\sigma(\mathbf{y}_i, \mathbf{y}_j)$, so that s_1 in Equation (2.2) remains unchanged. One can easily see that s_2 also remains unchanged as well.

(b) Without loss of generality, we may assume that the first coordinates of the $\mathbf{x}_i$'s are in ascending order. Now, suppose that every other coordinate of the $\mathbf{x}_i$'s is in a strictly monotonic relation with the first coordinate; then, for $j = 2, 3, \ldots, p$, the j-th coordinates of the $\mathbf{x}_i$'s will be in either ascending or descending order. By monotonic transformation invariance property, we may assume, without loss of generality, that all the coordinates of the $\mathbf{x}_i$'s are in ascending order. But then, the $\mathbf{y}_i$'s are clearly given by $y_i^{(j)} = \frac{i}{n}$, for all j and one can then see that $s_1 = v_1$ and $s_2 = v_2$, whence it follows that $\widehat{I}_{\sigma,n}(\mathbf{X}) = 1$. □

Proof of Theorem 2.4. For independent random vectors $(T_1^{(1)}, T_1^{(2)})$ and $(T_2^{(1)}, T_2^{(2)})$ both having distribution C_n, one has

$$\begin{aligned}\gamma_{K_\sigma}^2(\mathsf{C}_n, \Pi_n) =& \mathrm{E}\left[K_\sigma(T_1^{(1)}, T_2^{(1)}) K_\sigma(T_1^{(2)}, T_2^{(2)})\right] \\ &- 2\mathrm{E}\left[\mathrm{E}\left[K_\sigma(T_1^{(1)}, T_2^{(1)}) \middle| T_1^{(1)}\right] \mathrm{E}\left[K_\sigma(T_1^{(2)}, T_2^{(2)}) \middle| T_1^{(2)}\right]\right] \\ &+ \mathrm{E}\left[K_\sigma(T_1^{(1)}, T_2^{(1)})\right] \mathrm{E}\left[K_\sigma(T_1^{(2)}, T_2^{(2)})\right].\end{aligned}$$

Using the above, and from Lemma 2.1, we get $\gamma_{K_\sigma}^2(\mathsf{C}_n, \Pi_n) = \mathrm{E}[V^{(1)} V^{(2)}]$, where

$$V^{(i)} = K_\sigma(T_1^{(i)}, T_2^{(i)}) - \mathrm{E}\left[K_\sigma(T_1^{(i)}, T_2^{(i)}) \middle| T_1^{(i)}\right] - \mathrm{E}\left[K_\sigma(T_1^{(i)}, T_2^{(i)}) \middle| T_2^{(i)}\right] + \mathrm{E}\left[K_\sigma(T_1^{(i)}, T_2^{(i)})\right]$$

for $i = 1, 2$. Thus, we have $\gamma_{K_\sigma}^2(\mathsf{C}_n, \Pi_n) = \frac{1}{n^2} \sum\limits_{1 \le i,j \le n} v_{i,j}^{(1)} v_{i,j}^{(2)}$. Similarly, one can show that $\gamma_{K_\sigma}^2(\mathsf{M}_n, \Pi_n) = \frac{1}{n^2} \sum\limits_{1 \le i,j \le n} \left(v_{i,j}^{(1)}\right)^2 = \frac{1}{n^2} \sum\limits_{1 \le i,j \le n} \left(v_{i,j}^{(2)}\right)^2$.

Cauchy-Schwartz inequality immediately gives $\widehat{I}_{\sigma,n}(\mathbf{X})^2 \leq 1$. Further, by the necessary and sufficient condition for equality in the Cauchy-Schwartz inequality and using the fact that $\sum_{1\leq i,j\leq n}\left(v_{i,j}^{(1)}\right)^2 = \sum_{1\leq i,j\leq n}\left(v_{i,j}^{(2)}\right)^2$, one gets that $\widehat{I}_{\sigma,n}(\mathbf{X}) = 1$ if and only if $v_{i,j}^{(1)} = v_{i,j}^{(2)}\ \forall\ i,j$.

Now, if one coordinate of the observation vectors is a monotone function of the other coordinate, then either $y_i^{(2)} = y_i^{(1)}\ \forall\ i$ or $y_i^{(2)} = \frac{n+1}{n} - y_i^{(1)}\ \forall\ i$. In either case, $|y_i^{(1)} - y_j^{(1)}| = |y_i^{(2)} - y_j^{(2)}|\ \forall\ i,j$, which will clearly imply that $v_{i,j}^{(1)} = v_{i,j}^{(2)}\ \forall\ i,j$.

To prove the converse, first observe that for any i,

$$\sum_{l=1}^{n} K_\sigma(y_i^{(1)}, y_l^{(1)}) = \sum_{l=1}^{n} K_\sigma(y_l^{(1)}, y_i^{(1)}) = \sum_{l=1}^{n} K_\sigma(y_i^{(1)}, l/n) \text{ and}$$
$$\sum_{l=1}^{n} K_\sigma(y_i^{(2)}, y_l^{(2)}) = \sum_{l=1}^{n} K_\sigma(y_l^{(2)}, y_i^{(2)}) = \sum_{l=1}^{n} K_\sigma(y_i^{(2)}, l/n).$$

Now suppose that $v_{i,j}^{(1)} = v_{i,j}^{(2)}\ \forall\ i,j$. Then, taking $i = j$, one deduces that

$$\sum_{l=1}^{n} K_\sigma(y_i^{(1)}, l/n) = \sum_{l=1}^{n} K_\sigma(y_i^{(2)}, l/n) \quad \forall\ i \in \{1,2,\ldots,n\}. \tag{2.5}$$

Using this now in $v_{i,j}^{(1)} = v_{i,j}^{(2)}$, one gets

$$K_\sigma(y_i^{(1)}, y_j^{(1)}) = K_\sigma(y_i^{(2)}, y_j^{(2)}),\ \text{i.e.,}\ |y_i^{(1)} - y_j^{(1)}| = |y_i^{(2)} - y_j^{(2)}|\ \ \forall\ i,j. \tag{2.6}$$

We now claim that for $i, i^{'} \in \{1,2,\ldots,n\}$, $\sum_{l=1}^{n} K_\sigma(i/n, l/n) = \sum_{l=1}^{n} K_\sigma(i^{'}/n, l/n)$ if and only if either $i^{'} = i$ or $i^{'} = n+1-i$. The 'if' part is easy to see; if $i^{'} = n+1-i$, the equality is obtained by observing that $K_\sigma(i^{'}/n, j/n) = K_\sigma(i/n, (n+1-j)/n)\ \forall\ j$ and then making a change of variable ($j \mapsto n+1-j$) in the summation. The 'only if' part can now be completed by observing that whenever $i < n+1-i$, $\sum_{l=1}^{n} K_\sigma((i+1)/n, l/n) - \sum_{l=1}^{n} K_\sigma(i/n, l/n) = e^{-\frac{i^2}{2n^2\sigma^2}} - e^{-\frac{(n-i)^2}{2n^2\sigma^2}} > 0$, implying that $\sum_{l=1}^{n} K_\sigma(i/n, l/n)$ is strictly increasing in i whenever $i < n+1-i$.

Using this, (2.5) implies that for each i, we have either $y_i^{(2)} = y_i^{(1)}$ or $y_i^{(2)} = 1 + \frac{1}{n} - y_i^{(1)}$. Next, let i be such that $y_i^{(1)} = 1/n$. We know that either $y_i^{(2)} = y_i^{(1)}$ or $y_i^{(2)} = 1 + 1/n - y_i^{(1)}$. Suppose first that $y_i^{(2)} = y_i^{(1)}$. Now, take any $j \neq i$. We know $y_j^{(2)}$ equals either $y_j^{(1)}$ or $1+1/n-y_j^{(1)}$. But then (2.6) rules out the possibility that $y_j^{(2)} = 1+1/n-y_j^{(1)}$. Thus we have $y_j^{(2)} = y_j^{(1)}$ for all j. Similarly, if $y_i^{(2)} = 1+1/n-y_i^{(1)}$, one can show that $y_j^{(2)} = 1+1/n-y_j^{(1)}$ for all j. Thus we conclude that either $y_j^{(2)} = y_j^{(1)}\ \forall\ j$ or $y_j^{(2)} = 1 + 1/n - y_j^{(1)}\ \forall\ j$. But this means that one coordinate of the observation vectors is either an increasing or a decreasing function of the other coordinate. □

The following well-known result, which can be found in Tsukahara (2005), is crucial in our derivation of the limiting distributions of $\widehat{I}_{\sigma,n}(\mathbf{X})$ in both under the null and the alternative hypotheses.

Theorem 2.9 (Weak convergence of copula process)**.** *Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be independent observations on the random vector $\mathbf{X}$ with copula distribution C and let C_n be the empirical copula based on $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. If for all $i = 1, 2, \ldots, p$, the i-th partial derivatives $D_i\mathsf{C}(\mathbf{u})$ of C exist and are continuous, then the process $\sqrt{n}(\mathsf{C}_n - \mathsf{C})$ converges weakly in $l^\infty([0,1]^p)$ to the process $\mathbb{G}_\mathsf{C}$ given by $\mathbb{G}_\mathsf{C}(\mathbf{u}) = \mathbb{B}_\mathsf{C}(\mathbf{u}) - \sum_{i=1}^p D_i\mathsf{C}(\mathbf{u})\mathbb{B}_\mathsf{C}(\mathbf{u}_{(i)})$, where $\mathbb{B}_\mathsf{C}$ is a p-dimensional Brownian bridge on $[0,1]^p$ with covariance function $\mathrm{E}[\mathbb{B}_\mathsf{C}(\mathbf{u})\mathbb{B}_\mathsf{C}(\mathbf{v})] = \mathsf{C}(\mathbf{u}) \wedge \mathsf{C}(\mathbf{v}) - \mathsf{C}(\mathbf{u})\mathsf{C}(\mathbf{v})$, and for each i, $\mathbf{u}_{(i)}$ denotes the vector obtained from $\mathbf{u}$ by replacing its all coordinates, except the i-th one, by 1.*

For $i = 1, 2, \ldots, n$, define vectors $\mathbf{z}_i = (z_i^{(1)}, z_i^{(2)}, \ldots, z_i^{(p)})$ such that for $j = 1, 2, \ldots, p$, $z_i^{(j)} = F_j(x_i^{(j)})$. Denote the empirical distribution based on $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ by $\mathsf{C}_{\mathbf{z},n}$.

Lemma 2.2. *Assume that $\{P_n\}_{n\geq 1}$ is a sequence of distributions over $[0,1]^p$. Then*

1. $\left|\gamma^2_{K_\sigma}(\Pi_n, P_n) - \gamma^2_{K_\sigma}(\Pi, P_n)\right| = \mathcal{O}(n^{-2})$
2. $\left|\gamma^2_{K_\sigma}(\mathsf{M}_n, P_n) - \gamma^2_{K_\sigma}(\mathsf{M}, P_n)\right| = \mathcal{O}(n^{-2})$.

Proof. We prove the first part only. The proof of the second part is similar.

$$\begin{aligned}\left|\gamma^2_{K_\sigma}(\Pi_n, P_n) - \gamma^2_{K_\sigma}(\Pi, P_n)\right| \leq{}& \left|\mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim\Pi_n\otimes\Pi_n}[K_\sigma(\mathbf{S},\mathbf{S}_*)] - \mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim\Pi\otimes\Pi}[K_\sigma(\mathbf{S},\mathbf{S}_*)]\right| \\ &+ 2\mathrm{E}_{\mathbf{T}\sim P_n}\left|\mathrm{E}_{\mathbf{S}\sim\Pi_n}[K_\sigma(\mathbf{S},\mathbf{T})] - \mathrm{E}_{\mathbf{S}\sim\Pi}[K_\sigma(\mathbf{S},\mathbf{T})]\right|.\end{aligned}$$

One can notice that the first term on the right hand side of the above inequality is bounded above by

$$\frac{1}{n^{2p}} \sum_{\substack{\boldsymbol{\mu}=(i_1/n, i_2/n, \ldots, i_p/n) \\ 1\leq i_1, i_2, \ldots, i_p \leq n}} \sum_{\substack{\boldsymbol{\nu}=(j_1/n, j_2/n, \ldots, j_p/n) \\ 1\leq j_1, j_2, \ldots, j_p \leq n}} \int_{[\boldsymbol{\mu}-1/n\mathbf{1}, \boldsymbol{\mu}]} \int_{[\boldsymbol{\nu}-1/n\mathbf{1}, \boldsymbol{\nu}]} |K_\sigma(\boldsymbol{\mu}, \boldsymbol{\nu}) - K_\sigma(\boldsymbol{\zeta}, \boldsymbol{\eta})| \, d\boldsymbol{\zeta} \, d\boldsymbol{\eta},$$

where for any $\mathbf{u} = (u^{(1)}, u^{(2)}, \ldots, u^{(p)}) \in [0,1]^p$ and $\delta > 0$, $[\mathbf{u} - \delta\mathbf{1}, \mathbf{u}]$ denotes the rectangle $[u^{(1)} - \delta, u^{(1)}] \times [u^{(2)} - \delta, u^{(2)}] \times \ldots \times [u^{(p)} - \delta, u^{(p)}]$. The last expression is clearly bounded above by

$$\max_{\substack{\boldsymbol{\mu}=(i_1/n, i_2/n, \ldots, i_p/n) \\ 1\leq i_1, i_2, \ldots, i_p \leq n}} \max_{\substack{\boldsymbol{\nu}=(j_1/n, j_2/n, \ldots, j_p/n) \\ 1\leq j_1, j_2, \ldots, j_p \leq n}} \sup_{\boldsymbol{\zeta}\in[\boldsymbol{\mu}-1/n\mathbf{1}, \boldsymbol{\mu}]} \sup_{\boldsymbol{\eta}\in[\boldsymbol{\nu}-1/n\mathbf{1}, \boldsymbol{\nu}]} |K_\sigma(\boldsymbol{\mu}, \boldsymbol{\nu}) - K_\sigma(\boldsymbol{\zeta}, \boldsymbol{\eta})|.$$

Using Lemma 6 of Póczos *et al.* (2012), one can further deduce that the last expression is bounded above by pn^{-2}. Similar technique can be used for the second term to get the upper bound

$$2\,\mathrm{E}_{\mathbf{T}\sim P_n} \max_{\substack{\boldsymbol{\mu}=(i_1/n,i_2/n,\ldots,i_p/n)\\ 1\le i_1,i_2,\ldots,i_p\le n}} \sup_{\boldsymbol{\eta}\in[\boldsymbol{\mu}-1/n\mathbf{1},\boldsymbol{\mu}]} |K_\sigma(\boldsymbol{\mu},\mathbf{T}) - K_\sigma(\boldsymbol{\eta},\mathbf{T})| \le 2pn^{-2}.$$

Combining these two bounds, we get $\left|\gamma^2_{K_\sigma}(\Pi_n, P_n) - \gamma^2_{K_\sigma}(\Pi, P_n)\right| = \mathcal{O}(n^{-2})$. □

Lemma 2.3. $\gamma_{K_\sigma}(\mathsf{C}_n, \mathsf{C}_{\mathbf{z},n}) \to 0$ *almost surely as* $n \to \infty$.

Sketch of the proof. Since the essential idea of the proof is contained in Póczos *et al.* (2012) [Appendix E], we only describe the two main steps.

First, we use the definition of $\mathsf{C}_{\mathbf{z},n}$ and Lemma 6 of Póczos *et al.* (2012) to get the inequality $\gamma^2_{K_\sigma}(\mathsf{C}_n, \mathsf{C}_{\mathbf{z},n}) \le 2\sqrt{p}L \max_{1\le j\le p} \sup_{x\in\mathbb{R}} |\hat{F}_j(x) - F_j(x)|$, where $\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_p$ are the empirical distributions of $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$ respectively.

Then using the above inequality and the Kiefer-Dvoretzky-Wolfowitz Theorem (see Massart (1990), Page 1269), for any $\epsilon > 0$, we get $\Pr\left[\gamma^2_{K_\sigma}(\mathsf{C}_n, \mathsf{C}_{\mathbf{z},n}) > \epsilon\right] \le 2p \exp\left(-\frac{n\epsilon^2}{2pL^2}\right)$. The result now follows from the Borel-Cantelli Lemma. □

The next lemma and its proof are based on the ideas in Gretton *et al.* (2012) [see Appendix A2 in Gretton *et al.* (2012)].

Lemma 2.4. $\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C}) \to 0$ *almost surely as* $n \to \infty$.

Proof. It is enough to prove $\Pr\left[\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C}) - \mathrm{E}[\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C})] > \epsilon\right] \le \exp\left(-\frac{n\epsilon^2}{2}\right)$ and $\mathrm{E}\left[\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C})\right] \le \frac{2}{\sqrt{n}}$.

Denoting $\mathcal{F}$ to be the unit ball in the RKHS associated to the kernel K_σ on $\mathbb{R}^p$, one gets $\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C}) = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i) - \mathrm{E}_{\mathbf{Z}\sim\mathsf{C}} f(\mathbf{Z})\right|$ (See, Sriperumbudur *et al.*, 2010).

Let $\mathbf{z}_i^*$ $(i = 1, 2, \ldots, n)$ be independent and identically distributed with same distribution as $\mathbf{z}_i$ $(i = 1, 2, \ldots, n)$ and δ_i $(1, 2, \ldots, n)$ be i.i.d. random variables taking values ± 1 with equal probabilities. If the $\mathbf{z}_i$'s and δ_i's are independent, it is easy to see that

$$\begin{aligned}
\mathrm{E}[\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C})] &= \mathrm{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i) - \mathrm{E}_{\mathbf{Z}\sim\mathsf{C}} f(\mathbf{Z})\right|\right] \le \mathrm{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i) - \frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i^*)\right|\right] \\
&= \mathrm{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \delta_i \left(f(\mathbf{z}_i) - f(\mathbf{z}_i^*)\right)\right|\right] \underset{(a)}{\le} \frac{2}{\sqrt{n}}.
\end{aligned}$$

For the last inequality (a), we used a well-known result referred to as "Bound on Rademacher Complexity" (see Bartlett and Mendelson (2003), Page 478) .

We next calculate the upper bound of change in magnitude due to change in a particular coordinate. Consider $\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C})$ as a function of $\mathbf{z}_i$. It is easy to verify that changing any coordinate of $\mathbf{z}_i$, the change in $\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C})$ will be at most $2n^{-1}$. We use now the well-known McDiarmid's inequality (see McDiarmid (1989), Page 149) to get

$$\Pr\left[\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C}) - \mathrm{E}[\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C})] > \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{n.(2/n)^2}\right) = \exp\left(-\frac{n\epsilon^2}{2}\right). \qquad \square$$

Lemma 2.5. *Suppose that the assumptions of Theorem 2.9 hold. Then, we have the following results.*

(a) If $\mathsf{C} \neq \Pi$, $\sqrt{n}(\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi) - \gamma^2_{K_\sigma}(\mathsf{C}, \Pi)) \xrightarrow{\mathcal{D}} N(0, \delta_0^2)$, *where*
$\delta_0^2 = 4 \int\limits_{[0,1]^p} \int\limits_{[0,1]^p} g(\mathbf{u}) g(\mathbf{v})\, \mathrm{E}[\, d\mathbb{G}_\mathsf{C}(\mathbf{u})\, d\mathbb{G}_\mathsf{C}(\mathbf{v})]$ *and* $g(\mathbf{u}) = \int\limits_{[0,1]^p} K_\sigma(\mathbf{u}, \mathbf{v})\, d(\mathsf{C} - \Pi)(\mathbf{v})$.

(b) If $\mathsf{C} = \Pi$, $n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi) \xrightarrow{\mathcal{D}} \int\limits_{[0,1]^p} \int\limits_{[0,1]^p} K_\sigma(\mathbf{u}, \mathbf{v})\, d\mathbb{G}_\Pi(\mathbf{u})\, d\mathbb{G}_\Pi(\mathbf{v})$, *where* $\mathbb{G}_\Pi$ *is the Gaussian process* $\mathbb{G}_\mathsf{C}$ *for* $\mathsf{C} = \Pi$.

Proof. When $\mathsf{C} \neq \Pi$: Denoting $\mathcal{D}([0,1]^p)$ to be the space of right continuous real valued uniformly bounded functions on $[0,1]^p$ with left limits, equipped with max-sup norm, one can easily verify that the function $\psi(\mathsf{D}) = \gamma^2_{K_\sigma}(\mathsf{D}, \Pi)$ on $\mathcal{D}([0,1]^p)$ is Hadamard-differentiable and the derivative at C is given by

$$\psi'_\mathsf{C}(\mathsf{D}) = 2 \int\limits_{[0,1]^p} \int\limits_{[0,1]^p} K_\sigma(\mathbf{u}, \mathbf{w})\, d(\mathsf{C} - \Pi)(\mathbf{w})\, d\mathsf{D}(\mathbf{u}).$$

To prove this, consider a real sequence $\{t_n\}$ converging to 0 and a $\mathcal{D}([0,1]^p)$-valued sequence $\{D_n\}$ converging to $D \in \mathcal{D}([0,1]^p)$ such that $\mathsf{C} + t_n D_n \in \mathcal{D}([0,1]^p)$. For any $D \in \mathcal{D}([0,1]^p)$, define $\mu_D(\mathbf{w}) = \int\limits_{[0,1]^p} K_{\frac{\sigma}{\sqrt{2}}}(\mathbf{u}, \mathbf{w})\, dD(\mathbf{u})\ \forall \mathbf{w} \in \mathbb{R}^p$. Then

$$\begin{aligned}
&\frac{\psi(\mathsf{C} + t_n D_n) - \psi(\mathsf{C})}{t_n} \\
&\quad = \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^p \frac{1}{t_n} \int\limits_{\mathbb{R}^p} \left(\mu_\mathsf{C}(\mathbf{w}) + t_n \mu_{D_n}(\mathbf{w}) - \mu_\Pi(\mathbf{w})\right)^2 d\mathbf{w} \\
&\qquad\qquad - \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^p \frac{1}{t_n} \int\limits_{\mathbb{R}^p} \left(\mu_\mathsf{C}(\mathbf{w}) - \mu_\Pi(\mathbf{w})\right)^2 d\mathbf{w}
\end{aligned}$$

$$= \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^p \frac{1}{t_n} \int_{\mathbb{R}^p} t_n \mu_{D_n}(\mathbf{w}) \left(2\mu_{\mathsf{C}}(\mathbf{w}) + t_n \mu_{D_n}(\mathbf{w}) - 2\mu_{\Pi}(\mathbf{w})\right) d\mathbf{w}$$

$$= \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^p \left[2\int_{\mathbb{R}^p} \mu_{D_n}(\mathbf{w}) \left(\mu_{\mathsf{C}}(\mathbf{w}) - \mu_{\Pi}(\mathbf{w})\right) d\mathbf{w} + t_n \int_{\mathbb{R}^p} \mu^2_{D_n}(\mathbf{w})\, d\mathbf{w}\right] \quad \text{(E6)}$$

Now, using the fact $\int_{\mathbb{R}^p} \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^d K_{\frac{\sigma}{\sqrt{2}}}(\mathbf{u},\mathbf{w}) K_{\frac{\sigma}{\sqrt{2}}}(\mathbf{v},\mathbf{w})\, d\mathbf{w} = K_\sigma(\mathbf{u},\mathbf{v})$, it is quite straightforward to check that

$$\left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^p \int_{\mathbb{R}^p} \mu_{D_n}(\mathbf{w}) \left(\mu_{\mathsf{C}}(\mathbf{w}) - \mu_{\Pi}(\mathbf{w})\right) d\mathbf{w} = \int_{[0,1]^p}\int_{[0,1]^p} K_\sigma(\mathbf{u},\mathbf{v})\, dD_n(\mathbf{u})\, d(\mathsf{C}-\Pi)(\mathbf{v}).$$

From this identity and Equation (E6), we get

$$\psi'_{\mathsf{C}}(\mathsf{D}) = \lim_{n\to\infty} \frac{\psi(\mathsf{C}+t_n D_n) - \psi(\mathsf{C})}{t_n} = 2\int_{[0,1]^p}\int_{[0,1]^p} (\mathbf{u},\mathbf{v})\, dD(\mathbf{u})\, d(\mathsf{C}-\Pi)(\mathbf{v}).$$

This Lemma then follows easily from Theorem 2.9 and the functional delta method. The only thing that one needs to verify here is that $\psi'_{\mathsf{C}}(\mathbb{G}_{\mathsf{C}})$ is normally distributed with the mean 0 and the variance

$$\delta_0^2 = 4\int_{[0,1]^p}\int_{[0,1]^p} g(\mathbf{u})g(\mathbf{v})\, \mathrm{E}[\, d\mathbb{G}_{\mathsf{C}}(\mathbf{u})\, d\mathbb{G}_{\mathsf{C}}(\mathbf{v})] \quad \text{where} \quad g(\mathbf{u}) = \int_{[0,1]^p} K_\sigma(\mathbf{u},\mathbf{w})\, d(\mathsf{C}-\Pi)(\mathbf{w}).$$

But this is straightforward from the formula for the derivative ψ'_{C}.

When $\mathsf{C}=\Pi$: Clearly the map $\mathsf{D} \to \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^p \int_{\mathbb{R}^p} \left(\int_{[0,1]^p} K_{\frac{\sigma}{\sqrt{2}}}(\mathbf{u},\mathbf{v})\, d\mathsf{D}(\mathbf{u})\right)^2 d\mathbf{v}$ from $\mathcal{D}([0,1]^p)$ to $\mathbb{R}$ is continuous. So, the fact that $\sqrt{n}(\mathsf{C}_n - \Pi) \to \mathbb{G}_\Pi$ and the continuous mapping theorem gives

$$n\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi) \xrightarrow{\mathcal{D}} \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^p \int_{\mathbb{R}^p} \left(\int_{[0,1]^p} K_{\frac{\sigma}{\sqrt{2}}}(\mathbf{u},\mathbf{v})\, d\mathbb{G}_\Pi(\mathbf{u})\right)^2 d\mathbf{v}$$

$$= \int_{[0,1]^p}\int_{[0,1]^p} K_\sigma(\mathbf{u},\mathbf{v})\, d\mathbb{G}_\Pi(\mathbf{u})\, d\mathbb{G}_\Pi(\mathbf{v}). \qquad \square$$

Proof of Theorem 2.5

When $\mathsf{C}\neq\Pi$: Write $\sqrt{n}(\widehat{I}^2_{\sigma,n}(\mathbf{X}) - I^2_\sigma(\mathbf{X})) = \sqrt{n}\left(\frac{\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi_n)}{\gamma^2_{K_\sigma}(\mathsf{M}_n,\Pi_n)} - \frac{\gamma^2_{K_\sigma}(\mathsf{C},\Pi)}{\gamma^2_{K_\sigma}(\mathsf{M},\Pi)}\right)$ as $A_{1,n} + A_{2,n} + A_{3,n}$, where

$$A_{1,n} = \sqrt{n}\left(\frac{\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi_n)}{\gamma^2_{K_\sigma}(\mathsf{M}_n,\Pi_n)} - \frac{\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi_n)}{\gamma^2_{K_\sigma}(\mathsf{M},\Pi)}\right), \qquad A_{2,n} = \sqrt{n}\left(\frac{\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi_n)}{\gamma^2_{K_\sigma}(\mathsf{M},\Pi)} - \frac{\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi)}{\gamma^2_{K_\sigma}(\mathsf{M},\Pi)}\right)$$

and $A_{3,n} = \sqrt{n}\left(\frac{\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi)}{\gamma^2_{K_\sigma}(\mathsf{M},\Pi)} - \frac{\gamma^2_{K_\sigma}(\mathsf{C},\Pi)}{\gamma^2_{K_\sigma}(\mathsf{M},\Pi)}\right)$.

Clearly $A_{2,n} \to 0$ almost surely by Lemma 2.2. The same is true of $A_{1,n}$ as well, once again by Lemma 2.2 because it is bounded above by

$$\underbrace{\frac{\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)}{\gamma^2_{K_\sigma}(\mathsf{M}_n, \Pi_n)\gamma^2_{K_\sigma}(\mathsf{M}, \Pi)}}_{\text{bounded sequence}} \left(\underbrace{\sqrt{n}\left|\gamma^2_{K_\sigma}(\mathsf{M}_n, \Pi_n) - \gamma^2_{K_\sigma}(\mathsf{M}, \Pi_n)\right|}_{\text{goes to 0}} + \underbrace{\sqrt{n}\left|\gamma^2_{K_\sigma}(\mathsf{M}, \Pi_n) - \gamma^2_{K_\sigma}(\mathsf{M}, \Pi)\right|}_{\text{goes to 0}} \right).$$

Therefore, using Lemma 2.5, we can conclude that $\sqrt{n}(\widehat{I}^2_{\sigma,n}(\mathbf{X}) - I^2_\sigma(\mathbf{X})) \xrightarrow{\mathcal{D}} N(0, C^{-2}_{\sigma,p}.\delta^2_0)$. Now, applying the delta method, one gets $\sqrt{n}(\widehat{I}_{\sigma,n}(\mathbf{X}) - I_\sigma(\mathbf{X})) \xrightarrow{\mathcal{D}} N(0, \delta^2)$, where $\delta^2 = \gamma^{-2}_{K_\sigma}(\mathsf{C}, \Pi) \int\limits_{[0,1]^p} \int\limits_{[0,1]^p} g(\mathbf{u})g(\mathbf{v})\, \mathrm{E}[\, d\mathbb{G}_\mathsf{C}(\mathbf{u})\, d\mathbb{G}_\mathsf{C}(\mathbf{v})]$.

When $\mathsf{C} = \Pi$: As a consequence of the Lemma 2.5 and Lemma 2.2, under null hypothesis and assumptions of Theorem 2.9, we have

$$\begin{aligned} n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n) &= n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi) + \left(n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n) - n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi)\right) \\ &\xrightarrow{\mathcal{D}} \int\limits_{[0,1]^p} \int\limits_{[0,1]^p} K_\sigma(\mathbf{u}, \mathbf{v})\, d\mathbb{G}_\Pi(\mathbf{u})\, d\mathbb{G}_\Pi(\mathbf{v}). \end{aligned}$$

It is enough to show that $\int\limits_{[0,1]^p} \int\limits_{[0,1]^p} K_\sigma(\mathbf{u}, \mathbf{v})\, d\mathbb{G}_\Pi(\mathbf{u})\, d\mathbb{G}_\Pi(\mathbf{v}) \overset{\mathcal{D}}{=} \sum\limits_{i=1}^{\infty} \alpha_i Z_i^2$, for some $\alpha_i > 0$ and $Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$. Then the actual result will follow by putting $\lambda_i = \alpha_i C^{-1}_{\sigma,p}$.

To this end, we define $X(\mathbf{w}) := \int\limits_{[0,1]^p} K_{\frac{\sigma}{\sqrt{2}}}(\mathbf{u}, \mathbf{w})\, d\mathbb{G}_\Pi(\mathbf{u}),\ \forall \mathbf{w} \in \mathbb{R}^p$. So, $\{X(\mathbf{w}) : \mathbf{w} \in \mathbb{R}^p\}$ is then a zero-mean continuous path Gaussian process. This implies that

$$\int\limits_{\mathbb{R}^p} \left(\int\limits_{[0,1]^p} K_{\frac{\sigma}{\sqrt{2}}}(\mathbf{u}, \mathbf{w})\, d\mathbb{G}_\Pi(\mathbf{u}) \right)^2 d\mathbf{w} = \int\limits_{\mathbb{R}^p} (X(\mathbf{w}))^2\, d\mathbf{w} \overset{\mathcal{D}}{=} \sum_{i=1}^{\infty} \beta_i Z_i^2,$$

where the β_i's are the eigenvalues of the covariance operator associated with the Gaussian process (see, e.g. Ferreira and Menegatto, 2012; Serfling, 1980). Now, using the fact that $\left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^p \int\limits_{\mathbb{R}^p} K_{\frac{\sigma}{\sqrt{2}}}(\mathbf{u}, \mathbf{w}) K_{\frac{\sigma}{\sqrt{2}}}(\mathbf{v}, \mathbf{w})\, d\mathbf{w} = K_\sigma(\mathbf{u}, \mathbf{v})$, one can easily see that the last equality yields the desired result

$$\int\limits_{[0,1]^p} \int\limits_{[0,1]^p} K_\sigma(\mathbf{u}, \mathbf{v})\, d\mathbb{G}_\Pi(\mathbf{u})\, d\mathbb{G}_\Pi(\mathbf{v}) \overset{\mathcal{D}}{=} \sum_{i=1}^{\infty} \alpha_i Z_i^2, \quad \text{with } \alpha_i = \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^p \beta_i. \qquad \square$$

Proof of Theorem 2.6 Triangle inequality and $|a-b|^2 \le |a^2 - b^2|$ for $a, b \ge 0$ give

$$|\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n) - \gamma_{K_\sigma}(\mathsf{M}, \Pi)| \le \left|\gamma^2_{K_\sigma}(\mathsf{M}_n, \Pi_n) - \gamma^2_{K_\sigma}(\mathsf{M}_n, \Pi)\right|^{\frac{1}{2}} + \left|\gamma^2_{K_\sigma}(\mathsf{M}_n, \Pi) - \gamma^2_{K_\sigma}(\mathsf{M}, \Pi)\right|^{\frac{1}{2}}.$$

Using Lemma 2.2, we get $\lim_{n\to\infty} |\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n) - \gamma_{K_\sigma}(\mathsf{M}, \Pi)| = 0$ a.s.. Using again the same inequalities and the fact that γ_{K_σ} is a metric, one gets

$$|\gamma_{K_\sigma}(\mathsf{C}_n, \Pi_n) - \gamma_{K_\sigma}(\mathsf{C}, \Pi)| \leq \left|\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n) - \gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi)\right|^{\frac{1}{2}} + \gamma_{K_\sigma}(\mathsf{C}_n, \mathsf{C}_{\mathbf{z},n}) + \gamma_{K_\sigma}(\mathsf{C}_{\mathbf{z},n}, \mathsf{C}).$$

Again, using Lemmas 2.2, 2.3 and 2.4, we get $|\gamma_{K_\sigma}(\mathsf{C}_n, \Pi_n) - \gamma_{K_\sigma}(\mathsf{C}, \Pi)| \overset{\text{a.s.}}{\to} 0$ as $n \to 0$, and as a consequence, we conclude that as $n \to \infty$,

$$\widehat{I}_{\sigma,n}(\mathbf{X}) = \frac{\gamma_{K_\sigma}(\mathsf{C}_n, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n)} \to \frac{\gamma_{K_\sigma}(\mathsf{C}, \Pi)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)} = I_\sigma(\mathbf{X}) \text{ almost surely.} \qquad \square$$

Lemma 2.6. *Let P_n and Q_n be sequence of probability distribution over $[0,1]^p$. Let σ_n be a sequence of positive real numbers that converges to $\sigma_0 > 0$. Then as $n \to \infty$, $|\gamma^2_{K_{\sigma_n}}(P_n, Q_n) - \gamma^2_{K_{\sigma_0}}(P_n, Q_n)| \to 0$.*

Proof. First we observe that

$$\begin{aligned}|\gamma^2_{K_{\sigma_n}}(P_n, Q_n) - \gamma^2_{K_{\sigma_0}}(P_n, Q_n)| \leq\ & \mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim P_n\otimes P_n}|K_{\sigma_n}(\mathbf{S}, \mathbf{S}_*) - K_{\sigma_0}(\mathbf{S}, \mathbf{S}_*)| \\ & + 2\mathrm{E}_{(\mathbf{S},\mathbf{T})\sim P_n\otimes Q_n}|K_{\sigma_n}(\mathbf{S}, \mathbf{T}) - K_{\sigma_0}(\mathbf{S}, \mathbf{T})| \\ & + \mathrm{E}_{(\mathbf{T},\mathbf{T}_*)\sim Q_n\otimes Q_n}|K_{\sigma_n}(\mathbf{S}, \mathbf{T}_*) - K_{\sigma_0}(\mathbf{T}, \mathbf{T}_*)|.\end{aligned}$$

Applying Lemma 6 of Póczos *et al.* (2012), one can get an upper bound of $|K_{\sigma_n}(\mathbf{S}, \mathbf{T}) - K_{\sigma_0}(\mathbf{S}, \mathbf{T})|$ in the following way

$$|K_{\sigma_n}(\mathbf{S}, \mathbf{T}) - K_{\sigma_0}(\mathbf{S}, \mathbf{T})| \leq L\left\|\frac{\mathbf{S}}{\sigma_n} - \frac{\mathbf{S}}{\sigma_0}\right\| + L\left\|\frac{\mathbf{T}}{\sigma_n} - \frac{\mathbf{T}}{\sigma_0}\right\| \leq 2L\sqrt{p}\left|\frac{1}{\sigma_n} - \frac{1}{\sigma_0}\right|,$$

where L is a constant. Thus we can conclude that $|\gamma^2_{K_{\sigma_n}}(P_n, Q_n) - \gamma^2_{K_{\sigma_0}}(P_n, Q_n)| \leq 8L\sqrt{p}\left|\frac{1}{\sigma_n} - \frac{1}{\sigma_0}\right|$. This completes the proof. $\square$

Proof of Theorem 2.7. Note that

$$|\gamma^2_{\sigma_n}(\mathsf{C}_n, \Pi_n) - \gamma^2_{\sigma_0}(\mathsf{C}, \Pi)| \leq |\gamma^2_{\sigma_n}(\mathsf{C}_n, \Pi_n) - \gamma^2_{\sigma_0}(\mathsf{C}_n, \Pi_n)| + |\gamma^2_{\sigma_0}(\mathsf{C}_n, \Pi_n) - \gamma^2_{\sigma_0}(\mathsf{C}, \Pi)|.$$

As $n \to \infty$, the first term on the right hand side goes to 0 due to Lemma 2.6 and the second term converges to 0 almost surely due to Theorem 2.6. Similarly, one can show that $|\gamma^2_{\sigma_n}(\mathsf{M}_n, \Pi_n) - \gamma^2_{\sigma_0}(\mathsf{M}, \Pi)| \to 0$ as $n \to \infty$. This implies $\widehat{I}_{\sigma_n,n}(\mathbf{X}) \to I_{\sigma_0}(\mathbf{X})$ almost surely. Because of the fact that $I_{\sigma_0}(\mathbf{X}) = 0$ if and only if the coordinates of $\mathbf{X}$ are independent, test of independence based on the statistic $\widehat{I}_{\sigma_n,n}(\mathbf{X})$ is consistent. $\square$

Proof of Theorem 2.8. From Theorem 2.7, it follows that $\widehat{I}_{\sigma^{(i)},n}(\mathbf{X})$'s are consistent test statistics for all $i = 1, 2, \ldots, m$. Since m is finite, by the virtue of the definition of $T_{\max,n}$ and $T_{\text{sum},n}$, they converge to 0 almost surely if and only if the coordinates of $\mathbf{X}$ are independent. Otherwise, they converge to positive quantities. This property makes the resulting tests consistent. Again, under the alternative hypothesis, for any i, the p-value p_i corresponding to the test statistic $\widehat{I}_{\sigma^{(i)},n}(\mathbf{X})$ converges to zero almost surely. So, for sufficiently large n, almost surely, there would exist at least one i such that p_i is less than α/m, which makes the set $\{i : p_{(i)} < \alpha/m\}$ non-empty. Thus the power of the test based on FDR tends to 1 as sample size tends to infinity. □

Chapter 3

Test of Independence among Random Variables with Arbitrary Probability Distributions

In order to develop our copula based tests in Chapter 2, we assumed all underlying variables $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$ to be continuous so that ties occur with zero probability, and the ranks can be uniquely defined. Other rank based dependency measures and associated tests like those based on generalizations of Spearman's ρ, Kendall's τ, Blomqvist's β and Hoeffding's Φ statistics (see, e.g., Joe, 1990; Nelsen, 1996; Úbeda-Flores, 2005; Gaißer *et al.*, 2010) also need this continuity assumption for their implementations. However, in practice, we often have data sets consisting of a mixture of continuous, discrete, ordinal and binary variables. To deal with such data sets, recently Genest *et al.* (2019) proposed a generalization of the Hoeffding's Φ-statistic based on checkerboard copula and developed a test of independence based on it. This measure and the resulting test can be used for random variables having arbitrary probability distributions. But neither this measure nor the test is invariant under strictly monotone transformations of the variables unless the same type of transformations (either strictly increasing or strictly decreasing) are used for all variables. Moreover, they are not very useful for detecting complex non-monotone relationships among the variables (see Section 3.2). In order to take care of these problems, in this chapter, we propose a new measure of dependence, which can be viewed as a checkerboard copula version of the measure proposed in Chapter 2. This measure and the resulting tests are invariant under permutations and strictly monotone transformations of the variables. Description of these tests is given in the following section.

3.1 The proposed measure and associated tests

Assume that $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(p)}) \sim F$ and $X^{(i)} \sim F_i$, for $i = 1, 2, \ldots, p$. Sklar's theorem (see, e.g., Nelsen, 2007) guarantees that there exists at least one distribution function C on $[0,1]^p$ with uniform marginals such that

$$F(u^{(1)}, u^{(2)}, \ldots, u^{(p)}) = \mathsf{C}(F_1(u^{(1)}), F_2(u^{(2)}), \ldots, F_p(u^{(p)})), \ \forall\ u^{(1)}, u^{(2)}, \ldots, u^{(p)} \in \mathbb{R}. \quad (3.1)$$

When the $X^{(i)}$'s are continuous, then C is unique, and C can be shown to be the joint distribution of $F_1(X^{(1)}), F_2(X^{(2)}), \ldots, F_p(X^{(p)})$. This C is known as the copula distribution of F. So, under the continuity assumption, the $X^{(i)}$'s are independent if and only if $\mathsf{C} = \Pi$, the uniform distribution function on $[0,1]^p$. In Chapter 2, we considered a discrepancy measure γ_{K_σ}, called MMD, to measure the difference between C and Π. Recall that MMD between two probability distributions P and Q on $\mathbb{R}^p$ is given by

$$\gamma_{K_\sigma}(P, Q) = [\mathrm{E}K(\mathbf{Y}, \mathbf{Y}_*) - 2\mathrm{E}K(\mathbf{Y}, \mathbf{Z}) + \mathrm{E}K(\mathbf{Z}, \mathbf{Z}_*)]^{\frac{1}{2}}, \quad (3.2)$$

where $\mathbf{Y}, \mathbf{Y}_* \overset{i.i.d.}{\sim} P$, $\mathbf{Z}, \mathbf{Z}_* \overset{i.i.d.}{\sim} Q$ are four independent random vectors, and $K_\sigma(\mathbf{x}, \mathbf{y}) = \exp\left(-\|\mathbf{x} - \mathbf{y}\|^2 / 2\sigma^2\right)$ is the Gaussian kernel. Since γ_{K_σ} is a metric (see, e.g., Sriperumbudur *et al.*, 2010, for more details), we have $\gamma_{K_\sigma}(\mathsf{C}, \Pi) \geq 0$, where the equality holds if and only if $\mathsf{C} = \Pi$. Therefore, $\gamma_{K_\sigma}(\mathsf{C}, \Pi)$ serves as a measure of dependence among the $X^{(i)}$'s.

But when the $X^{(i)}$'s are not continuous, the copula distribution C in Equation (3.1) is unique only on $\mathrm{range}(F_1) \times \mathrm{range}(F_2) \times \cdots \times \mathrm{range}(F_p)$. However, for each joint distribution function F, there exists a copula $\mathsf{C}^{\maltese}$, called 'checkerboard copula', which satisfies Equation (3.1) and $\mathsf{C}^{\maltese} = \Pi$ if and only if the $X^{(i)}$'s are independent (see Genest *et al.*, 2017). The definition of checkerboard copula is given below.

Definition 3.1. *Let $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(p)})$ be a p-dimensional random vector having the distribution function F with uniariate marginals $F_1, F_2, \ldots, F_p$, respectively. Let $\mathbf{\Lambda} = (\Lambda^{(1)}, \Lambda^{(2)}, \ldots, \Lambda^{(p)})$ be a random vector which is independent of $\mathbf{X}$ and follows uniform distribution over $[0,1]^p$. The checkerboard copula $\mathsf{C}^{\maltese}$ of F is defined as the distribution function of the random vector $\mathbf{U} = (U^{(1)}, U^{(2)}, \ldots, U^{(p)})$, where*

$$U^{(i)} = \Lambda^{(i)} F_i(X^{(i)}) + (1 - \Lambda^{(i)}) F_i(X^{(i)-}) \quad \textit{for each} \quad i = 1, 2, \ldots, p.$$

Since γ_{K_σ} is a metric, we have $\gamma_{K_\sigma}(\mathsf{C}^{\maltese}, \Pi) \geq 0$, where $\gamma_{K_\sigma}(\mathsf{C}^{\maltese}, \Pi) = 0$ holds if and only if the $X^{(i)}$'s are independent. Like Chapter 2, here we use a scaled version of this measure

$$I_\sigma^{\Pi}(\mathbf{X}) = \gamma_{K_\sigma}(\mathsf{C}^{\Pi}, \Pi)/\gamma_{K_\sigma}(\mathsf{M}, \Pi),$$

where M is the maximum copula (defined in Section 2.1). So, $I_\sigma^{\Pi}(\mathbf{X}) = 0$ if and only if $X^{(i)}$'s are mutually independent, and positive otherwise. From the definition, it immediately follows that $I_\sigma^{\Pi}(\mathbf{X})$ coincides with $I_\sigma(\mathbf{X})$ whenever all the $X^{(i)}$'s are continuous (see, Chapter 2). The scaling constant $\gamma_{K_\sigma}(\mathsf{M}, \Pi) = \sqrt{C_{\sigma,p}}$ (see Theorem 2.1 for the definition of $C_{\sigma,p}$) can be calculated easily. Note that $I_\sigma^{\Pi}(\mathbf{X})$ is invariant under permutation of the variables $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$. The translation invariant property of the Gaussian kernel also makes it invariant under strictly monotone transformations of the variables. These results are stated below as Theorem 3.1.

Theorem 3.1. *$I_\sigma^{\Pi}(\mathbf{X})$ is invariant under permutations and strictly monotone transformations of random variables $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$.*

Like $I_\sigma(\mathbf{X})$, the dependency measure $I_\sigma^{\Pi}(\mathbf{X})$ also enjoys some nice theoretical properties. For instance, it can also be viewed as a weighted squared distance between the characteristic functions of C^{Π} and Π . In the case of $p = 2$, it can be expressed as a correlation co-efficient between two random quantities. Here we skip the proofs of these results since they are similar to those of Theorem 2.1 and 2.2. For getting a data driven estimate of $I_\sigma^{\Pi}(\mathbf{X})$, one needs to get an empirical version of the checkerboard copula. The construction of this empirical copula is given below.

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be n independent observations of the random vector $\mathbf{X}$. For any fixed $j = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$, we define $r_i^{(j)} = \sum_{k=1}^{n} \mathbb{I}\left[x_k^{(j)} \leq x_i^{(j)}\right]$ to get $\mathbf{r}_i = (r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(p)})$, the coordinate-wise rank of $\mathbf{x}_i$. For $i = 1, 2, \ldots, n$, we define the normalized rank vectors $\mathbf{y}_1 = \mathbf{r}_1/n, \mathbf{y}_2 = \mathbf{r}_2/n, \ldots, \mathbf{y}_n = \mathbf{r}_n/n$ as before. For $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, we also define $s_i^{(j)}$ as $s_i^{(j)} = \sum_{k=1}^{n} \mathbb{I}\left[x_k^{(j)} < x_i^{(j)}\right]$ to get $\mathbf{s}_i = (s_i^{(1)}, s_i^{(2)}, \ldots, s_i^{(p)})$ and its normalized version $\mathbf{z}_i = \mathbf{s}_i/n$ for $i = 1, 2, \ldots, n$. Now, for $a < b$, define $L_{a,b}$ as the distribution function of an uniform distribution on $[a, b]$. The empirical version of the checkerboard copula can be defined as

$$\mathsf{C}_n^{\Pi}(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} L_{z_i^{(j)}, y_i^{(j)}}(u^{(j)}) \quad \forall\ \mathbf{u} = (u^{(1)}, u^{(2)}, \ldots, u^{(p)}) \in [0, 1]^p.$$

This is equivalent to the definition of empirical checkerboard copula given in Genest *et al.* (2017). An interesting property of this empirical copula C_n^{Π} is that it admits a density c_n^{Π} on $(0, 1)^p$, which is given by

$$c_n^{\Pi}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}\frac{I\left[z_i^{(j)} < u^{(j)} \le y_i^{(j)}\right]}{\left(y_i^{(j)} - z_i^{(j)}\right)} \quad \forall\ \mathbf{u} = (u^{(1)}, u^{(2)}, \ldots, u^{(p)}) \in (0,1)^p. \tag{3.3}$$

Using this density, we get the following closed form expression for $\gamma_{K_\sigma}(\mathsf{C}_n^{\Pi}, \Pi)$.

Lemma 3.1. *Define*

$$V_\sigma(a_1, a_2, b_1, b_2) := \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(-1)^{i+j-1}\sqrt{2\pi}\sigma^2}{(a_2-a_1)(b_2-b_1)}\left[\left(\frac{a_i-b_j}{\sigma}\right)\Phi\left(\frac{a_i-b_j}{\sigma}\right) + \phi\left(\frac{a_i-b_j}{\sigma}\right)\right],$$

where Φ and ϕ are the distribution function and the density function of the standard normal variate, respectively. Then, $\gamma^2_{K_\sigma}(\mathsf{C}_n^{\Pi}, \Pi)$ is given by

$$\gamma^2_{K_\sigma}(\mathsf{C}_n^{\Pi}, \Pi) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\prod_{j=1}^{p} V_\sigma\left(z_i^{(j)}, y_i^{(j)}, z_k^{(j)}, y_k^{(j)}\right) - \frac{2}{n}\sum_{i=1}^{n}\prod_{j=1}^{p} V_\sigma\left(z_i^{(j)}, y_i^{(j)}, 0, 1\right) + [V_\sigma(0,1,0,1)]^p.$$

The above result shows that the computing cost for $\gamma_{K_\sigma}(\mathsf{C}_n^{\Pi}, \Pi)$ is of the order $\mathcal{O}(pn^2)$. Now we define $\widehat{I}^{\Pi}_{\sigma,n}(\mathbf{X}) = \gamma_{K_\sigma}(\mathsf{C}_n^{\Pi}, \Pi)/\gamma_{K_\sigma}(\mathsf{M}, \Pi)$ as an the estimator for $I^{\Pi}_{\sigma}(\mathbf{X})$. Like $I^{\Pi}_{\sigma}(\mathbf{X})$, its estimate $\widehat{I}^{\Pi}_{\sigma,n}(\mathbf{X})$ is also invariant under permutation and strictly monotone transformation. This result is asserted by the following theorem.

Theorem 3.2. *$\widehat{I}^{\Pi}_{\sigma,n}(\mathbf{X})$ is invariant under permutations and strictly monotonic transformations of the variables $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$.*

Genest *et al.* (2017) proved that $\|\mathsf{C}_n^{\Pi} - \mathsf{C}^{\Pi}\|$, the L_2-norm between C_n^{Π} and C^{Π}, converges to 0 in probability. In fact, under quite general conditions, they showed the weak convergence of the empirical checkerboard copula process $\mathbb{C}_n^{\Pi} = \sqrt{n}(\mathsf{C}_n^{\Pi} - \mathsf{C}^{\Pi})$ to a centered Gaussian process. Using their results, we can prove the consistency of our estimator $\widehat{I}^{\Pi}_{\sigma,n}(\mathbf{X})$. This result is stated below.

Theorem 3.3. *$\widehat{I}^{\Pi}_{\sigma,n}(\mathbf{X})$ is a consistent estimator of $I^{\Pi}_{\sigma}(\mathbf{X})$.*

We can use $T_n^{\Pi} = \widehat{I}^{\Pi}_{\sigma,n}(\mathbf{X})$ as a test statistic and reject $\mathbb{H}_0$, the null hypothesis of mutual independence, for large value of T_n^{Π}. The cut-off can be computed using the permutation principle discussed before (see Section 2.3). For any fixed parameter $\sigma > 0$, the large sample consistency of the test based on T_n^{Π} follows from the convergence of $T_n^{\Pi} = \widehat{I}^{\Pi}_{\sigma,n}(\mathbf{X})$ to $I^{\Pi}_{\sigma}(\mathbf{X})$ (see Theorem 3.3) and the fact that $I^{\Pi}_{\sigma}(\mathbf{X}) = 0$ under $\mathbb{H}_0$, while it is positive under the alternative hypothesis.

However, the finite sample performance of the test depends on the choice of σ. Like Chapter 2, here also one can choose σ based on median heuristic (see, e.g., Gretton

et al., 2008). But our empirical experience in the previous chapter suggests that the use of smaller bandwidths sometimes leads to good performance, especially when the variables have complex non-monotone relationships. So, one can also go for the multi-scale methods discussed before. In such cases, we consider the results for m different bandwidths $\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(m)}$ to come up with the test statistic $T^{\Cross}_{\text{sum},n} := \sum_{i=1}^{m} \widehat{I}^{\Cross}_{n,\sigma^{(i)}}(\mathbf{X})$ or $T^{\Cross}_{\text{max},n} := \max_{1\le i\le m} \widehat{I}^{\Cross}_{n,\sigma^{(i)}}(\mathbf{X})$ and reject $\mathbb{H}_0$ when the observed value of the test statistic is larger than the corresponding cut-off determined by the permutation method. The multi-scale method based on FDR can be used as well. For all three aggregation methods, the resulting tests turn out to be consistent. This consistency result is stated below.

Theorem 3.4. *Powers of the tests based on $T^{\Cross}_{\text{sum},n}$, $T^{\Cross}_{\text{max},n}$ and FDR converge to 1 as the sample size tends to infinity.*

3.2 Results from analysis of simulated and real data sets

We analyzed several simulated and real data sets to compare the performance of our proposed tests with some existing methods that can be used for testing independence among several random variables having arbitrary distributions. In particular, we considered the test proposed by Genest *et al.* (2019), the dHSIC test (see Pfister *et al.*, 2018) and the JdCov test (see Chakraborty and Zhang, 2019) for comparison. We also report the results of the HHG test (see Heller *et al.*, 2013) for bivariate data sets. Since the ranks are not uniquely defined here, the rank-JdCov test could not be used.

For our multi-scale tests, we started with the bandwidth based on median heuristic and choose other bandwidths following the method discussed in Chapter 2. The median of the distribution of $\|\mathbf{Z} - \mathbf{Z}_*\|^2$, where $\mathbf{Z}, \mathbf{Z}_* \overset{i.i.d.}{\sim} \Pi$, was used for median heuristic. For all other competing tests, we used the same set up as in Chapter 2. Throughout this chapter, cut-offs of all tests are computed based on 1000 random permutations. In each of the simulated examples, empirical powers of different tests are calculated based on 10000 simulations as before.

3.2.1 Analysis of simulated data sets

We began with eight examples involving bivariate data. Figure 3.1 provides a visual representation of these data sets. The first six examples (labeled as 'Four Clouds', 'W', 'Diamond', 'Parabola', 'Two Parabolas', and 'Circle') are discretized versions of the examples

considered by Newton (2009) and used in Section 2.5. Recall that in all these examples, $X^{(1)}$ and $X^{(2)}$ are uncorrelated, but barring the first example, they are not independent. After generating $(X^{(1)}, X^{(2)})$ using Newton's algorithm, we discretized them using the transformation $(X^{(1)}, X^{(2)}) \mapsto (\lfloor 5X^{(1)} \rfloor, \lfloor 5X^{(2)} \rfloor)$, where $\lfloor t \rfloor$ denotes the largest integer not exceeding t. In the 'Hyperplane' example, we have $X^{(1)} = \lfloor 5U \rfloor$ and $X^{(2)} = \lfloor 5U + 5V \rfloor$, where $U, V \overset{i.i.d.}{\sim} U(0,1)$. In the 'Correlated Normal' example, we have $X^{(1)} = \lfloor 5U \rfloor, X^{(2)} = \lfloor 5V \rfloor$, where (U, V) follows a bivariate normal distribution with correlation coefficient 0.4. So, in these two examples, $X^{(1)}$ and $X^{(2)}$ have positive correlations, which is evident from 3.1. For our proposed tests, we created an R packages 'GCGK' containing all necessary codes. This package is available at `https://github.com/angshumanroycode/GCGK`.

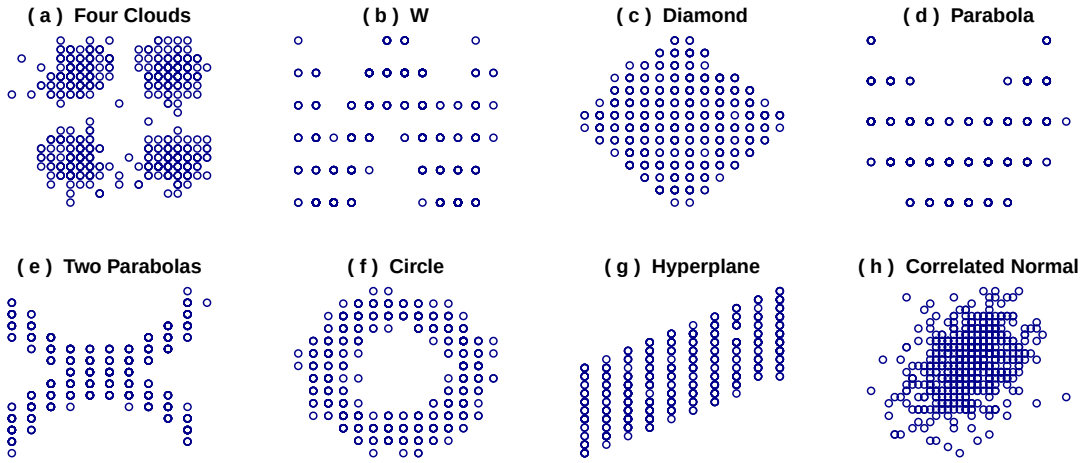


FIGURE 3.1: Scatter plots of eight bivariate data sets from distributions with discrete marginals.

Figure 3.2 shows the empirical powers of different tests on these eight data sets. In the first example, where $X^{(1)}$ and $X^{(2)}$ are independent, almost all tests had empirical powers close to the nominal level of 0.05 (See Figure 3.2(a)). Only the test based on FDR had slightly lower power because of its conservative nature.

In 'W' and 'Circle' examples, the proposed test based on $T^{\maltese}_{\max,n}$ outperformed all other tests considered here (see Figures 3.2(b) and 3.2(f)). The test based on FDR had the second best performance in the 'Circle' example. In the 'W' example also, it performed well, where this test, the dHSIC test and the HHG test had comparable powers. The HHG test had the highest power in 'Parabola', 'Diamond' and 'Two Parabolas' examples (see Figures 3.2(c)- 3.2(e)). The tests based on $T^{\maltese}_{\max,n}$ and FDR, particularly the former one, also performed well. The dHSIC test also had excellent performance in these three examples. Unfortunately, the test based on $T^{\maltese}_{n}$, the JdCov test and the test proposed

by Genest *et al.* (2019) did not have satisfactory performance in these five examples (see Figures 3.2(b)- 3.2(f)). However, in the last two examples, where $X^{(1)}$ and $X^{(2)}$ have positive correlations, they performed well (see Figures 3.2(g) and 3.2(h)). In these two data sets, all testing procedures had reasonable performance, though the powers of HHG and dHSIC tests were slightly lower. Among our multi-scale methods, the one based on $T^{✠}_{\mathrm{sum},n}$ had a slight edge.

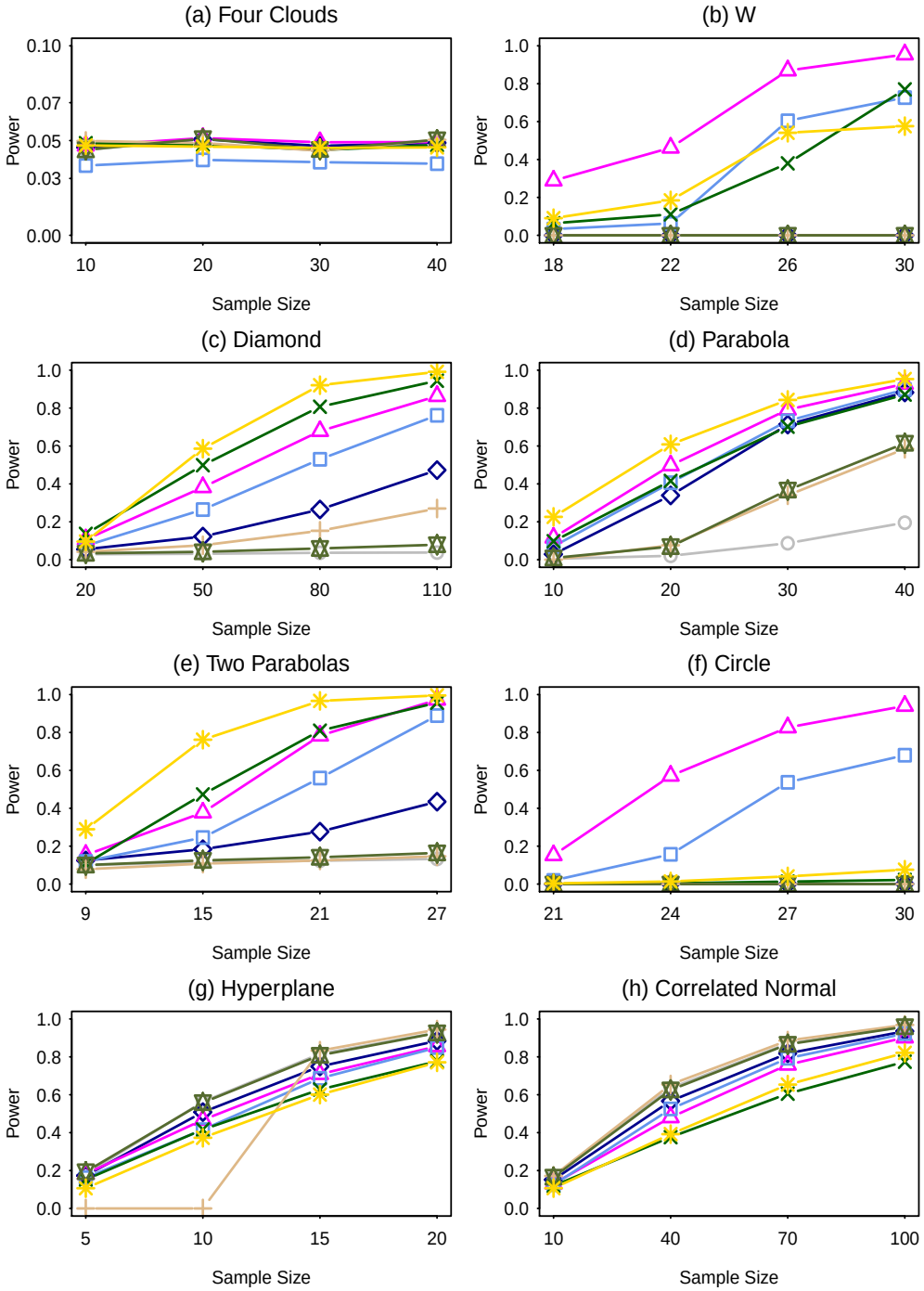


FIGURE 3.2: Powers of $T^{✠}_{n}$ (○), $T^{✠}_{\mathrm{sum},n}$ (◇), $T^{✠}_{\mathrm{max},n}$ (△), FDR (□), dHSIC (×), JdCov (+), Genest (⊠) and HHG (✳) tests in data sets generated from discrete bivariate distributions.

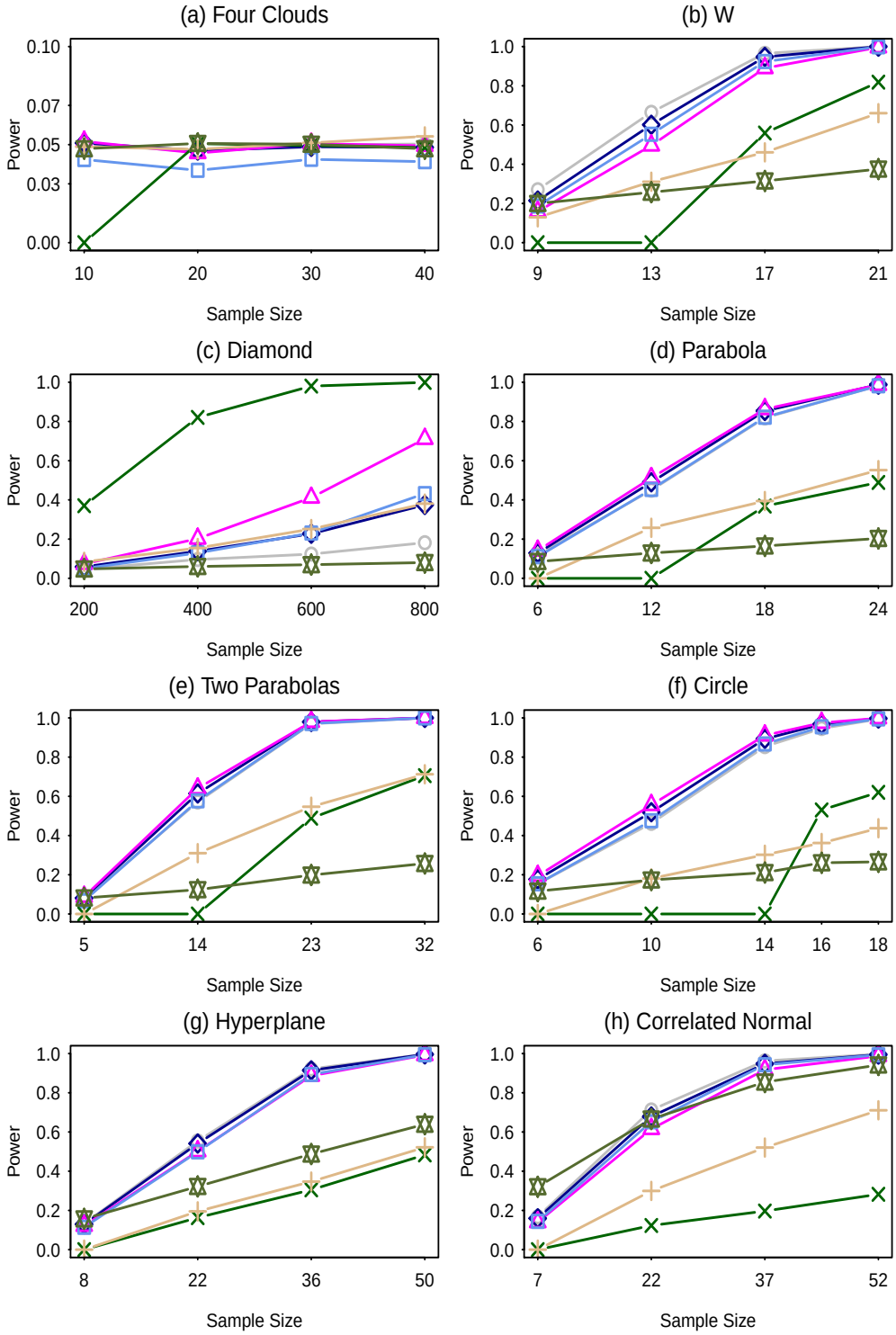


FIGURE 3.3: Powers of $T_n^{\maltese}$ (○), $T_{\mathrm{sum},n}^{\maltese}$ (◇), $T_{\max,n}^{\maltese}$ (△), FDR (□), dHSIC (×), JdCov (+) and Genest (⧖) tests in data sets generated from discrete eight-dimensional distributions.

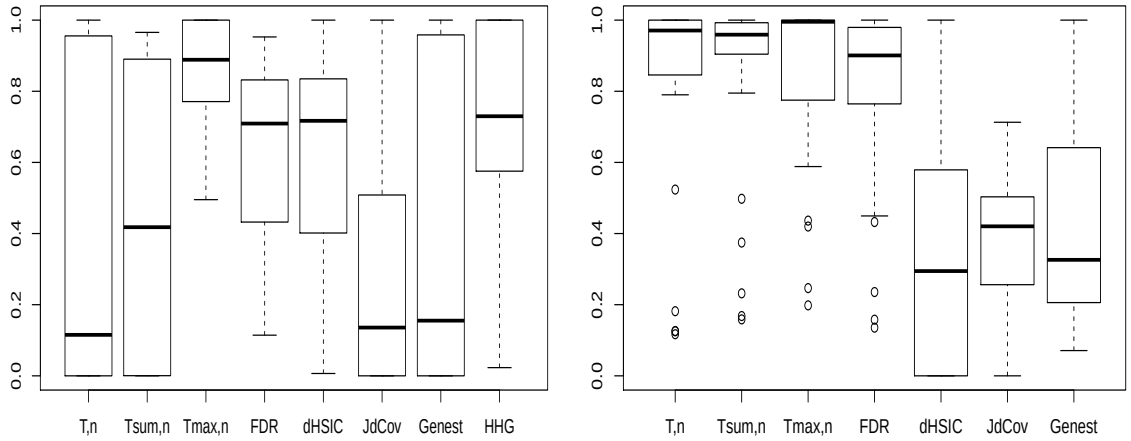
Next, we analyzed some eight-dimensional data sets, which can be viewed as noisy multivariate extensions of the bivariate examples discussed above. For each of the first six examples, we generated two independent observations from the discrete bivariate distribution (see Figure 3.1), and then four independent noise variables were augmented to get a

vector of dimension eight. Each noise variable was distributed as $\lfloor 5U \rfloor$, where $U \sim N(0,1)$. In the 'Hyperplane' example, we generated seven i.i.d. $U(0,1)$ variables $U^{(2)}, U^{(3)}, \ldots, U^{(8)}$ and took $U^{(1)} = (U^{(2)} + U^{(3)} + \cdots + U^{(8)}) + \varepsilon$, where $\varepsilon \sim U(0,1)$. In the 'Correlated Normal' example, $\mathbf{U} = (U^{(1)}, U^{(2)}, \ldots, U^{(8)})$ was generated from a eight-dimensional normal distribution with the mean vector $\mathbf{0}$ and the dispersion matrix $\mathbf{\Sigma} = ((a_{i,j}))$, where $a_{i,j} = 0.4^{|i-j|}$ for all $i, j \in \{1, 2, \ldots, 8\}$. In these two examples, observations on $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(8)})$ were obtained by using the transformation $X^{(i)} = \lfloor 5U^{(i)} \rfloor$ for $i = 1, 2, \ldots, 8$.

Figure 3.3 shows the powers of different tests on these data sets. Recall that the dHSIC needs the sample size to be at least twice the number of variables. So, it could not be used for sample size smaller than 16. As expected, in the case of 'Four Clouds' example, all tests barring the test based on FDR had power closed to the nominal level (See Figure 3.3(a)). In five out of the remaining seven examples, our proposed tests clearly outperformed their competitors. In 'Diamond' data set, the dHSIC test had the highest power, but the power of the test based on $T^{\maltese}_{\max,n}$ was higher than all other tests considered here. In the case of 'Correlated Normal' example, the test proposed by Genest *et al.* (2019) and our proposed tests had similar powers, and their performance was much better than JdCov and dHSIC tests. Note that unlike bivariate examples, the test based on $T^{\maltese}_{n}$ had very good performance in these eight-dimensional data sets. Its performance was comparable to its multi-scale versions based on $T^{\maltese}_{\text{sum},n}, T^{\maltese}_{\max,n}$ and FDR.

To compare the overall performance of different tests in a comprehensive way, we used the boxplots of efficiency scores as discussed in Section 2.5. However, here the comparison was carried out among all competing tests. Recall that for a given data set and a given sample size, we defined the efficiency score of a test as its observed power divided by the maximum power obtained for that experiment. These efficiency scores were computed for different data sets (barring the 'Four Clouds' example, where $\mathbf{X}$ and $\mathbf{Y}$ are independent), and they are presented using boxplots in Figure 3.4. This figure suggests that both for $p = 2$ and $p = 8$, the overall performance of the proposed test based on $T^{\maltese}_{\max,n}$ was much better than all other tests considered here. In the case of bivariate data sets, the HHG test had the second best performance. The dHSIC test and the test based on FDR also worked well. But the performance of all other tests including the one based on $T^{\maltese}_{n}$ was not satisfactory. Except for 'Hyperplane' and 'Correlated Normal' examples, this single-scale method could not perform well. This shows the necessity of the multi-scale approach.

However, in cases of eight dimensional data sets, this single-scale test and its all multi-scale analogs had excellent performance, and they outperformed their competitors. This is consistent with what we observed in Chapter 2.



(A) Discrete bivariate data sets.

(B) Discrete eight-dimensional data sets.

FIGURE 3.4: Boxplots of efficiency score for overall comparison among different tests.

3.2.2 Analysis of real data sets

For further evaluation of our proposed tests, we analyzed three real data sets taken from the UCI machine learning repository https://archive.ics.uci.edu/ml/. Each of these data sets contains some discrete variables with ties. Brief description of these data sets is given below.

BUPA Liver Disorders data were collected by BUPA Medical Research Ltd. This data set has 345 observations on seven variables. For our study, we consider the first five variables (Mean Corpuscular Volume, Alkaline Phosphatase, Alanine Aminotransferase, Aspartate Aminotransferase and Gamma-Glutamyl Transpeptidase) related to different blood tests, which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. The natural question that arises here is whether these variables have dependence among them.

Challenger Space Shuttle data were recorded and assessed by Draper (1995). This data set contains information on shuttle launching temperature, leak-check pressure and number of field joints experiencing thermal stress for the 23 NASA space shuttle flights before the challenger disaster happened in 1986. Here one may be interested to know whether there is any dependence among these three variables.

Haberman's Survival data set was analyzed by Haberman (1976). It contains 306 cases from a study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients, who had undergone surgery for breast cancer. Here we investigated whether there is any relationship among the three variables: age of patient at the time of operation, number of positive axillary nodes detected and whether the patient survived 5 years or longer.

In cases of BUPA Liver Disorder data and Haberman's Survival data, when we used the full data set for testing, all tests rejected the null hypothesis of independence. In the case of Challenger Space Shuttle data, only the test of Genest *et al.* (2019) failed to do so. Based on that single experiment in each data set, it was not possible to compare among different test procedures. So, we carried out our experiments using randomly chosen subsets from the full data set. For each sub-sample size, the experiment was repeated 5000 times to compute the empirical powers of different tests, and they are shown in Figure 3.5.

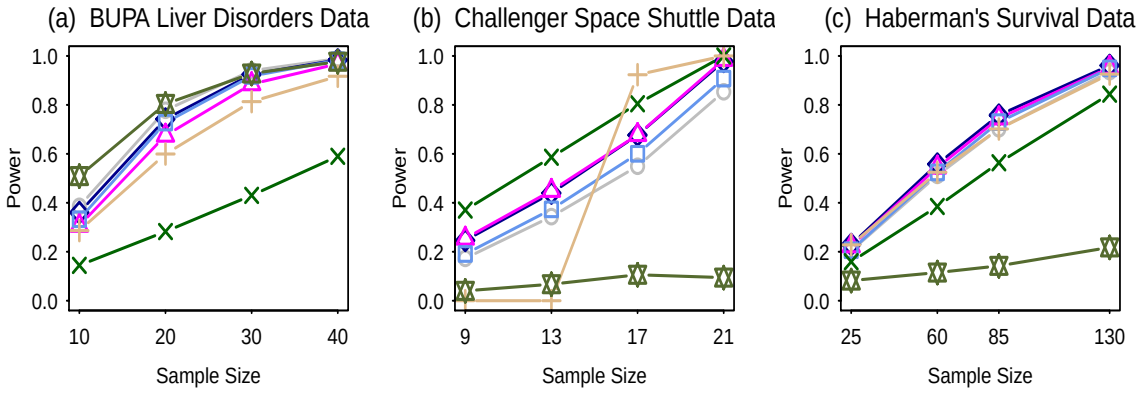


FIGURE 3.5: Powers of $T_n^{\maltese}$ (○), $T_{\text{sum},n}^{\maltese}$ (◇), $T_{\max,n}^{\maltese}$ (△), FDR (□), dHSIC (×), JdCov (+) and Genest (⊠) tests in real data sets.

The test proposed by Genest *et al.* (2019) had the highest power in BUPA Liver Disorders data set, but it performed poorly in other two cases. Interestingly, all our proposed tests had good overall performances in all data sets. In the case of Challenger Space Shuttle Data, the dHSIC test had the best overall performance, but it performed poorly otherwise. In this example, the JdCov test outperformed its competitors when larger samples were used, but unfortunately, it could not be used for smaller sample sizes (the R codes provided by the author returned error message). The tests based on $T_{\text{sum},n}^{\maltese}$, $T_{\max,n}^{\maltese}$ and FDR had excellent performance in Haberman's Survival data set. The test based on $T_n^{\maltese}$ and the JdCov test also had competitive powers.

3.3 Proofs and mathematical details

Proof of Theorem 3.1. For any permutation ξ on $\mathbb{R}^p$, we have $K_\sigma(\xi(\mathbf{x}), \xi(\mathbf{y})) = K_\sigma(\mathbf{x}, \mathbf{y})$ and also, $\mathbf{T} \sim \Pi$ implies $\xi(\mathbf{T}) \sim \Pi$. Using these, one gets

$$\mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim \mathsf{C}^{\Psi}_{\xi(\mathbf{X})}\otimes \mathsf{C}^{\Psi}_{\xi(\mathbf{X})}}[K_\sigma(\mathbf{S},\mathbf{S}_*)]=\mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim \mathsf{C}^{\Psi}_{\mathbf{X}}\otimes \mathsf{C}^{\Psi}_{\mathbf{X}}}[K_\sigma(\xi(\mathbf{S}),\xi(\mathbf{S}_*))]=\mathrm{E}_{(\mathbf{S},\mathbf{S}_*)\sim \mathsf{C}^{\Psi}_{\mathbf{X}}\otimes \mathsf{C}^{\Psi}_{\mathbf{X}}}[K_\sigma(\mathbf{S},\mathbf{S}_*)]$$

and $\mathrm{E}_{(\mathbf{S},\mathbf{T})\sim \mathsf{C}^{\Psi}_{\xi(\mathbf{X})}\otimes \Pi}[K_\sigma(\mathbf{S},\mathbf{T})] = \mathrm{E}_{(\mathbf{S},\mathbf{T})\sim \mathsf{C}^{\Psi}_{\mathbf{X}}\otimes \Pi}[K_\sigma(\xi(\mathbf{S}),\xi(\mathbf{T}))] = \mathrm{E}_{(\mathbf{S},\mathbf{T})\sim \mathsf{C}^{\Psi}_{\mathbf{X}}\otimes \Pi}[K_\sigma(\mathbf{S},\mathbf{T})]$.

From these, it follows that $\gamma_{K_\sigma}(\mathsf{C}^{\Psi}_{\xi(\mathbf{X})}, \Pi) = \gamma_{K_\sigma}(\mathsf{C}^{\Psi}, \Pi)$ and hence $I^{\Psi}_\sigma(\xi(\mathbf{X})) = I^{\Psi}_\sigma(\mathbf{X})$.

Consider any fixed set $A \subseteq \{1, 2, \ldots, p\}$ and a function $\mathbf{f} : \mathbb{R}^p \mapsto \mathbb{R}^p$ such that $\mathbf{f}(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = (f_1(x^{(1)}), f_2(x^{(2)}), \ldots, f_p(x^{(p)}))$, where for each $i \in A$, $f_i : \mathbb{R} \mapsto \mathbb{R}$ is strictly increasing and for each $i \notin A$, $f_i : \mathbb{R} \mapsto \mathbb{R}$ is strictly decreasing. Also define a function $\mathbf{g}(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = (g_1(x^{(1)}), g_2(x^{(2)}), \ldots, g_p(x^{(p)}))$ with $g_i(x) = x \ \forall i \in A$ and $g_i(x) = 1 - x \ \forall i \notin A$. Let C^{Ψ} and $\mathsf{C}^{\Psi}_{\mathbf{f}}$ be the checkerboard copulas corresponding to $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(p)})$ and $\mathbf{f}(\mathbf{X})$, respectively. At first, we will show that if $\mathbf{U} = (U^{(1)}, U^{(2)}, \ldots, U^{(p)})$ follows C^{Ψ}, then $\mathbf{g}(\mathbf{U})$ follows $\mathsf{C}^{\Psi}_{\mathbf{f}}$.

Assume that $\mathbf{V} = (V^{(1)}, V^{(2)}, \ldots, V^{(p)})$ follows $\mathsf{C}^{\Psi}_{\mathbf{f}}$ and let $G_1, G_2, \ldots, G_p$ be the cumulative distribution functions of $f_1(X^{(1)}), f_2(X^{(2)}), \ldots, f_p(X^{(p)})$, respectively. From the definition of checkerboard copula, one can check that $V^{(i)}$ has the same distribution as $\Psi^{(i)}G_i(f_i(X^{(i)})) + (1 - \Psi^{(i)})G_i(f_i(X^{(i)})^-)$, where $\mathbf{\Psi} = (\Psi^{(1)}, \Psi^{(2)}, \ldots, \Psi^{(p)})$ is independent of $\mathbf{f}(\mathbf{X})$ and follows uniform distribution over $[0, 1]^p$ (see, Definition 3.1).

Now, if $i \in A$, f_i is strictly increasing. So, we have

$$\begin{aligned}
&\Psi^{(i)}G_i(f_i(x)) + (1 - \Psi^{(i)})G_i(f_i(x)^-) \\
&= \Psi^{(i)}\mathrm{Pr}\left[f_i(X^{(i)}) \le f_i(x)\right] + (1 - \Psi^{(i)})\mathrm{Pr}\left[f_i(X^{(i)}) < f_i(x)\right] \\
&= \Psi^{(i)}\mathrm{Pr}\left[X^{(i)} \le x\right] + (1 - \Psi^{(i)})\mathrm{Pr}\left[X^{(i)} < x\right] = \Psi^{(i)}F_i(x) + (1 - \Psi^{(i)})F_i(x^-).
\end{aligned}$$

Thus $V^{(i)}$ has the same distribution as $\Psi^{(i)}F_i(X^{(i)}) + (1 - \Psi^{(i)})F_i(X^{(i)-})$, which in turn has the same distribution as $U^{(i)} = g_i(U^{(i)})$ (since $g_i(x) = x$ for $i \in A$).

Again, if $i \notin A$, f_i is strictly decreasing. So, we have

$$\begin{aligned}
&\Psi^{(i)}G_i(f_i(x)) + (1 - \Psi^{(i)})G_i(f_i(x)^-) \\
&= \Psi^{(i)}\mathrm{Pr}\left[f_i(X^{(i)}) \le f_i(x)\right] + (1 - \Psi^{(i)})\mathrm{Pr}\left[f_i(X^{(i)}) < f_i(x)\right] \\
&= \Psi^{(i)}\mathrm{Pr}\left[X^{(i)} \ge x\right] + (1 - \Psi^{(i)})\mathrm{Pr}\left[X^{(i)} > x\right]
\end{aligned}$$

$$= \Psi^{(i)} - \Psi^{(i)}\Pr\left[X^{(i)} < x\right] + (1-\Psi^{(i)}) - (1-\Psi^{(i)})\Pr\left[X^{(i)} \le x\right]$$
$$= 1 - (1-\Psi^{(i)})\Pr\left[X^{(i)} \le x\right] - \Psi^{(i)}\Pr\left[X^{(i)} < x\right] = 1 - (1-\Psi^{(i)})F_i(x) - \Psi^{(i)}F_i(x^-).$$

Thus $V^{(i)}$ has the same distribution as $1-(1-\Psi^{(i)})F_i(X^{(i)}) - \Psi^{(i)}F_i(X^{(i)-})$, which in turn has the same distribution as $1 - U^{(i)} = g_i(U^{(i)})$ (since $g_i(x) = 1-x$ for $i \notin A$).

It shows that if $\mathbf{U} = (U^{(1)}, U^{(2)}, \ldots, U^{(p)})$ follows C^{Ψ}, then $\mathbf{g}(\mathbf{U})$ follows $\mathsf{C}_{\mathbf{f}}^{\Psi}$. Next, we shall show that $\gamma_{K_\sigma}(\mathsf{C}_{\mathbf{f}}^{\Psi}, \Pi) = \gamma_{K_\sigma}(\mathsf{C}^{\Psi}, \Pi)$. Note that since the Gaussian kernel is translation invariant, we have $K_\sigma(\mathbf{g}(\mathbf{x}), \mathbf{g}(\mathbf{y})) = K_\sigma(\mathbf{x}, \mathbf{y})$. Also note that if $\mathbf{Y} \sim \Pi$, then $\mathbf{g}(\mathbf{Y}) \sim \Pi$. The rest of the proof follows from the facts that

$$\mathrm{E}_{(\mathbf{Y},\mathbf{Y}_*)\sim \mathsf{C}_{\mathbf{f}}^{\Psi}\otimes\mathsf{C}_{\mathbf{f}}^{\Psi}} K_\sigma(\mathbf{Y},\mathbf{Y}_*) = \mathrm{E}_{(\mathbf{Y},\mathbf{Y}_*)\sim \mathsf{C}^{\Psi}\otimes\mathsf{C}^{\Psi}} K_\sigma\left(\mathbf{g}(\mathbf{Y}),\mathbf{g}(\mathbf{Y}_*)\right) = \mathrm{E}_{(\mathbf{Y},\mathbf{Y}_*)\sim \mathsf{C}^{\Psi}\otimes\mathsf{C}^{\Psi}} K_\sigma(\mathbf{Y},\mathbf{Y}_*)$$

and $\mathrm{E}_{(\mathbf{Y},\mathbf{Z})\sim \mathsf{C}_{\mathbf{f}}^{\Psi}\otimes\Pi} K_\sigma(\mathbf{Y},\mathbf{Z}) = \mathrm{E}_{(\mathbf{Y},\mathbf{Z})\sim \mathsf{C}^{\Psi}\otimes\Pi} K_\sigma\left(\mathbf{g}(\mathbf{Y}),\mathbf{g}(\mathbf{Z})\right) = \mathrm{E}_{(\mathbf{Y},\mathbf{Z})\sim \mathsf{C}^{\Psi}\otimes\Pi} K_\sigma(\mathbf{Y},\mathbf{Z})$. □

Proof of Lemma 3.1 Note that

$$\int_{a_1}^{a_2}\int_{b_1}^{b_2} e^{-\frac{(x-y)^2}{2\sigma^2}}\,dx\,dy = \sqrt{2\pi}\sigma \int_{a_1}^{a_2}\left[\int_{b_1}^{b_2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-x)^2}{2\sigma^2}}\,dy\right]dx$$
$$= \sqrt{2\pi}\sigma \int_{a_1}^{a_2}\left[\Phi\left(\frac{b_2-x}{\sigma}\right) - \Phi\left(\frac{b_1-x}{\sigma}\right)\right]dx.$$

Using the fact $\int \Phi(x)\,dx = x\Phi(x) + \phi(x) + c$ (where c is an integration constant), we get

$$\int_{a_1}^{a_2}\int_{b_1}^{b_2} e^{-\frac{(x-y)^2}{2\sigma^2}}\,dx\,dy = \sqrt{2\pi}\sigma^2 \sum_{i=1}^{2}\sum_{j=1}^{2}(-1)^{i+j-1}\left[\left(\frac{a_i-b_j}{\sigma}\right)\Phi\left(\frac{a_i-b_j}{\sigma}\right) + \phi\left(\frac{a_i-b_j}{\sigma}\right)\right]$$
$$= (a_2-a_1)(b_2-b_1)V_\sigma(a_1,a_2,b_1,b_2).$$

Now, assume that $\mathbf{S}, \mathbf{S}_*, \mathbf{T}, \mathbf{T}_*$ are four independent random variables such that $\mathbf{S}, \mathbf{S}_* \sim \mathsf{C}_n^{\Psi}$ and $\mathbf{T}, \mathbf{T}_* \sim \Pi$. From definition of γ_{K_σ} in Equation (3.2), we obtain $\gamma_{K_\sigma}^2(\mathsf{C}_n^{\Psi}, \Pi) = S_1 - 2S_2 + S_3$, where $S_1 = \mathrm{E}K_\sigma(\mathbf{S}, \mathbf{S}_*)$, $S_2 = \mathrm{E}K_\sigma(\mathbf{S}, \mathbf{T})$ and $S_3 = \mathrm{E}K_\sigma(\mathbf{T}, \mathbf{T}_*)$. Now we simplify S_1, S_2 and S_3 further using the expression of empirical checkerboard copula density c_n^{Ψ} given in Equation (3.3).

$$S_1 = \mathrm{E}K_\sigma(\mathbf{S}, \mathbf{S}_*) = \int_{[0,1]^p}\int_{[0,1]^p} K_\sigma(\mathbf{u},\mathbf{v}) c_n^{\Psi}(\mathbf{u}) c_n^{\Psi}(\mathbf{v})\,d\mathbf{u}\,d\mathbf{v}$$
$$= \int_{[0,1]^p}\left[\int_{[0,1]^p} K_\sigma(\mathbf{u},\mathbf{v})\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}\frac{\mathbb{I}\left[z_i^{(j)} < u^{(j)} \le y_i^{(j)}\right]}{\left(y_i^{(j)} - z_i^{(j)}\right)}\,d\mathbf{u}\right] c_n^{\Psi}(\mathbf{v})\,d\mathbf{v}$$

$$= \int_{[0,1]^p} \left[\frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} \frac{1}{(y_i^{(j)} - z_i^{(j)})} \int_{z_i^{(j)}}^{y_i^{(j)}} e^{-\frac{(u^{(j)} - v^{(j)})^2}{2\sigma^2}} du^{(j)} \right] \frac{1}{n} \sum_{k=1}^{n} \prod_{j=1}^{p} \frac{\mathbb{I}\left[z_k^{(j)} < v^{(j)} \le y_k^{(j)}\right]}{\left(y_k^{(j)} - z_k^{(j)}\right)} d\mathbf{v}$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} \prod_{j=1}^{p} \frac{1}{(y_i^{(j)} - z_i^{(j)}) \left(y_k^{(j)} - z_k^{(j)}\right)} \int_{z_i^{(j)}}^{y_i^{(j)}} \int_{z_k^{(j)}}^{y_k^{(j)}} e^{-\frac{(u^{(j)} - v^{(j)})^2}{2\sigma^2}} du^{(j)} dv^{(j)}$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} \prod_{j=1}^{p} V_\sigma \left(z_i^{(j)}, y_i^{(j)}, z_k^{(j)}, y_k^{(j)}\right).$$

$$S_2 = \mathrm{E} K_\sigma(\mathbf{S}, \mathbf{T}) = \int_{[0,1]^p} \int_{[0,1]^p} K_\sigma(\mathbf{u}, \mathbf{v}) c_n^{\maltese}(\mathbf{u}) \, d\mathbf{u} \, d\mathbf{v}$$

$$= \int_{[0,1]^p} \left[\int_{[0,1]^p} K_\sigma(\mathbf{u}, \mathbf{v}) \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} \frac{\mathbb{I}\left[z_i^{(j)} < u^{(j)} \le y_i^{(j)}\right]}{\left(y_i^{(j)} - z_i^{(j)}\right)} d\mathbf{u} \right] d\mathbf{v}$$

$$= \int_{[0,1]^p} \left[\frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} \frac{1}{(y_i^{(j)} - z_i^{(j)})} \int_{z_i^{(j)}}^{y_i^{(j)}} e^{-\frac{(u^{(j)} - v^{(j)})^2}{2\sigma^2}} du^{(j)} \right] d\mathbf{v}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} \frac{1}{(y_i^{(j)} - z_i^{(j)})} \int_{z_i^{(j)}}^{y_i^{(j)}} \int_{0}^{1} e^{-\frac{(u^{(j)} - v^{(j)})^2}{2\sigma^2}} du^{(j)} dv^{(j)} = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} V_\sigma \left(z_i^{(j)}, y_i^{(j)}, 0, 1\right).$$

$$S_3 = \mathrm{E} K_\sigma(\mathbf{T}, \mathbf{T}_*) = \int_{[0,1]^p} \int_{[0,1]^p} K_\sigma(\mathbf{u}, \mathbf{v}) \, d\mathbf{u} \, d\mathbf{v} = \int_{[0,1]^p} \left[\prod_{j=1}^{p} \int_{0}^{1} e^{-\frac{(u^{(j)} - v^{(j)})^2}{2\sigma^2}} du^{(j)} \right] d\mathbf{v}$$

$$= \prod_{j=1}^{p} \left[\int_{0}^{1} \int_{0}^{1} e^{-\frac{(u^{(j)} - v^{(j)})^2}{2\sigma^2}} du^{(j)} dv^{(j)} \right] = [V_\sigma(0,1,0,1)]^p. \qquad \square$$

Proof of Theorem 3.2. Permutation invariance of $\widehat{I}_{\sigma,n}^{\maltese}(\mathbf{X})$ follows immediately from the form of $\gamma_{K_\sigma}^2(\mathsf{C}_n^{\maltese}, \Pi)$ given in Lemma 3.1.

Note that if the j-th variable (for $j = 1, 2, \ldots, p$) undergoes a strictly increasing transformation, the values of $y_i^{(j)}$ and $z_i^{(j)}$ remain unchanged for all $i = 1, 2, \ldots, n$. If it undergoes a strictly decreasing transformation, then modified values of $y_i^{(j)}$ and $z_i^{(j)}$ turn out to be $y_i^{*(j)} = 1 - z_i^{(j)}$ and $z_i^{*(j)} = 1 - y_i^{(j)}$, respectively. Now, the proof follows from the expression of $\gamma_{K_\sigma}^2(\mathsf{C}_n^{\maltese}, \Pi)$ given in Lemma 3.1, and the observation that for $a, b, c, d \in (0, 1)$, $V_\sigma(1-b, 1-a, 1-d, 1-c) = V_\sigma(a, b, c, d)$ and $V_\sigma(1-b, 1-a, 0, 1) = V_\sigma(a, b, 0, 1)$. $\square$

Lemma 3.2. *Let A and B be two non-empty subsets of $\{1, 2, \ldots, p\}$. Now given $\mathbf{u} = (u^{(1)}, u^{(2)}, \ldots, u^{(p)}) \in [0,1]^p$ and $\mathbf{v} = (v^{(1)}, v^{(2)}, \ldots, v^{(p)}) \in [0,1]^p$, we define two vectors $\mathbf{u}_A = (u_A^{(1)}, u_A^{(2)}, \ldots, u_A^{(p)}) \in [0,1]^p$ and $\mathbf{v}_B = (v_B^{(1)}, v_B^{(2)}, \ldots, v_B^{(p)}) \in [0,1]^p$ as*

$$u_A^{(j)} = \begin{cases} u^{(j)} \text{ if } j \in A \\ 1 \text{ if } j \notin A \end{cases} \quad \text{and} \quad v_B^{(j)} = \begin{cases} v^{(j)} \text{ if } j \in B \\ 1 \text{ if } j \notin B. \end{cases}$$

Also define $K_{A,B} : [0,1]^p \times [0,1]^p \mapsto \mathbb{R}$ *such that* $K_{A,B}(\mathbf{u}, \mathbf{v}) = \prod_{j=1}^p K_{A,B}^{(j)}(u^{(j)}, v^{(j)})$, *where*

$$K_{A,B}^{(j)}(u,v) = \begin{cases} 1 & \text{if } j \notin A \cup B \\ \frac{(u-1)}{\sigma^2} e^{-\frac{1}{2\sigma^2}(u-1)^2} & \text{if } j \in A \setminus B \\ \frac{(v-1)}{\sigma^2} e^{-\frac{1}{2\sigma^2}(v-1)^2} & \text{if } j \in B \setminus A \\ \left\{ \frac{1}{\sigma^2} - \frac{(u-v)^2}{\sigma^4} \right\} e^{-\frac{1}{2\sigma^2}(u-v)^2} & \text{if } j \in A \cap B. \end{cases}$$

If we use the notation $d\mathbf{u}_A := \prod_{j \in A} du^{(j)}$ *and* $d\mathbf{v}_B := \prod_{j \in B} dv^{(j)}$, *then*

$$\int_{[0,1]^p} \int_{[0,1]^p} e^{-\frac{\|\mathbf{u}-\mathbf{v}\|^2}{2\sigma^2}} \, d\mathbb{C}_n^{\Pi}(\mathbf{u}) \, d\mathbb{C}_n^{\Pi}(\mathbf{v}) = \sum_A \sum_B \int_{[0,1]^{|A|}} \int_{[0,1]^{|B|}} K_{A,B}(\mathbf{u}_A, \mathbf{v}_B) \mathbb{C}_n^{\Pi}(\mathbf{u}_A) \mathbb{C}_n^{\Pi}(\mathbf{v}_B) \, d\mathbf{u}_A \, d\mathbf{v}_B.$$

Proof. Here we shall prove it for $p = 2$. The proof for general dimension can be obtained similarly by repeated applications of Fubini's theorem and integration by parts formula.

As C^{Π} and C_n^{Π} both are copula distributions, we have the following results:

$$\begin{aligned} \mathbb{C}_n^{\Pi}(0, u^{(2)}) &= \sqrt{n} \left(\mathsf{C}_n^{\Pi}(0, u^{(2)}) - \mathsf{C}^{\Pi}(0, u^{(2)}) \right) = \sqrt{n}(0-0) = 0 \\ \mathbb{C}_n^{\Pi}(u^{(1)}, 0) &= \sqrt{n} \left(\mathsf{C}_n^{\Pi}(u^{(1)}, 0) - \mathsf{C}^{\Pi}(u^{(1)}, 0) \right) = \sqrt{n}(0-0) = 0 \\ \mathbb{C}_n^{\Pi}(1, u^{(2)}) &= \sqrt{n} \left(\mathsf{C}_n^{\Pi}(1, u^{(2)}) - \mathsf{C}^{\Pi}(1, u^{(2)}) \right) = \sqrt{n}(u^{(2)} - u^{(2)}) = 0 \\ \mathbb{C}_n^{\Pi}(u^{(1)}, 1) &= \sqrt{n} \left(\mathsf{C}_n^{\Pi}(u^{(1)}, 1) - \mathsf{C}^{\Pi}(u^{(1)}, 1) \right) = \sqrt{n}(u^{(1)} - u^{(1)}) = 0. \end{aligned}$$

Since $\mathbb{C}_n^{\Pi}$ is a signed measure, we can always write the differential $d\mathbb{C}_n^{\Pi}(u^{(1)}, u^{(2)}) = d\mathbb{C}_n^{\Pi}(u^{(1)}|u^{(2)}) \, d\mathbb{C}_n^{\Pi}(u^{(2)})$, where $\mathbb{C}_n^{\Pi}(u^{(1)}|u^{(2)})$ is the conditional signed measure and $\mathbb{C}_n^{\Pi}(u^{(2)})$ is the marginal signed measure. So, for $v^{(1)}, v^{(2)} \in [0,1]$, using Fubini's theorem, we get

$$\begin{aligned} &\int_0^1 \int_0^1 e^{-\frac{1}{2\sigma^2}\{(u^{(1)}-v^{(1)})^2 + (u^{(2)}-v^{(2)})^2\}} \, d\mathbb{C}_n^{\Pi}(u^{(1)}, u^{(2)}) \\ &= \int_0^1 \left\{ \int_0^1 e^{-\frac{1}{2\sigma^2}(u^{(1)}-v^{(1)})^2} \, d\mathbb{C}_n^{\Pi}(u^{(1)}|u^{(2)}) \right\} e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2} \, d\mathbb{C}_n^{\Pi}(u^{(2)}). \end{aligned} \tag{3.4}$$

For the integral inside braces in Equation (3.4), we now use the integral by parts formula.

$$\begin{aligned} &\int_0^1 e^{-\frac{1}{2\sigma^2}(u^{(1)}-v^{(1)})^2} \, d\mathbb{C}_n^{\Pi}(u^{(1)}|u^{(2)}) \\ &= \left[e^{-\frac{1}{2\sigma^2}(u^{(1)}-v^{(1)})^2} \mathbb{C}_n^{\Pi}(u^{(1)}|u^{(2)}) \right]_0^1 + \int_0^1 \frac{(u^{(1)} - v^{(1)})}{\sigma^2} e^{-\frac{1}{2\sigma^2}(u^{(1)}-v^{(1)})^2} \mathbb{C}_n^{\Pi}(u^{(1)}|u^{(2)}) \, du^{(1)} \end{aligned}$$

$$= e^{-\frac{1}{2\sigma^2}(1-v^{(1)})^2}\mathbb{C}_n^{\Psi}(1|u^{(2)}) - e^{-\frac{1}{2\sigma^2}\left(v^{(1)}\right)^2}\mathbb{C}_n^{\Psi}(0|u^{(2)}) + \int_0^1 \frac{(u^{(1)}-v^{(1)})}{\sigma^2} e^{-\frac{1}{2\sigma^2}(u^{(1)}-v^{(1)})^2}\mathbb{C}_n^{\Psi}(u^{(1)}|u^{(2)})\,du^{(1)}. \tag{3.5}$$

So, from (3.4) and (3.5), we get $\int_0^1\int_0^1 e^{-\frac{1}{2\sigma^2}\{(u^{(1)}-v^{(1)})^2+(u^{(2)}-v^{(2)})^2\}}\,d\mathbb{C}_n^{\Psi}(u^{(1)},u^{(2)})$

$$\begin{aligned}
&= \int_0^1 e^{-\frac{1}{2\sigma^2}(1-v^{(1)})^2}\mathbb{C}_n^{\Psi}(1|u^{(2)})e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\,d\mathbb{C}_n^{\Psi}(u^{(2)}) \\
&\quad - \int_0^1 e^{-\frac{1}{2\sigma^2}\left(v^{(1)}\right)^2}\mathbb{C}_n^{\Psi}(0|u^{(2)})e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\,d\mathbb{C}_n^{\Psi}(u^{(2)}) \\
&\quad + \int_0^1\int_0^1 \frac{(u^{(1)}-v^{(1)})}{\sigma^2} e^{-\frac{1}{2\sigma^2}(u^{(1)}-v^{(1)})^2}\mathbb{C}_n^{\Psi}(u^{(1)}|u^{(2)})\,du^{(1)}e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\,d\mathbb{C}_n^{\Psi}(u^{(2)}) \\
&= e^{-\frac{1}{2\sigma^2}(1-v^{(1)})^2}\underbrace{\int_0^1 e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(1|u^{(2)})\,d\mathbb{C}_n^{\Psi}(u^{(2)})}_{I_1} \\
&\quad - e^{-\frac{1}{2\sigma^2}\left(v^{(1)}\right)^2}\underbrace{\int_0^1 e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(0|u^{(2)})\,d\mathbb{C}_n^{\Psi}(u^{(2)})}_{I_2} \\
&\quad + \int_0^1 \frac{(u^{(1)}-v^{(1)})}{\sigma^2} e^{-\frac{1}{2\sigma^2}(u^{(1)}-v^{(1)})^2}\underbrace{\left\{\int_0^1 e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(u^{(1)}|u^{(2)})\,d\mathbb{C}_n^{\Psi}(u^{(2)})\right\}}_{I_3}\,du^{(1)}
\end{aligned} \tag{3.6}$$

Note that

$$\begin{aligned}
I_1 &= \int_0^1 e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(1|u^{(2)})\,d\mathbb{C}_n^{\Psi}(u^{(2)}) \\
&= \left[e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(1,u^{(2)})\right]_0^1 + \int_0^1 \frac{(u^{(2)}-v^{(2)})}{\sigma^2} e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(1,u^{(2)})\,du_2 = 0.
\end{aligned}$$

Similarly, one can show that $I_2 = 0$, and

$$\begin{aligned}
I_3 &= \int_0^1 e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(u^{(1)}|u^{(2)})\,d\mathbb{C}_n^{\Psi}(u^{(2)}) \\
&= \left[e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(u^{(1)},u^{(2)})\right]_0^1 + \int_0^1 \frac{(u^{(2)}-v^{(2)})}{\sigma^2} e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(u^{(1)},u^{(2)})\,du_2 \\
&= \int_0^1 \frac{(u^{(2)}-v^{(2)})}{\sigma^2} e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(u^{(1)},u^{(2)})\,du^{(2)}.
\end{aligned}$$

Plugging the values of I_1, I_2 and I_3 in Equation (3.6), we get

$$\begin{aligned}
&\int_0^1\int_0^1 e^{-\frac{1}{2\sigma^2}\{(u^{(1)}-v^{(1)})^2+(u^{(2)}-v^{(2)})^2\}}\, d\mathbb{C}_n^{\Psi}(u^{(1)},u^{(2)}) \\
&= \int_0^1\int_0^1 \frac{(u^{(1)}-v^{(1)})}{\sigma^2}\frac{(u^{(2)}-v^{(2)})}{\sigma^2} e^{-\frac{1}{2\sigma^2}\{(u^{(1)}-v^{(1)})^2+(u^{(2)}-v^{(2)})^2\}}\mathbb{C}_n^{\Psi}(u^{(1)},u^{(2)})\, du^{(1)}\, du^{(2)}.
\end{aligned} \tag{3.7}$$

Now, using similar tricks, one can show that for any $u^{(1)}, u^{(2)} \in [0,1]$,

$$\begin{aligned}
&\int_0^1\int_0^1 \frac{(u^{(1)}-v^{(1)})}{\sigma^2}\frac{(u^{(2)}-v^{(2)})}{\sigma^2} e^{-\frac{1}{2\sigma^2}\{(u^{(1)}-v^{(1)})^2+(u^{(2)}-v^{(2)})^2\}}\, d\mathbb{C}_n^{\Psi}(v^{(1)},v^{(2)}) \\
&= \int_0^1\int_0^1 \left\{\frac{1}{\sigma^2}-\frac{(u^{(1)}-v^{(1)})^2}{\sigma^4}\right\} e^{-\frac{1}{2\sigma^2}(u^{(1)}-v^{(1)})^2} \\
&\qquad \times \left\{\frac{1}{\sigma^2}-\frac{(u^{(2)}-v^{(2)})^2}{\sigma^4}\right\} e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(v^{(1)},v^{(2)})\, dv^{(1)}\, dv^{(2)}.
\end{aligned} \tag{3.8}$$

Now, using (3.7) and (3.8), we finally get

$$\begin{aligned}
&\int_0^1\int_0^1\int_0^1\int_0^1 e^{-\frac{1}{2\sigma^2}\{(u^{(1)}-v^{(1)})^2+(u^{(2)}-v^{(2)})^2\}}\, d\mathbb{C}_n^{\Psi}(u^{(1)},u^{(2)})\, d\mathbb{C}_n^{\Psi}(v^{(1)},v^{(2)}) \\
&= \int_0^1\int_0^1\int_0^1\int_0^1 \left\{\frac{1}{\sigma^2}-\frac{(u^{(1)}-v^{(1)})^2}{\sigma^4}\right\} e^{-\frac{1}{2\sigma^2}(u^{(1)}-v^{(1)})^2} \times \left\{\frac{1}{\sigma^2}-\frac{(u^{(2)}-v^{(2)})^2}{\sigma^4}\right\} \\
&\qquad e^{-\frac{1}{2\sigma^2}(u^{(2)}-v^{(2)})^2}\mathbb{C}_n^{\Psi}(u^{(1)},u^{(2)})\mathbb{C}_n^{\Psi}(v^{(1)},v^{(2)})\, du^{(1)}\, du^{(2)}\, dv^{(1)}\, dv^{(2)}.
\end{aligned}$$

□

Lemma 3.3. *The sequence* $\{n\gamma_{K_\sigma}^2(\mathsf{C}_n^{\Psi}, \mathsf{C}^{\Psi})\}_{n\geq 1}$ *is a tight sequence.*

Proof. Form Lemma 3.2, using the same set of notations, we get that

$$\begin{aligned}
n\gamma_{K_\sigma}^2(\mathsf{C}_n^{\Psi}, \mathsf{C}^{\Psi}) &= \int_{[0,1]^p}\int_{[0,1]^p} e^{-\frac{\|\mathbf{u}-\mathbf{v}\|^2}{2\sigma^2}}\, d\mathbb{C}_n^{\Psi}(\mathbf{u})\, d\mathbb{C}_n^{\Psi}(\mathbf{v}) \\
&= \sum_A\sum_B \int_{[0,1]^{|A|}}\int_{[0,1]^{|B|}} K_{A,B}(\mathbf{u}_A,\mathbf{v}_B)\mathbb{C}_n^{\Psi}(\mathbf{u}_A)\mathbb{C}_n^{\Psi}(\mathbf{v}_B)\, d\mathbf{u}_A\, d\mathbf{v}_B.
\end{aligned} \tag{3.9}$$

It can be shown that $K_{A,B} : [0,1]^p \times [0,1]^p \mapsto \mathbb{R}$ is bounded i.e., there exists $M > 0$ such that $\forall\ \mathbf{u}, \mathbf{v} \in [0,1]^p$, $|K_{A,B}(\mathbf{u},\mathbf{v})| \leq M$. Thus for any two given subsets A and B, we have

$$\left| \int_{[0,1]^{|A|}} \int_{[0,1]^{|B|}} K_{A,B}(\mathbf{u}_A, \mathbf{v}_B) \mathbb{C}_n^{\Pi}(\mathbf{u}_A) \mathbb{C}_n^{\Pi}(\mathbf{v}_B)\, d\mathbf{u}_A\, d\mathbf{v}_B \right|$$
$$\leq \int_{[0,1]^{|A|}} \int_{[0,1]^{|B|}} |K_{A,B}(\mathbf{u}_A, \mathbf{v}_B)| \left| \mathbb{C}_n^{\Pi}(\mathbf{u}_A) \right| \left| \mathbb{C}_n^{\Pi}(\mathbf{v}_B) \right| d\mathbf{u}_A\, d\mathbf{v}_B$$
$$\leq M \int_{[0,1]^{|A|}} \left| \mathbb{C}_n^{\Pi}(\mathbf{u}_A) \right| d\mathbf{u}_A \int_{[0,1]^{|B|}} \left| \mathbb{C}_n^{\Pi}(\mathbf{v}_B) \right| d\mathbf{v}_B \quad \leq M \left\| \mathbb{C}_n^{\Pi}(\mathbf{u}_A) \right\| \left\| \mathbb{C}_n^{\Pi}(\mathbf{v}_B) \right\|.$$

Since $\|\mathbb{C}_n^{\Pi}\|$ is tight (see, e.g., Genest *et al.*, 2017), from above equation, we get the tightness of $\int_{[0,1]^{|A|}} \int_{[0,1]^{|B|}} K_{S,T}(\mathbf{u}_A, \mathbf{v}_B) \mathbb{C}_n^{\Pi}(\mathbf{u}_A) \mathbb{C}_n^{\Pi}(\mathbf{v}_B)\, d\mathbf{u}_A\, d\mathbf{v}_B$, which in turn implies the tightness of the sequence $\{n\gamma_{K_\sigma}^2(\mathsf{C}_n^{\Pi}, \mathsf{C}^{\Pi})\}_{n\geq 1}$ (follows from Equation (3.9)). □

Proof of Theorem 3.3. Lemma 3.3 guarantees that the sequence $\{n\gamma_{K_\sigma}^2(\mathsf{C}_n^{\Pi}, \mathsf{C}^{\Pi})\}_{n\geq 1}$ is tight. So, for a given $\varepsilon > 0$, there exists a $M_\varepsilon > 0$ such that $\Pr\left[n\gamma_{K_\sigma}^2(\mathsf{C}_n^{\Pi}, \mathsf{C}^{\Pi}) > M_\varepsilon\right] < \varepsilon$ for all $n \geq 1$. Thus for any given $\varepsilon > 0$ and $\delta > 0$, if we choose N such that $N\delta > M_\varepsilon$, then for all $n > N$, $\Pr\left[\gamma_{K_\sigma}^2(\mathsf{C}_n^{\Pi}, \mathsf{C}^{\Pi}) > \delta\right] < \varepsilon$. Hence $\gamma_{K_\sigma}^2(\mathsf{C}_n^{\Pi}, \mathsf{C}^{\Pi}) \xrightarrow{\Pr} 0$ and therefore $\gamma_{K_\sigma}(\mathsf{C}_n^{\Pi}, \mathsf{C}^{\Pi}) \xrightarrow{\Pr} 0$.

Note that since γ_{K_σ} is a metric, we have $\left|\gamma_{K_\sigma}(\mathsf{C}_n^{\Pi}, \Pi) - \gamma_{K_\sigma}(\mathsf{C}^{\Pi}, \Pi)\right| \leq \gamma_{K_\sigma}(\mathsf{C}_n^{\Pi}, \mathsf{C}^{\Pi})$. Now, since $\gamma_{K_\sigma}(\mathsf{C}_n^{\Pi}, \mathsf{C}^{\Pi}) \xrightarrow{\Pr} 0$, we can conclude that $\gamma_{K_\sigma}(\mathsf{C}_n^{\Pi}, \Pi) \xrightarrow{\Pr} \gamma_{K_\sigma}(\mathsf{C}^{\Pi}, \Pi)$, which further implies that $\widehat{I}_{\sigma,n}^{\Pi}(\mathbf{X})$ converges to $I_\sigma^{\Pi}(\mathbf{X})$ in probability. □

Proof of Theorem 3.4. For any given σ, while $T_n^{\Pi} \xrightarrow{\Pr} 0$ under $\mathbb{H}_0$, under the alternative hypothesis, it converges to a positive constant. So, under the alternative, the p-value associated with the test based on T_n^{Π} converges to 0 as n tends to infinity.

Since m is finite, it is easy to check that under $\mathbb{H}_0$, both $T_{\text{sum},n}^{\Pi}$ and $T_{\text{max},n}^{\Pi}$ converge to 0 in probability, while they converge to some positive constants under the alternative hypothesis. This proves the consistency of the tests based on $T_{\text{sum},n}^{\Pi}$ and $T_{\text{max},n}^{\Pi}$.

Now, it is clear from the above discussion that $p_{(m)}$ and hence all other $p_{(i)}$'s (for $i < m$) converge to 0 in probability. As a result, the set $\{i : p_{(i)} < i\alpha/m\}$ becomes non-empty with probability tending to one. This implies the consistency of the test based on FDR. □

Chapter 4

Test of Independence among Randoms Vectors: Methods Based on One-dimensional Projections

In Chapters 2 and 3, we proposed some copula based methods for testing mutual independence among several random variables. We have also seen that there are several other methods available for this purpose (see, e.g., Nelsen, 1996; Úbeda-Flores, 2005; Gaißer *et al.*, 2010; Póczos *et al.*, 2012; Genest *et al.*, 2019). But, these tests dealing with univariate random variables do not have straightforward generalizations for random vectors. In this chapter, we propose some common recipes for their multivariate generalizations so that the resulting tests can used for testing independence among several random vectors of arbitrary dimensions. Our strategy is very simple; we use some suitable transformations to transform the observations on sub-vectors $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ into univariate observations and then use the existing univariate tests on those transformed observations. In this chapter, we adopt this strategy for multivariate generalizations of the the copula based tests proposed in Chapter 2. But from the description our proposed methods (given in the following sections), it will be clear that these methods can also be used for multivariate generalizations of other univariate tests mentioned above.

4.1 Method based on pairwise distances

This method is motivated by the result that p random vectors $\mathbf{X}^{(1)} \in \mathbb{R}^{d_1}, \mathbf{X}^{(2)} \in \mathbb{R}^{d_2}, \ldots, \mathbf{X}^{(p)} \in \mathbb{R}^{d_p}$ are mutually independent if and only if

$$\Pr\left[\|\mathbf{X}^{(i)} - \mathbf{a}^{(i)}\| < r_i,\ \forall\ 1 \le i \le p\right] = \prod_{i=1}^{p} \Pr\left[\|\mathbf{X}^{(i)} - \mathbf{a}^{(i)}\| < r_i\right] \tag{4.1}$$

for every $\mathbf{a}^{(1)} \in \mathbb{R}^{d_1}, \mathbf{a}^{(2)}, \in \mathbb{R}^{d_2} \ldots, \mathbf{a}^{(p)} \in \mathbb{R}^{d_p}$ and non-negative real numbers $r_1, r_2, \ldots, r_p$. This result follows from the fact that the collection of sets $\Big\{B(\mathbf{a}^{(1)}, r_1) \times B(\mathbf{a}^{(2)}, r_2) \times \cdots \times B(\mathbf{a}^{(p)}, r_p) : \mathbf{a}^{(1)} \in \mathbb{R}^{d_1}, \mathbf{a}^{(2)} \in \mathbb{R}^{d_2}, \ldots, \mathbf{a}^{(p)} \in \mathbb{R}^{d_p},\ r_1, r_2, \ldots, r_p \ge 0\Big\}$ generates the Borel σ-field on $\mathbb{R}^d$, where $B(\mathbf{c}, r) = \big\{\mathbf{u} : \|\mathbf{u} - \mathbf{c}\| < r\big\}$ is the open ball of radius $r > 0$ with center $\mathbf{c}$ (here we use the same notation $B(\cdot, \cdot)$ irrespective of the dimension of the ball). Thus testing for independence among $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ is equivalent to testing for independence among random variables $X^{(\mathbf{a},1)} = \|\mathbf{X}^{(1)} - \mathbf{a}^{(1)}\|, X^{(\mathbf{a},2)} = \|\mathbf{X}^{(2)} - \mathbf{a}^{(2)}\|, \ldots, X^{(\mathbf{a},p)} = \|\mathbf{X}^{(p)} - \mathbf{a}^{(p)}\|$ for every $\mathbf{a} = (\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(p)}) \in \mathbb{R}^d$. Now consider a functional $\mathcal{T}$, which measures dependence among several random variables and satisfies the following property mentioned as Assumption 4.1.

Assumption 4.1. *The dependency measure $\mathcal{T}$ is a functional such that for any p-dimensional ($p \ge 2$) random vector $\mathbf{Z} = (Z^{(1)}, Z^{(2)}, \ldots, Z^{(p)}) \sim G$, $\mathcal{T}(G)$ is non-negative and it takes the value 0 if and only if $Z^{(1)}, Z^{(2)}, \ldots, Z^{(p)}$ are mutually independent.*

So, if $F^{\mathbf{a}}$ denotes the joint distribution of $X^{(\mathbf{a},1)}, X^{(\mathbf{a},2)}, \ldots, X^{(\mathbf{a},p)}$, we have $\mathcal{T}(F^{\mathbf{a}}) \ge 0$, where the equality holds if and only if $X^{(\mathbf{a},1)}, X^{(\mathbf{a},2)}, \ldots, X^{(\mathbf{a},p)}$ are mutually independent. Therefore, from our above discussion, it follows that $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are mutually independent if and only if $\mathcal{T}(F^{\mathbf{a}}) = 0$ for every $\mathbf{a} \in \mathbb{R}^d$. Now, consider a probability measure P on $\mathbb{R}^d$ and define a functional

$$\zeta_{\mathcal{T}}^{P}(F) = \int_{\mathbb{R}^d} \mathcal{T}(F^{\mathbf{a}})\ dP(\mathbf{a}). \tag{4.2}$$

Clearly, $\zeta_{\mathcal{T}}^{P}(F)$ is non-negative, and it takes the value 0 if and only if $\mathcal{T}(F^{\mathbf{a}}) = 0$ for P-almost every $\mathbf{a}$. If P satisfies a certain property, then this in turn implies mutual independence among $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$. This is shown by the following theorem.

Theorem 4.1. *Let $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}) \sim F$ be a d-dimensional random vector, and $\mathcal{T}$ be a measure of dependence among p random variables, which satisfies Assumption 4.1. If P is not singular with respect to the Lebesgue measure, then $\zeta_{\mathcal{T}}^{P}(F) = 0$ if and only if $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are mutually independent.*

Thus, $\zeta_{\mathcal{T}}^{P}(F)$ can be viewed as a multivariate analog of $\mathcal{T}$, and it measures dependence among several random vectors. However, $\zeta_{\mathcal{T}}^{P}(F)$ involves unknown distributions $F^{\mathbf{a}}$ for

$\mathbf{a} \in \mathbb{R}^d$. So, one needs to estimate it from the data. We address this issue in the following subsection.

4.1.1 Estimation of $\zeta_{\mathcal{T}}^{P}(F)$

From Equation 4.2, it is clear that for the estimation of $\zeta_{\mathcal{T}}^{P}(F)$, one needs to estimate $\mathcal{T}(F^{\mathbf{a}})$ for different choices of $\mathbf{a} \in \mathbb{R}^d$. One can use $\mathcal{T}_n(F^{\mathbf{a}})$, a consistent estimator based on n independent observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ on $\mathbf{X}$, for this purpose. Note that $\zeta_{\mathcal{T}}^{P}(F)$ is the expectation of $\mathcal{T}(F^{\mathbf{a}})$ with respect to the probability measure P (see Equation 4.2). So, one can generate N independent observations $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N$ from P and estimate $\zeta_{\mathcal{T}}^{P}(F)$ by the sample average $\zeta_{\mathcal{T}_n}^{P_N}(F) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{T}_n(F^{\mathbf{a}_i})$. Here P_N stands for the empirical version of P with N mass points $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N$. If $\mathcal{T}_n(F^{\mathbf{a}})$ is a consistent estimator of $\mathcal{T}(F^{\mathbf{a}})$ for any fixed $\mathbf{a}$, and $\mathcal{T}(F^{\mathbf{a}})$ is uniformly bounded over $\mathbf{a}$, then under some suitable conditions, $\zeta_{\mathcal{T}_n}^{P_N}(F)$ converges to $\zeta_{\mathcal{T}}^{P}(F)$ either in probability or almost surely. This result is given by the following theorem.

Theorem 4.2. *Assume that $\mathcal{T}$ is a bounded functional and $\{\mathcal{T}_n : n \geq 1\}$ is a sequence of consistent estimators of $\mathcal{T}$. For any fixed $\delta > 0$, define the probability $\widetilde{p}_n(\delta, F) := \sup_{\mathbf{a}\in\mathbb{R}^d} \Pr\left[|\mathcal{T}_n(F^{\mathbf{a}}) - \mathcal{T}(F^{\mathbf{a}})| > \delta\right]$. If $N = N(n)$ is an increasing function of n such that $N(n) \to \infty$ and $N(n)\widetilde{p}_n(\delta, F) \to 0$ as $n \to \infty$, then $\zeta_{\mathcal{T}_n}^{P_N}(F)$ converges to $\zeta_{\mathcal{T}}^{P}(F)$ in probability. Further, if $\sum_{n=1}^{\infty} N(n)\widetilde{p}_n(\delta, F) < \infty$, $\zeta_{\mathcal{T}_n}^{P_N}(F)$ converges to $\zeta_{\mathcal{T}}^{P}(F)$ almost surely.*

If we use the copula based dependency measure $I_\sigma(\mathbf{X})$ and its estimator $\widehat{I}_{\sigma,n}(\mathbf{X})$ (discussed in Chapter 2) as $\mathcal{T}$ and $\mathcal{T}_n$, respectively, irrespective of the choice of F, the condition $\sum_{n=1}^{\infty} N(n)p_n(\delta, F) < \infty$ holds when N is a polynomial function of n (see Lemma 4.1 in Section 4.8). We use similar choices of $\mathcal{T}$ and $\mathcal{T}_n$ for the construction of our test statistic. The details are given below.

4.1.2 Construction of the test statistic

Consider a p dimensional random vector $\mathbf{Z} = (Z^{(1)}, Z^{(2)}, \ldots, Z^{(p)})$ following a distribution G with continuous univariate marginals. To measure dependence among the coordinate variables $Z^{(1)}, Z^{(2)}, \ldots, Z^{(p)}$, in Chapter 2, we used a copula based dependency measure $I_\sigma(\mathbf{Z})$ and its estimator $\widehat{I}_{\sigma,n}(\mathbf{Z})$. Since $I_\sigma(\mathbf{Z})$ is essentially a functional, here we denote it by $\mathbb{T}_\sigma(G)$, and its estimator $\widehat{I}_{\sigma,n}(\mathbf{Z})$ is denoted by $\mathbb{T}_{\sigma,n}(G)$. Also recall that in Chapter 2, we constructed a test statistics $T_n = \widehat{I}_{\sigma_n,n}(\mathbf{Z})$, where σ_n was chosen using "median

heuristic". We have also seen that σ_n is a non-random function of n, and as n tends to infinity, it converges to a constant σ_0. From the proof of Theorem 2.7, we also get that $\mathbb{T}_{\sigma_n,n}(G) = \widehat{I}_{\sigma_n,n}(\mathbf{Z})$ converges to $\mathbb{T}_{\sigma_0}(G) = I_{\sigma_0}(\mathbf{Z})$ almost surely. In this chapter, we shall use $\mathbb{T}_{\sigma_0}$ as $\mathcal{T}$, and $\mathbb{T}_{\sigma_n,n}$ as $\mathcal{T}_n$.

Recall that for the approximation of $\zeta_{\mathcal{T}}^{P}(F)$, we need to generate N i.i.d. observations from P, where N increases with n at an appropriate rate (e.g., polynomial rate of any order). However, we already have n i.i.d. observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from F at our disposal. Therefore, for the practical implementation, we choose $P = F$, $N = n$ and use $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ as $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N$, respectively. However, when $\mathbf{x}_i$ is used as $\mathbf{a}_i$, instead of $\mathcal{T}_n(F^{\mathbf{a}_i}) = \mathbb{T}_{\sigma_n,n}(F^{\mathbf{a}_i})$, we compute $\mathcal{T}_{n-1}^{(-i)}(F^{\mathbf{a}_i}) = \mathbb{T}_{\sigma_{n-1},n-1}^{(-i)}(F^{\mathbf{a}_i})$ based on the remaining $(n-1)$ observations leaving $\mathbf{x}_i$. For these choices of P and $\mathcal{T}$, we shall denote $\zeta_{\mathcal{T}}^{P}(F)$ as $\zeta(F)$, which is given by

$$\zeta(F) = \int_{\mathbb{R}^d} \mathcal{T}(F^{\mathbf{a}})\, dF(\mathbf{a}) = \int_{\mathbb{R}^d} \mathbb{T}_{\sigma_0}(F^{\mathbf{a}})\, dF(\mathbf{a}).$$

The corresponding estimator is denoted by $\zeta_n(F)$, and it can be expressed as

$$\zeta_n(F) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{T}_{n-1}^{(-i)}(F^{\mathbf{x}_i}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{T}_{\sigma_{n-1},n-1}^{(-i)}(F^{\mathbf{x}_i}).$$

Different steps of our algorithm is given below.

Algorithm

1. For a fixed i $(1 \le i \le n)$, compute p-dimensional vectors $\mathbf{z}_{1,i}, \ldots, \mathbf{z}_{i-1,i}, \mathbf{z}_{i+1,i}, \ldots, \mathbf{z}_{n,i}$, where that $\mathbf{z}_{j,i} = (\|\mathbf{x}_j^{(1)} - \mathbf{x}_i^{(1)}\|, \|\mathbf{x}_j^{(2)} - \mathbf{x}_i^{(2)}\|, \ldots, \|\mathbf{x}_j^{(p)} - \mathbf{x}_i^{(p)}\|)$ for $j = 1, 2, \ldots, n; j \ne i$.
2. Compute $\mathbb{T}_{\sigma_{n-1},n-1}^{(-i)}(F^{\mathbf{x}_i})$ based on these $(n-1)$ observations $\{\mathbf{z}_{j,i};\ 1 \le j \le n, j \ne i\}$ to measure dependence among p coordinate variables.
3. Repeat Steps 1 and 2 with $i = 1, 2, \ldots, n$ to compute the test statistic
$$\zeta_n(F) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{T}_{\sigma_{n-1},n-1}^{(-i)}(F^{\mathbf{x}_i}).$$

The null hypothesis $\mathbb{H}_0$ is rejected if the observed value of the test statistic ζ_n is larger than the cut-off, which is computed based on the permutation principle discussed before. Note that the statistic ζ_n is based on distances among the observations. So, it can be

conveniently used for data of arbitrary dimensions. When F is absolutely continuous, consistency of ζ_n can be derived from Theorems 4.1 and 4.2. Consistency of the resulting test follows from that. This result is asserted by the following theorem.

Theorem 4.3. *Suppose that* $\mathbf{X}$ *follows an absolutely continuous distribution* F*. Then, under any fixed alternative, the power of the right-tailed test based on* ζ_n *converges to 1 as* n *tends to infinity.*

4.2 Analysis of simulated data sets

We analyzed some simulated data sets to compare the performance of our proposed test based on ζ_n with the JdCov test, the rank-JdCov test (Chakraborty and Zhang, 2019) and the dHSIC test (Pfister *et al.*, 2018). For problems involving two sub-vectors, we also used the HHG test (Heller *et al.*, 2013) for comparison. For our proposed tests, we created an R package 'MCGK' containing all necessary codes. This package is available at `https://github.com/angshumanroycode/MCGK`. Cut-offs of all these tests were computed based on 1000 random permutations as before.

First, we considered six examples involving two random vectors each of dimension 5. For each of these examples, we considered samples of different sizes, and the powers of different tests (levels, if $\mathbb{H}_0$ is true) were computed based on 1000 Monte Carlo simulations. These results are reported in Figure 4.1.

To study the level properties of different tests, we began with an example, where observations were generated from the 10-dimensional standard 'Normal' distribution. While the vector $\mathbf{X}^{(1)}$ consisted of the first five variables, $\mathbf{X}^{(2)}$ was formed by the rest. Clearly, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent in this example. So, as expected, all tests rejected $\mathbb{H}_0$ in nearly 5% cases (see Figure 4.1(a)).

To study the power properties of the tests, next we generated observations from the 10-dimensional standard 't_5' distribution (Student's t distribution with 5 degrees of freedom), and split them into two sub-vectors of dimension 5 each. Note that in this example, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are uncorrelated but not independent. So, different tests can be compared based on their powers. Figure 4.1(b) clearly shows that in this example, the HHG test and our proposed test performed much better than dHSIC, JdCov and rank-JdCov tests.

The next two examples deal with mixtures of 10-dimensional normal distributions. In 'Mixture Normal-1', we generated observations from an equal mixture of $N_{10}(\mathbf{0}, \mathbf{I}_{10})$ and

$N_{10}(\mathbf{0}, 4\mathbf{I}_{10})$ distributions, where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. In 'Mixture Normal-2', they were generated from an equal mixture of $N_{10}(\mathbf{0}, \mathbf{\Sigma}_1)$ and $N_{10}(\mathbf{0}, \mathbf{\Sigma}_2)$ distributions with $\mathbf{\Sigma}_1 = \text{diag}(\mathbf{I}_5, 4\mathbf{I}_5)$ and $\mathbf{\Sigma}_2 = \text{diag}(4\mathbf{I}_5, \mathbf{I}_5)$. In both of these examples, our proposed test had an edge over the HHG test (see Figures 4.1(c) and 4.1(d)). JdCov, rank-JdCov and dHSIC tests had very poor performance in the 'Mixture Normal-2' example. The JdCov test and its rank version performed poorly in the 'Mixture Normal-1' example as well.

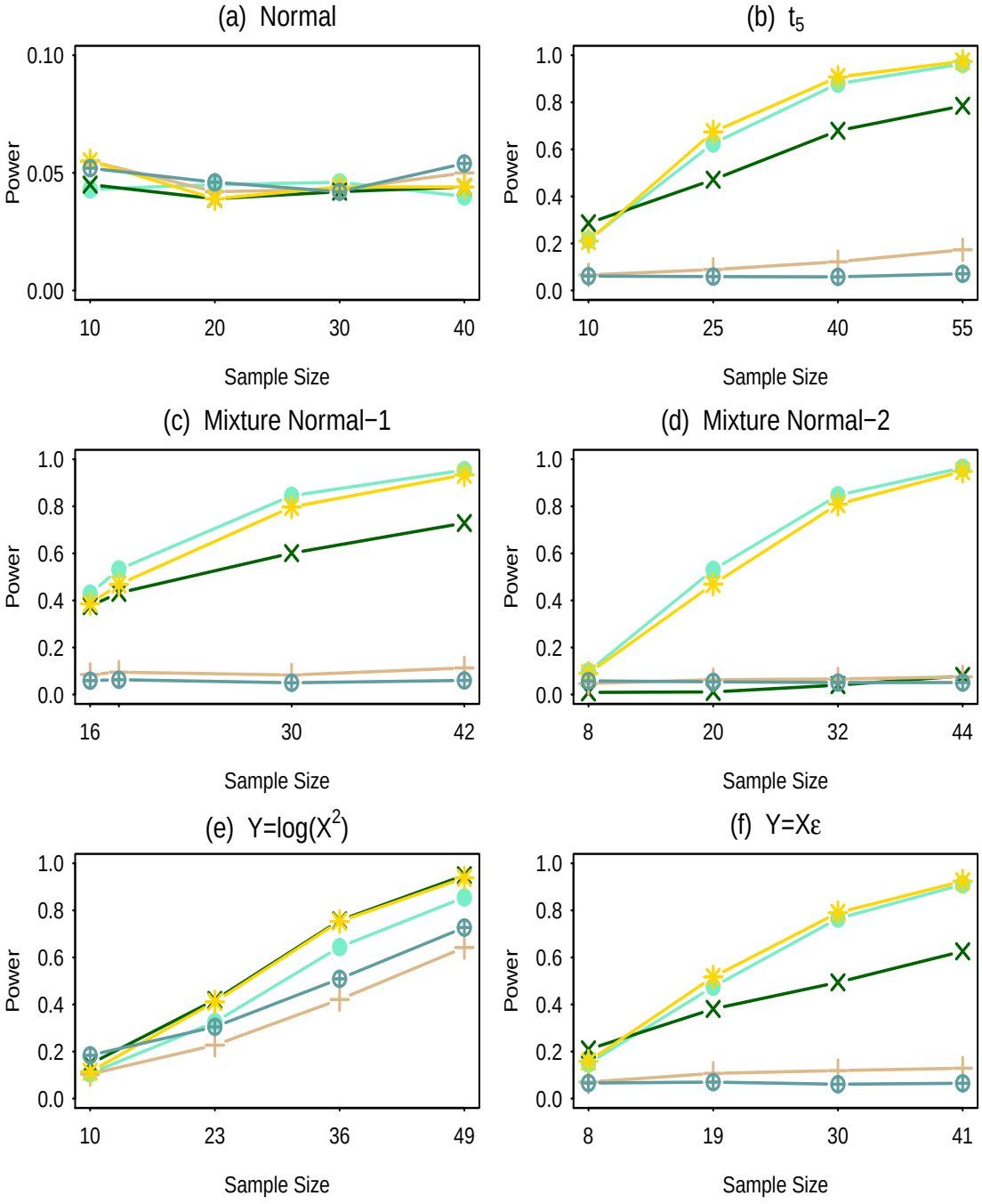


FIGURE 4.1: Powers of JdCov (+), rank-JdCov (⊕), dHSIC (×), HHG (✳) tests and the proposed test based on ζ_n (●) in simulated data sets with two sub-vectors ($p = 2$, $d_1 = d_2 = 5$).

We also considered two other examples, which can be viewed as multi-dimensional versions of the examples used in Székely *et al.* (2007). In both of these examples, observations on $\mathbf{X}^{(1)}$ were generated from the 5-dimensional standard normal distribution. In '$Y = \log(X^2)$' example, each coordinate of $\mathbf{X}^{(2)}$ was obtained from the corresponding coordinate of $\mathbf{X}^{(1)}$ by using the transformation $y = \log(x^2)$. In '$Y = X\epsilon$' example, each coordinate of $\mathbf{X}^{(2)}$ was generated by multiplying the corresponding coordinate of $\mathbf{X}^{(1)}$ with an independent $N(0, 1)$ noise. In the '$Y = \log(X^2)$' example, the dHSIC test had the best performance closely followed by HHG and our proposed test (see Figure 4.1(e)). The JdCov test had relatively low power. In the '$Y = X\epsilon$' example, the HHG test and the proposed test outperformed their competitors (see Figure 4.1(f)). JdCov and rank-JdCov tests had miserable performance in this example.

Next, we considered some problems involving 4 random vectors, each of dimension 5. In each of these examples, we generated observations from a 20-dimensional distribution. The first five variables formed the sub-vector $\mathbf{X}^{(1)}$, the next five formed the sub-vector $\mathbf{X}^{(2)}$, and so on. Note that the HHG test could not be used in these examples. So, we compared the performance of our test with JdCov, rank-JdCov and dHSIC tests. Again we considered six examples. They are labeled as 'Normal', 't_5', 'Mixture Normal-1', 'Mixture Normal-2', 'Hypersphere' and 'L_1 Ball'. The first four examples can be viewed as four-component extensions of the examples considered in Figures 4.1(a)-4.1(d).

In the 'Normal' example, we generated observations from the 20-dimensional standard normal distribution. In this example, since the $\mathbf{X}^{(i)}$'s ($i = 1, 2, 3, 4$) are independent, one expects all tests to have powers close to the nominal level of 0.05, and we observed the same in our experiment (see Figure 4.2(a)). Next, we replaced the normal distribution by 20-dimensional standard t_5 distribution. In this example, $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}, \mathbf{X}^{(4)}$ are uncorrelated but not independent. The proposed test and the dHSIC test could identify this dependency well (see Figure 4.2(b)). Among them, the former one performed better.

The next two examples, 'Mixture Normal-1' and 'Mixture Normal-2', are similar to those used before. In 'Mixture Normal-1', we generated 20-dimensional observations from an equal mixture of $N_{20}(\mathbf{0}, \mathbf{I}_{20})$ and $N_{20}(\mathbf{0}, 2\mathbf{I}_{20})$ distributions. In 'Mixture Normal-2', they were generated from an equal mixture of $N_{20}(\mathbf{0}, \mathbf{\Sigma}_1^\circ)$ and $N_{20}(\mathbf{0}, \mathbf{\Sigma}_2^\circ)$ distributions with $\mathbf{\Sigma}_1^\circ = \text{diag}(\mathbf{I}_5, 2\mathbf{I}_5, 3\mathbf{I}_5, 4\mathbf{I}_5)$ and $\mathbf{\Sigma}_2^\circ = \text{diag}(4\mathbf{I}_5, 3\mathbf{I}_5, 2\mathbf{I}_5, \mathbf{I}_5)$. In these two examples, our proposed test performed much better than their competitors (see Figures 4.2(c) and

4.2(d)). JdCov and rank-JdCov tests had poor performance in both of these examples. The dHSIC test had somewhat reasonable performance in 'Mixture Normal-1'.

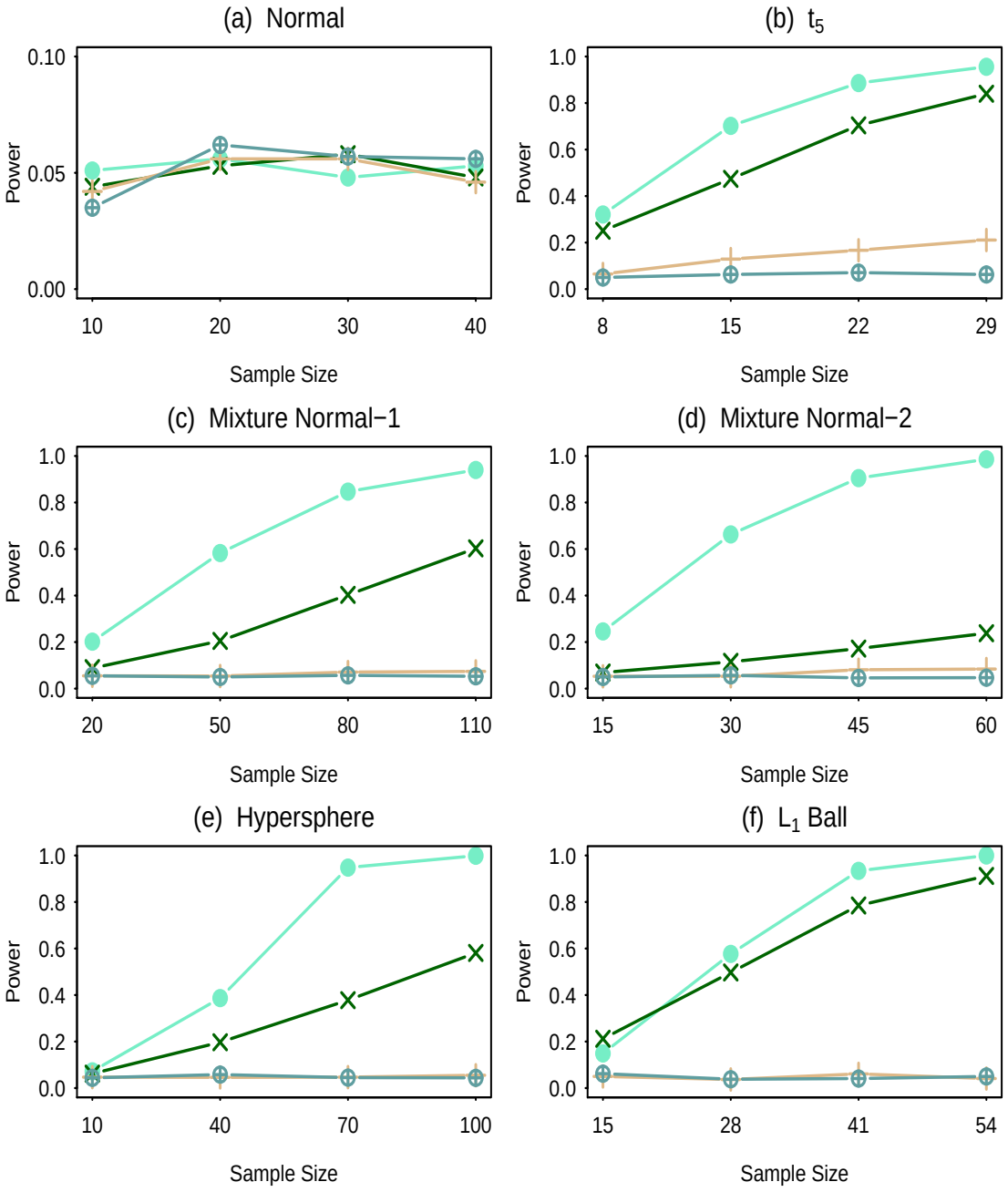


FIGURE 4.2: Powers of JdCov (+), rank-JdCov (⊕), dHSIC (×) tests and the proposed test based on ζ_n (●) in simulated data sets with four sub-vectors ($p = 4$, $d_1 = d_2 = d_3 = d_4 = 5$).

Next we considered two examples involving uniform distributions. In the 'Hypersphere' example, observations on $\mathbf{X}$ were generated from the 20-dimensional uniform distribution on $\{\mathbf{x} = (x_1, x_2, \ldots, x_{20}) : x_1^2 + x_2^2 + \cdots + x_{20}^2 \leq 1\}$, while in the '$L_1$ Ball' example, they were generated from the uniform distribution on $\{\mathbf{x} = (x_1, x_2, \ldots, x_{20}) : |x_1| + |x_2| + \cdots + |x_{20}| \leq 1\}$. Figures 4.2(e) and 4.2(f) show the superiority of our proposed test in these examples.

The dHSIC test also had competitive performance in the 'L_1 Ball' example, but JdCov and rank-JdCov tests did not have satisfactory performance in either of these examples.

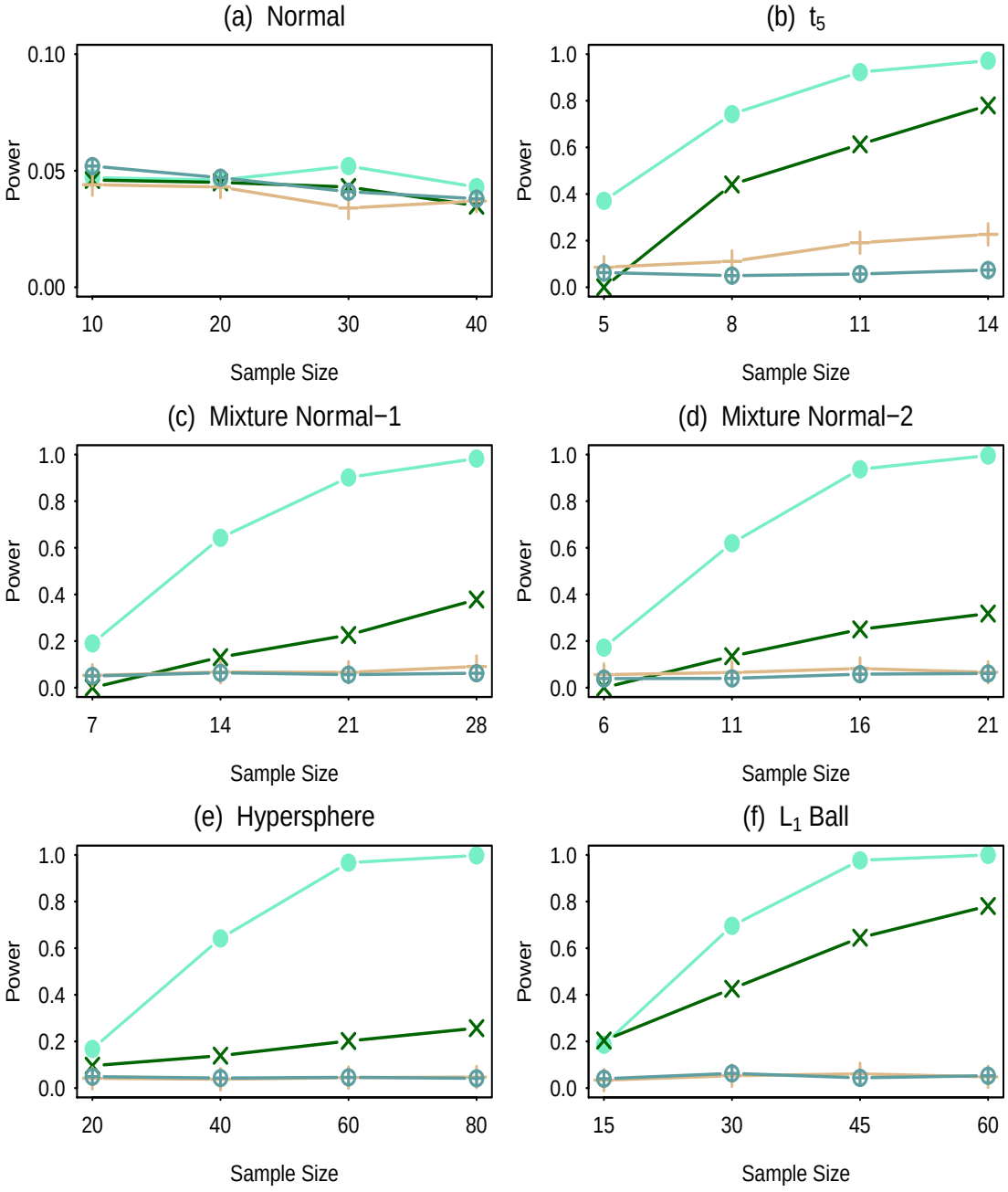


FIGURE 4.3: Powers of JdCov (+), rank-JdCov (⊕), dHSIC (×) tests and the proposed test based on ζ_n (●) in simulated data sets with 15-dimensional sub-vectors ($p = 4$, $d_1 = d_2 = d_3 = d_4 = 15$).

We observed similar results when we repeated these experiments with data of higher dimensions. Figure 4.3 shows the powers of different tests for 15-dimensional versions (i.e., $d_1 = d_2 = d_3 = d_4 = 15$) of the six data sets considered in Figure 4.2. The superiority of our proposed tests based on ζ_n is more evident in this figure.

It is easy to see that for each $\mathbf{a}_i = \mathbf{x}_i$, the computing cost of $\mathbb{T}^{(-i)}_{\sigma_{n-1},n-1}(F^{\mathbf{a}_i})$ is of the order $\mathcal{O}(n^2p)$, which is the same as that of dHSIC and JdCov tests. But, for the proposed test based on ζ_n, we need to repeat it n times (see Sub-section 4.1.2), and that increases its computing cost. However, the results reported in Figures 4.1-4.3 show that in many cases, it is worthy to go for this extra computation.

4.3 Method based on linear projections

So far, for the construction of our test, we have considered characterization of independence among the random vectors based on independence among the pairwise distances (see equation 4.1). Another characterization of independence is given by the Cramér-Wold device, which says that $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are independent if and only if $\widetilde{X}^{(\mathbf{a},1)} = \langle \mathbf{a}^{(1)}, \mathbf{X}^{(1)} \rangle$, $\widetilde{X}^{(\mathbf{a},2)} = \langle \mathbf{a}^{(2)}, \mathbf{X}^{(2)} \rangle, \ldots, \widetilde{X}^{(\mathbf{a},p)} = \langle \mathbf{a}^{(p)}, \mathbf{X}^{(p)} \rangle$ are independent for all $\mathbf{a} = (\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(p)}) \in \mathbb{R}^d$. So, instead of testing independence among Euclidean distances $X^{(\mathbf{a},1)}, X^{(\mathbf{a},2)}, \ldots, X^{(\mathbf{a},p)}$, one can also test for independence among the linear projections $\widetilde{X}^{(\mathbf{a},1)}, \widetilde{X}^{(\mathbf{a},2)}, \ldots, \widetilde{X}^{(\mathbf{a},p)}$ and aggregate the results for several choices of $\mathbf{a}$ as before to come up with a test statistic analogous to ζ_n. We call this test statistic $\widetilde{\zeta}_n$. Consistency results similar to Theorems 4.2 and 4.3 hold for the test based on $\widetilde{\zeta}_n$ as well, and they can easily be proved using similar line of arguments. So, we are not repeating them. However, the empirical performance of these projection-based tests was inferior to our tests based on pairwise distances in almost all simulated examples considered in Section 4.2. In many cases, they had poor performance like the JdCov test. That is why we did not report those results in Section 4.2.

To understand the difference between these two methods, first note that the copula based dependency measure $\mathbb{T}_{\sigma_n,n}(F^{\mathbf{a}})$ depends on $\left(x_i^{(\mathbf{a},1)}, x_i^{(\mathbf{a},2)}, \ldots, x_i^{(\mathbf{a},p)}\right)$ only through $\left(r_i^{(\mathbf{a},1)}, r_i^{(\mathbf{a},2)}, \ldots, r_i^{(\mathbf{a},p)}\right)$ $(i = 1, 2, \ldots, n)$, where $r_i^{(\mathbf{a},j)}$ $(j = 1, 2, \ldots, p)$ is the rank of $x_i^{(\mathbf{a},j)}$ in the set $\{x_1^{(\mathbf{a},j)}, x_2^{(\mathbf{a},j)}, \ldots, x_n^{(\mathbf{a},j)}\}$. So, this statistic is invariant under monotone transformation of coordinate variables (see Theorem 2.1 and also Roy *et al.*, 2020b, for details) and we will get the same result if, instead of $(x_i^{(\mathbf{a},1)}, x_i^{(\mathbf{a},2)}, \ldots, x_i^{(\mathbf{a},p)}) = (\|\mathbf{x}_i^{(1)} - \mathbf{a}^{(1)}\|, \|\mathbf{x}_i^{(2)} - \mathbf{a}^{(2)}\|, \ldots, \|\mathbf{x}_i^{(p)} - \mathbf{a}^{(p)}\|)$, we use $(\|\mathbf{x}_i^{(1)} - \mathbf{a}^{(1)}\|^2, \|\mathbf{x}_i^{(2)} - \mathbf{a}^{(2)}\|^2, \ldots, \|\mathbf{x}_i^{(p)} - \mathbf{a}^{(p)}\|^2)$ for $i = 1, 2, \ldots, n$. Now, observe that $\|\mathbf{x}_i^{(j)} - \mathbf{a}^{(j)}\|^2 = \|\mathbf{x}_i^{(j)}\|^2 + \|\mathbf{a}^{(j)}\|^2 - 2\widetilde{x}_i^{(\mathbf{a},j)}$. So, for any fixed j $(j = 1, 2, \ldots, p)$, the ranks $y_i^{(j)}$'s $(i = 1, 2, \ldots, n)$ depend not only on $\widetilde{x}_1^{(\mathbf{a},j)}, \widetilde{x}_2^{(\mathbf{a},j)}, \ldots, \widetilde{x}_n^{(\mathbf{a},j)}$ but also on the values of $\|\mathbf{x}_1^{(j)}\|^2, \|\mathbf{x}_2^{(j)}\|^2, \ldots, \|\mathbf{x}_n^{(j)}\|^2$. When these

norms are constant, the test based on pairwise distances and that based on linear projections lead to the same result. But in many cases, these norms carry significant information about dependency among the random vectors. In such cases, the tests based on pairwise distances are expected to perform better. We also observed the same in our experiments in Section 4.2.

To demonstrate these above mentioned facts, we considered three simple examples. In Example A, we generated n i.i.d. observations $\theta_1, \theta_2, \ldots, \theta_n$ from the $U(0, 2\pi)$ distribution and then computed $\phi_i = (\theta_i + \varepsilon_i)\ mod(2\pi)$ for $i = 1, 2, \ldots, n$, where the ε_i's were independently generated from a wrapped normal distribution (see, e.g., Breitenberger, 1963; Mardia and Jupp, 2009) with the location parameter 0 and the scale parameter 1. Observations on the two sub-vectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ were obtained using the transformations $\mathbf{x}_i^{(1)} = (\cos\theta_i, \sin\theta_i)$ and $\mathbf{x}_i^{(2)} = (\cos\phi_i, \sin\phi_i)$ for $i = 1, 2, \ldots, n$. This experiment was repeated 1000 times to compute the powers of the two tests. Note that in this example, we have $\|\mathbf{x}_i^{(1)}\| = \|\mathbf{x}_i^{(2)}\| = 1$ for all $i = 1, 2 \ldots, n$. Therefore, as expected, the test based on pairwise distances and that based on linear projections had the same power for all choices of n (see Figure 4.4(a)).

In Example B, $\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \ldots, \mathbf{x}_n^{(1)}$ were generated from the bivariate normal distribution with zero means, unit variances and correlation coefficient 0.5. Observations on $\mathbf{X}^{(2)}$ were obtained from the corresponding observations on $\mathbf{X}^{(1)}$ by using the transformation $\mathbf{x}_i^{(2)} = (\|\mathbf{x}_i^{(1)}\| \cos\theta_i,\ \|\mathbf{x}_i^{(1)}\| \sin\theta_i)$ $(i = 1, 2, \ldots, n)$, where $\theta_1, \ldots, \theta_n$ are i.i.d. $U(0, 2\pi)$ variables. So, in this example, the norms of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are non-constant, but they are equal with probability one. The tests based on ζ_n successfully utilized this information to come up with much better performance than the test based on $\widetilde{\zeta}_n$ (see Figure 4.4(b)). Similar phenomenon was observed in the examples considered in Section 4.2 as well.

However, the projection-based test can also have higher power than the pairwise distance-based test in some examples. To demonstrate this, in Example C, we considered the same setup as in Example A for generating the θ_i's and ϕ_i's for $i = 1, 2, \ldots, n$. In addition to that, we generated n independent observations (U_i, V_i) $(i = 1, 2, \ldots, n)$ from the uniform distribution on $[0, 10] \times [0, 10]$. Observations on $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ were obtained using the transformations $\mathbf{x}_i^{(1)} = (U_i \cos\theta_i, U_i \sin\theta_i)$ and $\mathbf{x}_i^{(2)} = (V_i \cos\phi_i, V_i \sin\phi_i)$ for $i = 1, 2, \ldots, n$. In this example, $\|\mathbf{x}_i^{(1)}\| = U_i$ and $\|\mathbf{x}_i^{(2)}\| = V_i$ being independent random variables, do not contain any information regarding dependence between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. In fact, they can be

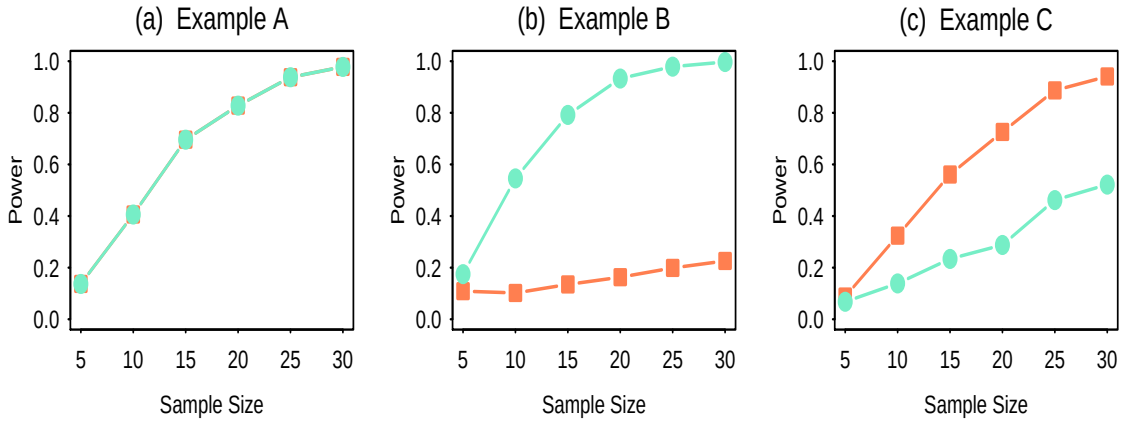


FIGURE 4.4: Powers of the tests based on pairwise distances ζ_n (●) and linear projections $\widetilde{\zeta}_n$ (■) in Examples A, B and C.

viewed as noise. In such a situation, the test based on linear projection performed better than the test based on pairwise distances (see Figure 4.4(c)).

For one-dimensional problems (i.e., $d_1 = d_2 = \cdots = d_p = 1$), the test based on $\widetilde{\zeta}_n$ usually yields similar results as obtained by the test based on T_n (considered in Chapter2). This is quite clear from the description of the test statistics. When all sub-vectors are one-dimensional, since $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(p)}$ are scalars and $\mathbb{T}_{\sigma_n,n}(F^{\mathbf{a}})$ is invariant under monotone transformation of the coordinate variables, for a fixed choice of $\mathbf{a}$, the test statistic based on linear projection can be viewed as the test statistic T_n computed using $n-1$ observations (leaving out the one which is used as $\mathbf{a}$). However, the statistic ζ_n leads to a different test in one-dimension, and it often yields different results.

To demonstrate this, we considered two simple examples each involving four-dimensional distributions. In these examples, each coordinate variable was used as a sub-vector (i.e. $d_1 = d_2 = d_3 = d_4 = 1$). In Example D1, observations were generated from a multivariate normal distribution with the mean vector $\mathbf{0} = (0,0,0,0)$ and the dispersion matrix $\mathbf{\Sigma} = ((\sigma_{ij}))$ with $\sigma_{ij} = 0.4^{|i-j|}$ for $i,j = 1,2,3,4$. In Example D2, they were generated from the standard Cauchy distribution. Each experiment was repeated 1000 times as before to compute the powers of different tests, and they are reported in Figure 4.5. This figure clearly shows that in both of these examples, the test based on $\widetilde{\zeta}_n$ had powers similar to that based on T_n, but the test based on ζ_n had different powers. In the second example, the tests based on pairwise distances outperformed the tests based on linear projections (see Figure 4.5(b)), but in the first example, the tests based on linear projections had better performance (see Figure 4.5(a)). This can be explained using the arguments given in the

earlier part of this section. In the example with Cauchy distribution, the average correlation coefficient between $|X^{(1)}|$ and $|X^{(2)}|$ over the 1000 trials was found to be more than 0.75. Since the four variables are exchangeable in this example, we observed similar correlations for other pairs of variables as well. This clearly indicates that the absolute values of the variables carried substantial information regarding dependence. But in the example with normal distribution, this dependence among the absolute values of the variables was very weak. We could not find any pattern in the scatter plots, and the maximum of the average pairwise correlations was found to be smaller than 0.2.

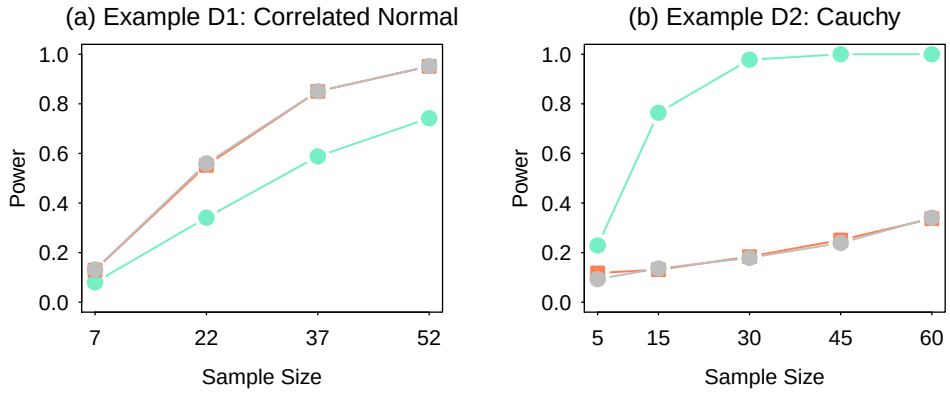


FIGURE 4.5: Powers of the tests based on pairwise distances ζ_n (●), linear projections $\widetilde{\zeta}_n$ (■) and that based on T_n (●) in Examples D1 and D2 involving four one-dimensional variables.

4.4 Results from the analysis of real data sets

We also analyzed three real data sets for further evaluation of our proposed methods. Description of the 'Combined Cycle Power Plant (CCPP) data' has already been given in Chapter 2. Like before, here also we did not consider the variable 'electric energy output' for our analysis and carried out tests of independence among the remaining four variables. The other two data sets, namely 'Tecator data' and 'Pollution data', are available at the CMU Data Set Archive (`http://lib.stat.cmu.edu/datasets/`). Brief description of these two data sets is given below.

Tecator data were recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850-1050 nm by the Near Infrared Transmission principle. This data set was analyzed by Ferraty and Vieu (2006). Here each sample contains finely chopped pure meat with different water, fat and protein contents. For each of the 215 meat samples, the data consist of 100 channel spectrum of absorbance, the content of water, the content of fat and that of protein. The absorbance is measured by the spectrometer, while the

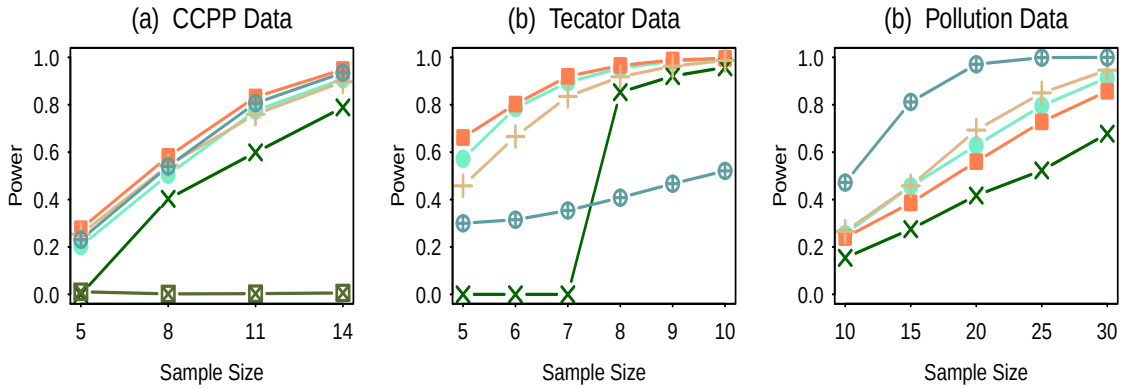


FIGURE 4.6: Powers of JdCov (+), rank-JdCov (⊕), dHSIC (×), Hoeffding (⊠) tests and the proposed tests based on ζ_n (●) and $\widetilde{\zeta}_n$ (■) in real data sets.

three contents are determined by analytic chemistry. Now, a natural question to ask is whether protein content, fat content or water content in the meat can be assessed based on the absorbance spectra. So, here, we want to test whether the relationship among these four variables is statistically significant.

Pollution data set contains measurements on four groups of variables: weather, socio-economic status, pollution and age-adjusted mortality rate. There are several variables under each of these groups. Data were collected from 60 different standard metropolitan statistical areas of the USA in the year 1960, and this data set first appeared in McDonald and Schwing (1973). Here one might be interested in testing whether there is any dependence among these four groups of variables.

Like before, in these examples also, when we used the full data sets for testing (for the CCPP data, where all random vectors are of dimension 1, test based on Hoeffding's Φ-statistic was also used as a competitor), all tests suggested significant dependence among the random vectors. Since it was not possible to compare among the performance of different tests based on those results on full data sets, following Sarkar and Ghosh (2018), we carried out our experiments with randomly chosen subsets of different sizes. For each subset size, the experiment was performed 1000 times to compute the empirical powers of different tests, which are shown in Figure 4.6.

In CCPP data, the test based on $\widetilde{\zeta}_n$ had the highest power, but the JdCov test, the rank-JdCov test and the test based on ζ_n also had competitive performance. The dHSIC test had relatively low power. The test based on Hoeffding's Φ-statistic had miserable performance. It had power close to zero for all four choices of the sample size considered here. In the case of Tecator data, our proposed tests based on ζ_n and $\widetilde{\zeta}_n$ had similar

performance, and they outperformed all other tests. Note that the dHSIC test could only be used for samples of size 8 or higher. For smaller sample sizes, its power was taken to be zero. In the case of Pollution data, the rank-JdCov test had the best performance. The test based on ζ_n and the JdCov test also performed well, and they had almost similar powers. The test based on $\widetilde{\zeta}_n$ had competitive performance as well, but the power of the dHSIC test was slightly lower.

4.5 Multi-scale versions of the proposed tests

For our proposed tests, so far, we have used the bandwidth chosen based on median heuristic. However, as we have seen before, this may not always be the best choice. When the relationships among the random vectors are nearly monotone (i.e., the conditional expectation of a variable is a monotone function of others), median heuristic usually yields good results. But, smaller bandwidths are often helpful for detecting complex non-monotone relationships. To take care of this problem, we adopt the multi-scale approach as before and aggregate the results for several choices of the bandwidth. For any fixed $\mathbf{a}_i = \mathbf{x}_i$, we consider the results obtained for m bandwidths $\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(m)}$ and use either $\mathbb{T}^{(-i)}_{\text{sum}}(F^{\mathbf{x}_i}) = \sum_{j=1}^{m} \mathbb{T}^{(-i)}_{\sigma^{(j)},n-1}(F^{\mathbf{x}_i})$ or $\mathbb{T}^{(-i)}_{\text{max}}(F^{\mathbf{x}_i}) = \max_{1\le j\le m} \mathbb{T}^{(-i)}_{\sigma^{(j)},n-1}(F^{\mathbf{x}_i})$ as the aggregated statistic. We take the average of the $\mathbb{T}^{(-i)}_{\text{sum}}(F^{\mathbf{x}_i})$'s (respectively, $\mathbb{T}^{(-i)}_{\text{max}}(F^{\mathbf{x}_i})$'s) over all $\mathbf{a}_i = \mathbf{x}_i$ to come up with the test statistic $\zeta_{\text{sum},n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{T}^{(-i)}_{\text{sum}}(F^{\mathbf{x}_i})$ (respectively, $\zeta_{\text{max},n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{T}^{(-i)}_{\text{sum}}(F^{\mathbf{x}_i})$). For the multi-scale versions of the tests based on linear projections, similar test statistics can be constructed, and we denote them by $\widetilde{\zeta}_{\text{sum},n}$ and $\widetilde{\zeta}_{\text{max},n}$, respectively. For the choice of m and $\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(m)}$, we adopt the same strategy as in Section 2.5. For any fixed m, large sample consistency of these tests follows from Theorems 4.1 and 4.2 using similar line of arguments as used in the proof of Theorem 4.3.

To demonstrate the utility of these multi-scale approaches, we used six examples each involving two bivariate random vectors. For generating observations on these sub-vectors, we considered the six unusual bivariate data sets of Newton (2009), namely, 'Four clouds', 'W', 'Parabola', 'Two parabolas', 'Diamond' and 'Circle' (see Figure 2.5 for the scatter plots of these data sets). Recall that in each of these examples, the two variables are uncorrelated, but barring the first example, they are not independent. We generated observations from these bivariate distributions and considered them as the first coordinates of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively. The second coordinates of these two vectors were generated independently

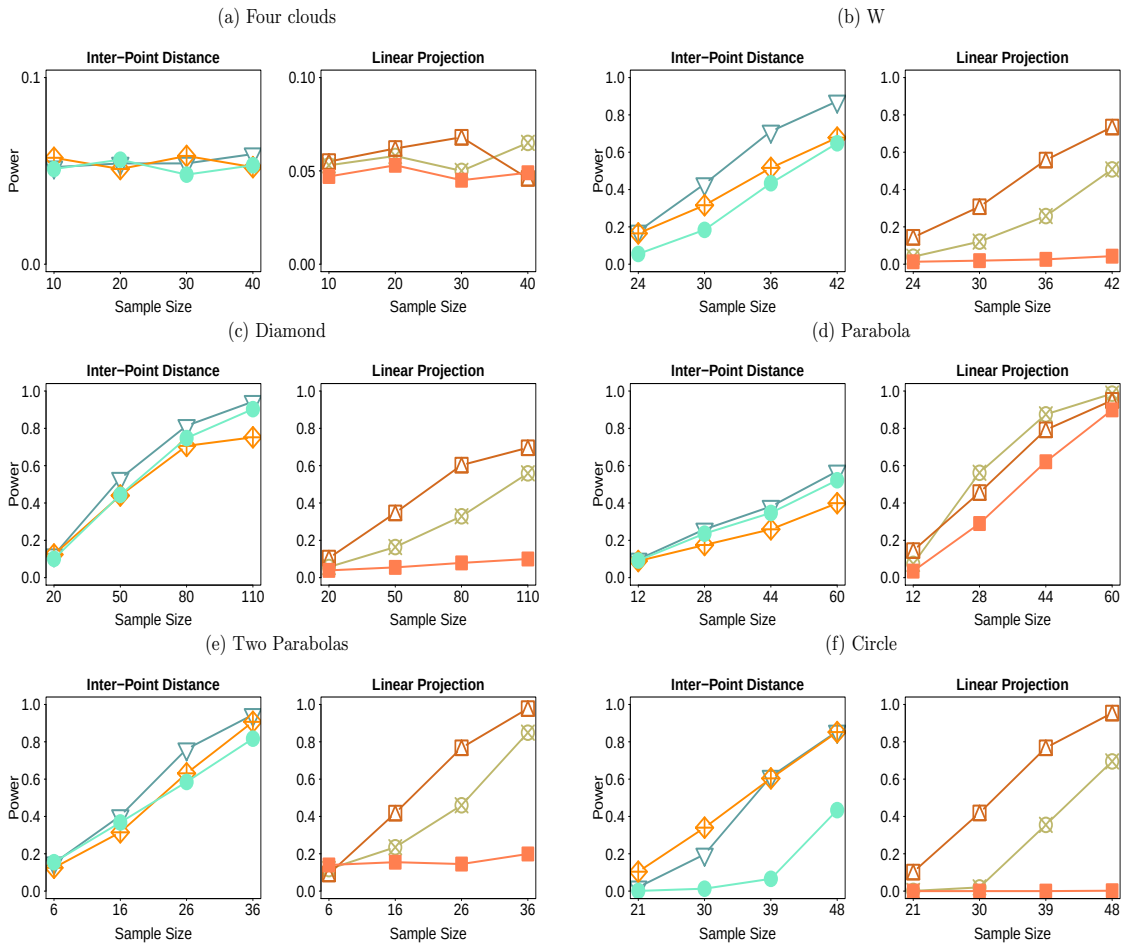


FIGURE 4.7: Powers of the single-scale and multi-scale tests based on pairwise distances (tests based on ζ_n (●), $\zeta_{\text{sum},n}$ (▽) and $\zeta_{\max,n}$ (◈)) and those based on linear projections (tests based on $\widetilde{\zeta}_n$ (■), $\widetilde{\zeta}_{\text{sum},n}$ (⊗) and $\widetilde{\zeta}_{\max,n}$ (⊠)) in simulated data sets with two sub-vectors.

from the $N(0, 1/9)$ distribution. We considered samples of different sizes and for each sample size, the experiment was repeated 1000 times to compute the powers of single-scale and multi-scale versions of the proposed tests. These powers are reported in Figure 8.

In the case of 'Four clouds' data set, two random vectors were independent. So, as expected, all tests had powers close to the nominal level of 0.05 (see Figure 4.7(a)). Figure 4.7 clearly shows that multi-scale versions of the proposed tests performed well in the other five examples. In many cases, they outperformed their corresponding single-scale analogs. Among the multi-scale methods based on pairwise distances, the test based on $\zeta_{\text{sum},n}$ had better performance. In all cases, it had higher power than its single-scale analog based on ζ_n (see Figures 4.7(b)-4.7(f)). Except for the 'Circle' data, in all other cases, it outperformed the test based on $\zeta_{\max,n}$ as well. While the single-scale method based on linear projections performed poorly in four out of five examples, its multi-scale versions based on $\widetilde{\zeta}_{\text{sum},n}$ and $\widetilde{\zeta}_{\max,n}$ worked well in all five examples. Among these two tests, the later one had an edge.

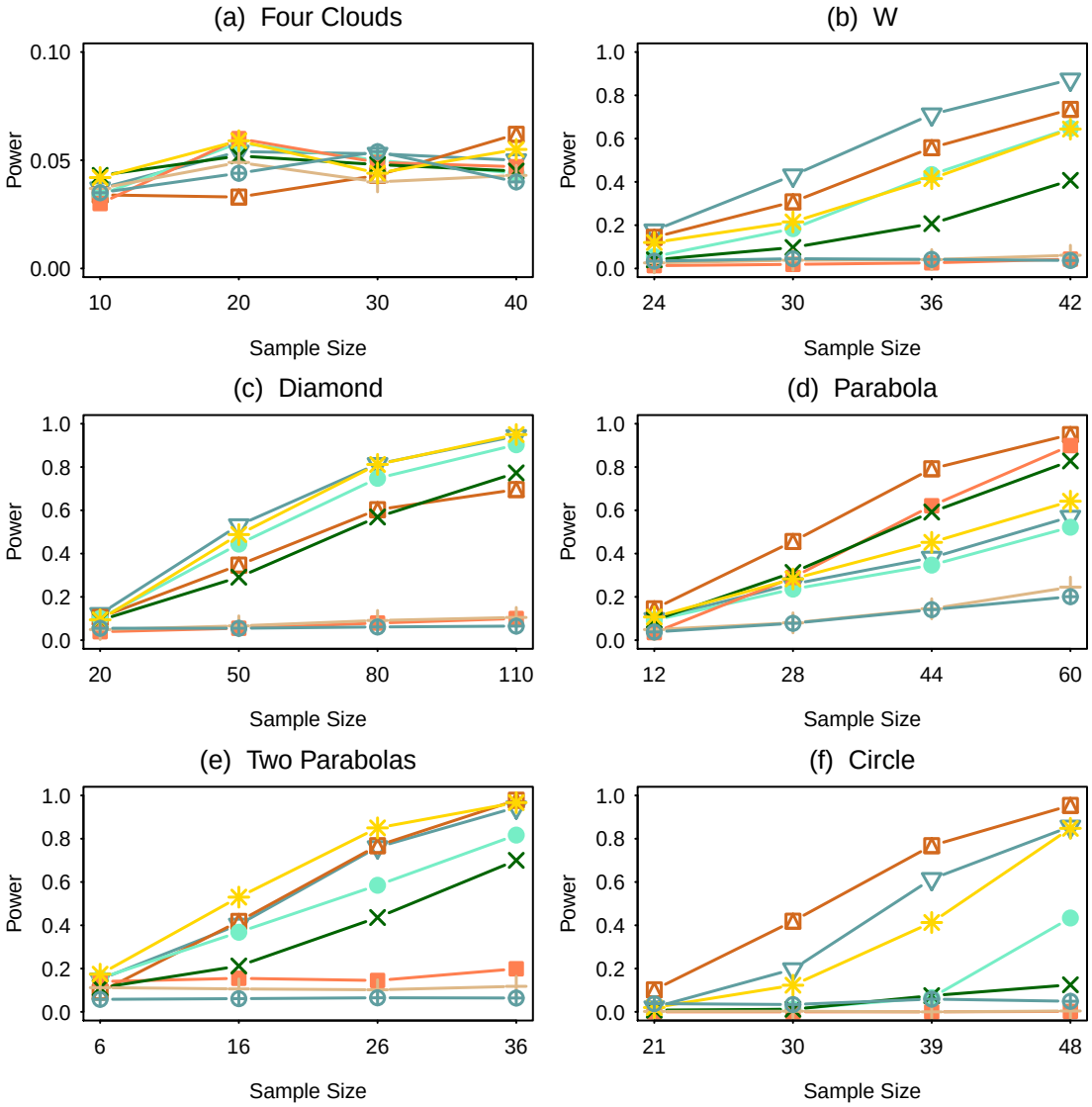


FIGURE 4.8: Powers of JdCov (+), rank-JdCov (⊕), dHSIC (×), HHG (✳) tests and the proposed tests based on ζ_n (●), $\zeta_{\text{sum},n}$ (▽), $\widetilde{\zeta}_n$ (■) and $\widetilde{\zeta}_{\text{max},n}$ (▵) in simulated data sets with two 2-dimensional sub-vectors ($p = 2$, $d_1 = d_2 = 2$).

Next we compared the performance of these tests with HHG, JdCov, rank-JdCov and dHSIC tests. Figure 4.8 shows the powers of these tests along with those of the single-scale and multi-scale tests based on pairwise distances and linear projections. Among the multi-scale methods, results are reported for $\zeta_{\text{sum},n}$ and $\widetilde{\zeta}_{\text{max},n}$ only because of their better performance. In all these examples, JdCov and rank-JdCov tests had miserable performance. Performance of the dHSIC tests was also unsatisfactory in some examples, especially in cases of 'W' and 'Circle' data. In the case of 'W' data, multi-scale versions of our tests outperformed their competitors. Among the rest, the HHG test and the single-scale test based on pairwise distances had competitive performance. Inter-point distance

based single-scale and multi-scale tests performed well in 'Diamond' data as well. These two tests and the HHG test had higher powers than the rest. In 'Parabola' data, projection based tests, especially the one based on $\widetilde{\zeta}_{\max,n}$ had superior performance. The dHSIC test had the third highest power in this example. In 'Two parabolas' data set, the HHG test had the best performance, but the performance of the multi-scale methods and the single-scale method based on ζ_n was also satisfactory. They performed better than the rest of their competitors. Our single-scale tests could not perform well in 'Circle' data, but their multi-scale analogs outperformed all other tests considered here.

These examples clearly show the usefulness of the multi-scale approach in complex data sets. In the examples considered in Sections 4.2 and 4.4, these multi-scale methods had almost similar or slightly improved performance compared to their single-scale analogs. That is why here we do not report them again.

4.6 Results from the analysis of functional data

From the description of our proposed tests based on pairwise distances, it is clear that they can also be used for testing independence among several random functions having distributions on infinite dimensional Banach spaces. If these Banach spaces are separable, under appropriate regularity conditions, we can prove the large sample consistency of the test based on ζ_n (see Theorems 4.4 and 4.5 in Section 4.8). Following similar line of arguments, one can establish consistency for the multi-scale versions of the test based on $\zeta_{\text{sum},n}$ and $\zeta_{\max,n}$ as well. Similarly, the tests based on linear projections (i.e., inner products) can be used for functions in infinite dimensional Hilbert spaces, and their consistency can be proved under similar regularity conditions. Now, one may be curious to know how these tests perform in practice for functional data. In this section, we analyze several simulated data sets to address this question.

Over the last couple of decades, several nonparametric methods have been developed in the literature for dealing with functional data (see, e.g. Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). But, unfortunately, the literature on test of independence between two or more random functions is almost non-existent. Lyons (2013) generalized the notion of the distance correlation for random functions in metric spaces of strong negative type (e.g., the Hilbert space of square integrable functions on $[0, 1]$), and hence they generalized the dCov test (Székely *et al.*, 2007) for testing independence between two random

functions. The underlying principle is simple. The dCov test statistic depends only on pairwise distances among the observations. So, using an appropriate distance function, this can be used for function valued data. Using the same idea, one can generalize the JdCov test (Chakraborty and Zhang, 2019) for testing independence among several random functions. Interestingly, the dHSIC test (Pfister *et al.*, 2018), when used with Gaussian kernel, depends only on pairwise distances. The HHG test (Heller *et al.*, 2013) based on 2×2 contingency tables also has the same property. So, these tests can also be used to test for independence between random functions. The consistency of HHG test for functional data can be established using the proof of Theorem 1 in Sarkar *et al.* (2020). In this section, we compare the performance of our proposed tests with that of JdCov, dHSIC and HHG tests, when they are used for testing independence between two random functions. Note that the rank-JdCov test can not be used for such functional data.

We used six examples each involving two random functions $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ in $\mathcal{L}_2[0,1]$, the space of square integrable functions on $[0,1]$. Let $\{\xi_j(t)\}_{j\geq 1}$ be an orthonormal basis of $\mathcal{L}_2[0,1]$. Consider two sequence $\{\nu_j^{(1)}\}_{j\geq 1}$ and $\{\nu_j^{(2)}\}_{j\geq 1}$ of positive real values such that $\sum_{j=1}^{\infty}\left(\nu_j^{(1)}\right)^2 < \infty$ and $\sum_{j=1}^{\infty}\left(\nu_j^{(2)}\right)^2 < \infty$. Also consider a sequence of independent pairs of random variables $\{(Z_j^{(1)}, Z_j^{(2)})\}_{j\geq 1}$. Note that $\mathbf{X}^{(1)}, \mathbf{X}^{(2)} \in \mathcal{L}_2[0,1]$ can be generated as using the formulae $\mathbf{X}^{(1)} = \sum_{j=1}^{\infty}\nu_j^{(1)}Z_j^{(1)}\xi_j(t)$ and $\mathbf{X}^{(2)} = \sum_{j=1}^{\infty}\nu_j^{(2)}Z_j^{(2)}\xi_j(t)$ for all $t \in [0,1]$. For our study, we considered the Fourier basis $\{1, \sqrt{2}\sin(2\pi jt), \sqrt{2}\cos(2\pi jt),\ j = 1,2,\ldots\}$ of $\mathcal{L}_2[0,1]$ and used $\nu_j^{(1)} = j^{-5/4}, \nu_j^{(2)} = j^{-3/2}$ for all $j \geq 1$. Independent bivariate observations on $(Z_1^{(1)}, Z_1^{(2)}), (Z_2^{(1)}, Z_2^{(2)}), \ldots, (Z_9^{(1)}, Z_9^{(2)})$ were generated from a bivariate distribution to get observations on the random functions $\mathbf{X}^{(1)}(t) = \sum_{j=1}^{9}\nu_j^{(1)}Z_j^{(1)}\xi_j(t)$ and $\mathbf{X}^{(2)}(t) = \sum_{j=1}^{9}\nu_j^{(2)}Z_j^{(2)}\xi_j(t)$. Note that because of the use of orthonormal basis, here all pairwise L_2 distances and inner products between the observations can be calculated easily. Here, we used six examples, where the observation on the $(Z_j^{(1)}, Z_j^{(2)})$'s were generated from the six bivariate distributions considered in Newton (2009) (see Figure 2.5). For each example, we considered samples of different sample sizes, and for each sample size, the experiment was repeated 1000 times to compute the powers of different tests. These powers are reported in Figure 4.9.

Note that $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent if and only if the two variables associated with the bivariate distribution are independent. Therefore, in the case of 'Four Clouds' data, all tests had powers close to 0.05 (see Figure 4.9(a)). In the other five examples,

$\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are dependent. Figure 4.9 clearly shows that in these examples, multi-scale versions of our tests performed better that their single-scale analogs. The tests based on $\zeta_{\text{sum},n}, \zeta_{\text{max},n}$ and $\widetilde{\zeta}_{\text{max},n}$ outperformed all other tests in 'W' and 'Circle' examples. In these two examples, dHSIC, JdCov and HHG tests did not perform well. In fact, except for the 'Parabola' example, dHSIC and JdCov tests performed poorly throughout. In the 'Parabola' example, the HHG test had the best overall performance, but all other methods barring the test based on $\widetilde{\zeta}_n$ had competitive powers. In 'Diamond' and 'Two Parabolas' examples, the HHG test and our proposed tests based on $\zeta_{\text{sum},n}$, $\zeta_{\text{max},n}$ and $\widetilde{\zeta}_{\text{max},n}$ had almost similar powers, and they performed better than all other tests considered here.

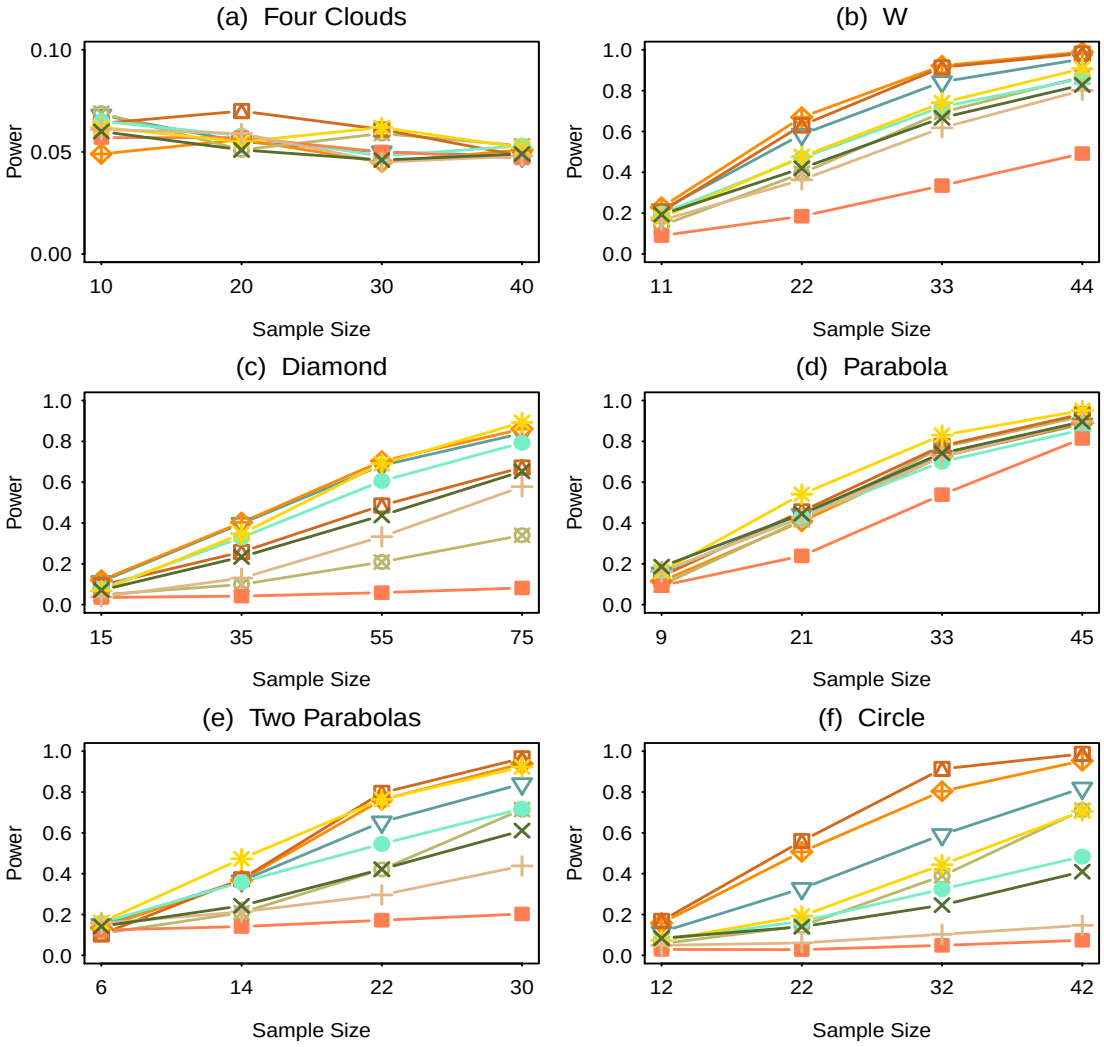


FIGURE 4.9: Powers of JdCov (+), dHSIC (×), HHG (✳) tests and the proposed tests based on ζ_n (●), $\zeta_{\text{sum},n}$ (▽), $\zeta_{\text{max},n}$ (◈), $\widetilde{\zeta}_n$ (■), $\widetilde{\zeta}_{\text{sum},n}$ (⊗) and $\widetilde{\zeta}_{\text{max},n}$ (◬) in functional data sets.

So far, for our data analysis, we assumed all functions to be fully known. But in practice, each function is usually observed only on some grid points, and one needs to approximate

the pairwise L_2 distances or the inner products using those observed values. So, next we considered the situation, where each function was observed on 101 equidistant grid points $\{0.00, 0.01, \ldots, 1.00\}$ on the $[0, 1]$ interval. In such cases, we estimate the pairwise L_2 distance $\|f - g\| = \left(\int_0^1 (f(t) - g(t))^2 dt\right)^{1/2}$ and the inner-product $< f,\ g >= \int_0^1 f(t)g(t)dt$ between two functions f and g by $\left(\frac{1}{101}\sum_{i=0}^{100}(f(t_i) - g(t_i))^2\right)^{1/2}$ and $\frac{1}{101}\sum_{i=0}^{100} f(t_i)g(t_i)$, respectively, where $t_i = 0.01i$ for $i = 0, 1, \ldots, 100$. We carried out our experiment with these estimated values of pairwise L_2 distances and inner-products, but relative performance of all tests were almost the same, and that is why we are not reporting those results again.

4.7 Application in causal discovery

In this section, we use our proposed tests of independence for discovery of causal relationship among the sub-vectors $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$. This type of application was considered in Pfister *et al.* (2018) and Chakraborty and Zhang (2019). In order to unveil the causal relationship among these p random vectors, we consider all possible structural equation models with additive noise, each of which lead to a directed acyclic graph (DAG) on p nodes. Any such structural equation model (SEM) has the following form:

$$\mathbf{X}^{(j)} = f_j(\mathbf{PA}^{(j)}) + \boldsymbol{\epsilon}^{(j)},\ j = 1, 2, \ldots, p; \tag{4.3}$$

where $\mathbf{PA}^{(j)}$ denotes the set of all parents of $\mathbf{X}^{(j)}$ according to the SEM, and the $\boldsymbol{\epsilon}^{(j)}$'s are independent additive noise vectors. If there are no parent nodes, we take f_j to be zero.

Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are independent observations on $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)})$. Given an SEM, for each $j = 1, 2, \ldots, p$, we construct $\hat{f}_j$, an estimate of f_j by regressing $\mathbf{X}^{(j)}$ on its parent nodes $\mathbf{PA}^{(j)}$ using a nonparametric method. Then, the residuals are computed as $\widehat{\boldsymbol{\epsilon}}_i^{(j)} := \mathbf{x}_i^{(j)} - \widehat{f}_j(\mathbf{PA}_i^{(j)})$ for $i = 1, 2, \ldots, n$. If the underlying SEM is correct, these residuals are supposed to be jointly independent. So, we perform a test of independence among the estimated residuals and compute the corresponding p-value. We do it for all possible SEMs that can be represented using DAGs, and finally, the model with the highest p-value (least evidence against independence) is selected. However, if this highest p-value is smaller than 0.05, none of the SEMs is selected. This step can be viewed as the first step used by Hochberg (1988)'s step-up method for multiple testing, which controls the family-wise error rate strongly at 5% level.

We considered two examples for our experiments in this context. In both cases, the true SEM satisfies the identifiability conditions (see Peters *et al.*, 2014, Corollary 31). First, we considered a model involving two bivariate random vectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. Observations on $\mathbf{X}^{(1)} = (U^{(1)}, U^{(2)})$ were generated from the standard bivariate normal distribution, and those on $\mathbf{X}^{(2)} = (V^{(1)}, V^{(2)})$ were obtained from them using the model $V^{(i)} = (U^{(i)})^2 + \varepsilon^{(i)}$, where the $\varepsilon^{(i)}$'s are i.i.d. $N(0, 0.01)$ random variables. So, the actual SEM consists of two nodes with an arrow from $\mathbf{X}^{(1)}$ to $\mathbf{X}^{(2)}$ indicating the dependence of $\mathbf{X}^{(2)}$ on $\mathbf{X}^{(1)}$.

In this example, there are three possible SEMs: (i) $\mathbf{X}^{(1)} \longrightarrow \mathbf{X}^{(2)}$, (ii) $\mathbf{X}^{(1)} \longleftarrow \mathbf{X}^{(2)}$ and (iii) a graph with no edge. We generated 10 observations on $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$, and different methods were used to select the true model. This experiment was repeated 100 times, and the results are reported in Table 4.1. In this example, the HHG test, our single scale method based on ζ_n and its multi-scale version based on $\zeta_{\text{sum},n}$ outperformed all other methods considered here. The JdCov test and its rank version did not perform well. For the test based on linear projections, the multi-scale method based on $\widetilde{\zeta}_{\text{max},n}$ had the best performance, and it performed slightly better than the dHSIC test. The test based on $\zeta_{\text{max},n}$ also outperformed the dHSIC test in this example.

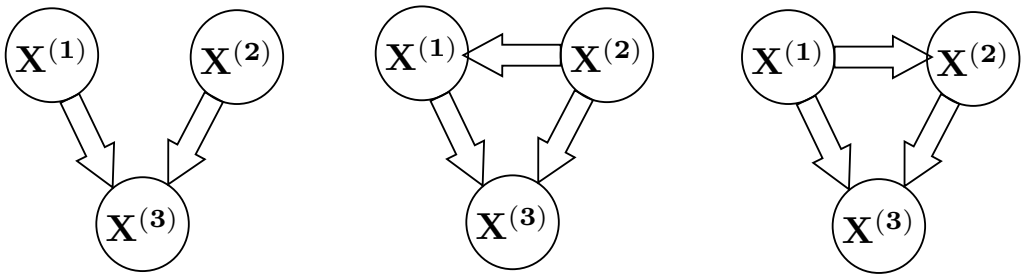
We obtained somewhat similar results when the observations on $\mathbf{X}^{(1)}$ were generated from the standard bivariate t_2 distribution (t distribution with 2 degrees of freedom). In that example also, the single-scale method based on ζ_n, its multi-scale analog based on $\zeta_{\text{sum},n}$ and the HHG test outperformed their competitors. The test based on $\zeta_{\text{max},n}$ and the dHSIC test had almost similar performance. The tests based on linear projections had slightly lower success rates, but they performed better than JdCov and rank-JdCov tests.

Next, we considered an example involving three random variables. In this example, we had two independent random variables $X^{(1)}$ and $X^{(2)}$ from the standard Cauchy distribution, and $X^{(3)} = (X^{(1)}X^{(2)})^3 + \varepsilon$, where $\varepsilon \sim N(0, 0.01)$. We generated 30 observations from the joint distribution of $(X^{(1)}, X^{(2)}, X^{(3)})$ and used different tests to identify the correct model (depicted in Figure 4.10(a)) out of 25 possible SEMs. This experiment was repeated

TABLE 4.1: Proportion of times different methods selected the correct model in the example involving two bivariate random vectors.

	ζ_n	$\zeta_{\text{sum},n}$	$\zeta_{\text{max},n}$	$\widetilde{\zeta}_n$	$\widetilde{\zeta}_{\text{sum},n}$	$\widetilde{\zeta}_{\text{max},n}$	dHSIC	JdCov	r-JdCov	HHG
Normal	0.89	0.87	0.82	0.61	0.63	0.75	0.70	0.47	0.48	0.88
t_2	0.97	0.96	0.89	0.76	0.76	0.78	0.86	0.61	0.57	0.98

100 times as before. Note that in this example, there are two other super models (see Figures 4.10(b) and 4.10(c)), which contain the true SEM. Therefore, instead of choosing the correct model, all these methods for causal inference sometimes selected one of these super models. So, in this example, we counted the proportion of times a method selected one of these three models depicted in Figure 4.10, and they are reported in Table 4.2. Since there are more than two sub-vectors (variables), the HHG test could not be used in this example.



(a) True Model (b) Super Model 1 (c) Super Model 2

FIGURE 4.10: DAGs corresponding to the true model and two super models in the example involving three random variables.

In this example, the multi-scale versions of the proposed tests performed slightly better than their single-scale analogs. Table 4.2 clearly shows the superiority of the tests based on pairwise distances. In this example, the dHSIC test and the JdCov test selected one of these three models in 26% and 34% cases only. The methods based on linear projections had success rates varying between 40% and 50%. The rank-JdCov test also had similar performance. But the multi-scale methods based on pairwise distances successfully selected one of the three models in more than 60% cases. The single scale method based on ζ_n chose one of the these models on 58 out of 100 occasions.

TABLE 4.2: Proportion of times different methods selected the true model and two super models in the example involving three random variables.

	ζ_n	$\zeta_{\text{sum},n}$	$\zeta_{\text{max},n}$	$\widetilde{\zeta}_n$	$\widetilde{\zeta}_{\text{sum},n}$	$\widetilde{\zeta}_{\text{max},n}$	dHSIC	JdCov	r-JdCov
True Model	0.10	0.13	0.14	0.17	0.14	0.12	0.06	0.13	0.16
Super Model 1	0.23	0.22	0.21	0.12	0.13	0.20	0.08	0.11	0.16
Super Model 2	0.25	0.28	0.27	0.12	0.19	0.17	0.12	0.10	0.14
Total	0.58	0.63	0.62	0.41	0.46	0.49	0.26	0.34	0.46

4.8 Proofs and mathematical details

Proof of Theorem 4.1. Since $\mathcal{T}$ satisfies Assumption 4.1, $\mathcal{T}(F^{\mathbf{a}}) \geq 0$ for every $\mathbf{a} \in \mathbb{R}^d$. This implies that $\zeta_{\mathcal{T}}^P(F) \geq 0$. Now, $\zeta_{\mathcal{T}}^P(F) = 0$ holds if and only if $\mathcal{T}(F^{\mathbf{a}}) = 0$ for P-almost every $\mathbf{a}$. So, $\|\mathbf{X}^{(1)} - \mathbf{a}^{(1)}\|, \|\mathbf{X}^{(2)} - \mathbf{a}^{(2)}\|, \ldots, \|\mathbf{X}^{(p)} - \mathbf{a}^{(p)}\|$ are mutually independent for every $\mathbf{a} \in \mathcal{S}$, where $\mathcal{S}$ is the support of P. That is, $\mathcal{L}(\mathbf{X}) = \mathcal{L}(\mathbf{X}^{(1)}) \otimes \mathcal{L}(\mathbf{X}^{(2)}) \otimes \cdots \otimes \mathcal{L}(\mathbf{X}^{(p)})$ on $B(\mathbf{a}^{(1)}, r_1) \times B(\mathbf{a}^{(2)}, r_2) \times \cdots \times B(\mathbf{a}^{(p)}, r_p)$ for all $(\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(p)}) \in \mathcal{S}$ and $r_1, r_2, \ldots, r_p \geq 0$, where $\mathcal{L}$ denotes the law (distribution) of a random vector. Now, for any $\mathbf{a} \in \mathcal{S}$ and $r \geq 0$, the set $B(\mathbf{a}, r)$ can be written as a countable union of sets of the form $B(\mathbf{a}^{(1)}, r_1) \times B(\mathbf{a}^{(2)}, r_2) \times \cdots \times B(\mathbf{a}^{(p)}, r_p)$. So, this in turn implies that $\mathcal{L}(\mathbf{X}) = \mathcal{L}(\mathbf{X}^{(1)}) \otimes \mathcal{L}(\mathbf{X}^{(2)}) \otimes \cdots \otimes \mathcal{L}(\mathbf{X}^{(p)})$ on $B(\mathbf{a}, r)$ for all $\mathbf{a} \in \mathcal{S}$ and $r \geq 0$. Since the Lebesgue measure of $\mathcal{S}$ is positive, following Rawat and Sitaram (2000), this implies that $\mathcal{L}(\mathbf{X}) = \mathcal{L}(\mathbf{X}^{(1)}) \otimes \mathcal{L}(\mathbf{X}^{(2)}) \otimes \cdots \otimes \mathcal{L}(\mathbf{X}^{(p)})$ throughout, i.e., $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are independent. □

Proof of Theorem 4.2. First observe that

$$\begin{aligned}
\left|\zeta_{\mathcal{T}_n}^{P_N}(F) - \zeta_{\mathcal{T}}^P(F)\right| &= \left|\int_{\mathbb{R}^d} \mathcal{T}_n(F^{\mathbf{a}})\, dP_N(\mathbf{a}) - \int_{\mathbb{R}^d} \mathcal{T}(F^{\mathbf{a}})\, dP(\mathbf{a})\right| \\
&\leq \left|\int_{\mathbb{R}^d} \mathcal{T}_n(F^{\mathbf{a}})\, dP_N(\mathbf{a}) - \int_{\mathbb{R}^d} \mathcal{T}(F^{\mathbf{a}})\, dP_N(\mathbf{a})\right| + \left|\int_{\mathbb{R}^d} \mathcal{T}(F^{\mathbf{a}})\, dP_N(\mathbf{a}) - \int_{\mathbb{R}^d} \mathcal{T}(F^{\mathbf{a}})\, dP(\mathbf{a})\right| \\
&\leq \frac{1}{N}\sum_{i=1}^N |\mathcal{T}_n(F^{\mathbf{a}_i}) - \mathcal{T}(F^{\mathbf{a}_i})| + \left|\frac{1}{N}\sum_{i=1}^N \mathcal{T}(F^{\mathbf{a}_i}) - \int_{\mathbb{R}^d} \mathcal{T}(F^{\mathbf{a}})\, dP(\mathbf{a})\right|. \qquad (4.4)
\end{aligned}$$

Now, as $\mathcal{T}(F^{\mathbf{a}_i})$'s are bounded i.i.d. random variables, it follows from the strong law of large numbers that the second term on the right side in Equation (4.4), i.e.

$$\left|\frac{1}{N}\sum_{i=1}^N \mathcal{T}(F^{\mathbf{a}_i}) - \int_{\mathbb{R}^d} \mathcal{T}(F^{\mathbf{a}})\, dP(\mathbf{a})\right| \to 0 \text{ almost surely as } N \to \infty.$$

Also, we have

$$\begin{aligned}
\Pr\left[\frac{1}{N}\sum_{i=1}^N |\mathcal{T}_n(F^{\mathbf{a}_i}) - \mathcal{T}(F^{\mathbf{a}_i})| > \delta\right] &\leq \Pr\left[\max_{1\leq i\leq N} |\mathcal{T}_n(F^{\mathbf{a}_i}) - \mathcal{T}(F^{\mathbf{a}_i})| > \delta\right] \\
&\leq \sum_{i=1}^N \Pr\left[|\mathcal{T}_n(F^{\mathbf{a}_i}) - \mathcal{T}(F^{\mathbf{a}_i})| > \delta\right] \leq N(n)\widetilde{p}_n(\delta, F).
\end{aligned}$$

Since $\delta > 0$ is arbitrary, $\frac{1}{N}\sum_{i=1}^N |\mathcal{T}_n(F^{\mathbf{a}_i}) - \mathcal{T}(F^{\mathbf{a}_i})|$ converges to 0 in probability if $\lim_{n\to\infty} N(n)\widetilde{p}_n(\delta, F) = 0$. If $\sum_{n=1}^\infty N(n)\widetilde{p}_n(\delta, F) < \infty$, the almost sure convergence of $\frac{1}{N}\sum_{i=1}^N |\mathcal{T}_n(F^{\mathbf{a}_i}) - \mathcal{T}(F^{\mathbf{a}_i})|$ to 0 follows from the Borel-Cantelli Lemma. Using these facts, we get the probability convergence and almost sure convergence of $\zeta_{\mathcal{T}_n}^{P_N}(F)$ to $\zeta_{\mathcal{T}}^P(F)$ in the respective cases. □

Lemma 4.1. *For any fixed $\delta > 0$, define $p_{\sigma,n}(\delta, F) = \sup_{\mathbf{a}\in\mathbb{R}^d} \Pr\left[|\mathbb{T}_{\sigma,n}(F^{\mathbf{a}}) - \mathbb{T}_{\sigma}(F^{\mathbf{a}})| > \delta\right]$. If $N(n)$ is a polynomial function of n, then $\sum_{n=1}^{\infty} N(n) p_{\sigma,n}(\delta, F) < \infty$.*

Proof. Let us denote the copula distribution of $F^{\mathbf{a}}$ by $\mathsf{C}^{\mathbf{a}}$ and the empirical copula distribution based on n independent observations from $F^{\mathbf{a}}$ by $\mathsf{C}_n^{\mathbf{a}}$. Thus,

$$\mathbb{T}_{\sigma}(F^{\mathbf{a}}) = \frac{\gamma_{K_\sigma}(\mathsf{C}^{\mathbf{a}}, \Pi)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)} \text{ and } \mathbb{T}_{\sigma,n}(F^{\mathbf{a}}) = \frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n)},$$

where M and Π are maximum and uniform copula as defined in Section 2.1, and M_n and Π_n are their empirical analogs defined in Section 2.2.

$$\begin{aligned}\text{Observe that } \Pr\left[|\mathbb{T}_{\sigma,n}(F^{\mathbf{a}}) - \mathbb{T}_{\sigma}(F^{\mathbf{a}})| > \delta\right] &= \Pr\left[\left|\frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n)} - \frac{\gamma_{K_\sigma}(\mathsf{C}^{\mathbf{a}}, \Pi)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)}\right| > \delta\right] \\ \leq \Pr\left[\left|\frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n)} - \frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)}\right| > \frac{\delta}{2}\right] &+ \Pr\left[\left|\frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)} - \frac{\gamma_{K_\sigma}(\mathsf{C}^{\mathbf{a}}, \Pi)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)}\right| > \frac{\delta}{2}\right]\end{aligned} \tag{4.5}$$

First consider the first term of Equation (4.5)

$$\left|\frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n)} - \frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)}\right| = \frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n)\gamma_{K_\sigma}(\mathsf{M}, \Pi)} \times |\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n) - \gamma_{K_\sigma}(\mathsf{M}, \Pi)|.$$

Note that the term $\left|\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)/[\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n)\gamma_{K_\sigma}(\mathsf{M}, \Pi)]\right|$ is uniformly bounded and the term $|\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n) - \gamma_{K_\sigma}(\mathsf{M}, \Pi)|$ is a non-random quantity that converges to 0 as n tends to infinity (see, Lemma 2.2). Therefore, there exists $n_0 \geq 1$ such that for all $n > n_0$,

$$\left|\frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n)} - \frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)}\right| < \frac{\delta}{2}, \text{ i.e., } \Pr\left[\left|\frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}_n, \Pi_n)} - \frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)}\right| > \frac{\delta}{2}\right] = 0.$$

Note that this n_0 does not depend on $\mathbf{a}$. Now consider the second term of Equation (4.5)

$$\Pr\left[\left|\frac{\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)} - \frac{\gamma_{K_\sigma}(\mathsf{C}^{\mathbf{a}}, \Pi)}{\gamma_{K_\sigma}(\mathsf{M}, \Pi)}\right| > \frac{\delta}{2}\right] = \Pr\left[|\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n) - \gamma_{K_\sigma}(\mathsf{C}, \Pi)| > \delta^*\right],$$

where $\delta^* = \gamma_{K_\sigma}(\mathsf{M}, \Pi)\frac{\delta}{2}$ is a positive constant. Now to prove this lemma, it is enough to show the finiteness of

$$\sum_{n=1}^{\infty} N(n) \sup_{\mathbf{a}\in\mathbb{R}^d} \Pr\left[|\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n) - \gamma_{K_\sigma}(\mathsf{C}^{\mathbf{a}}, \Pi)| > \delta^*\right].$$

For $j = 1, 2, \ldots, p$, let $F^{(\mathbf{a},j)}$ be the distribution function of $X^{(\mathbf{a},j)}$. For $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, denote $\|\mathbf{x}_i^{(j)} - \mathbf{a}_i^{(j)}\|$ as $x_i^{(\mathbf{a},j)}$. Also define $\mathsf{C}_n^{\mathbf{a}*}$ as the empirical joint distribution of $\left(F^{(\mathbf{a},1)}(x_1^{(\mathbf{a},1)}), \ldots, F^{(\mathbf{a},p)}(x_1^{(\mathbf{a},p)})\right), \ldots \left(F^{(\mathbf{a},1)}(x_n^{(\mathbf{a},1)}), \ldots, F^{(\mathbf{a},p)}(x_n^{(\mathbf{a},p)})\right)$. Then, following the proof of Theorem 2.6, we get

$$
\begin{aligned}
&\Pr\left[|\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n) - \gamma_{K_\sigma}(\mathsf{C}^{\mathbf{a}}, \Pi)| > \delta^*\right] \\
&\leq \Pr\left[\left|\gamma^2_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n) - \gamma^2_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi)\right|^{\frac{1}{2}} + \gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \mathsf{C}_n^{\mathbf{a}*}) + \gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}*}, \mathsf{C}^{\mathbf{a}}) > \delta^*\right] \\
&\leq \Pr\left[\left|\gamma^2_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n) - \gamma^2_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi)\right|^{\frac{1}{2}} > \frac{\delta^*}{3}\right] \\
&\qquad + \Pr\left[\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \mathsf{C}_n^{\mathbf{a}*}) > \frac{\delta^*}{3}\right] + \Pr\left[\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}*}, \mathsf{C}^{\mathbf{a}}) > \frac{\delta^*}{3}\right]
\end{aligned}
$$

Now, using Lemma 2.2, we can show that there exists $n_1 \geq 1$ such that for all $n > n_1$, $\Pr\left[\left|\gamma^2_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n) - \gamma^2_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \Pi)\right|^{\frac{1}{2}} > \delta^*/3\right] = 0$. Note that here the choice of n_1 does not depend on $\mathbf{a}$. Again from Lemmas 2.3 and 2.4, we have $\Pr\left[\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}}, \mathsf{C}_n^{\mathbf{a}*}) > \delta^*/3\right] < 2p\exp\left(-\frac{n\delta^{*2}}{18pL^2}\right)$ and $\Pr\left[\gamma_{K_\sigma}(\mathsf{C}_n^{\mathbf{a}*}, \mathsf{C}^{\mathbf{a}}) > \delta^*/3\right] < \exp\left(-\frac{n}{2}\left(\frac{\delta^*}{3} - \frac{2}{\sqrt{n}}\right)^2\right)$, respectively, where $L > 0$ is a constant independent of $\mathbf{a}$. So, to prove the lemma, it is sufficient to show that

$$
\sum_{n=1}^{\infty} N(n)\left[2p\exp\left(-\frac{n\delta^{*2}}{18pL^2}\right) + \exp\left(-\frac{n}{2}\left(\frac{\delta^*}{3} - \frac{2}{\sqrt{n}}\right)^2\right)\right] < \infty.
$$

Clearly, this is true since $N(n)$ is a polynomial function of n. □

Lemma 4.2. *Consider a sequence $\{\sigma_n\}_{n\geq 1}$, which converges to some $\sigma_0 > 0$. If $N(n)$ is a polynomial in n, then for any fixed $\delta > 0$, we have $\sum_{n=1}^{\infty} N(n) p_{\sigma_n, n}(\delta, F) < \infty$.*

Proof. Note that for any fixed $\mathbf{a}$,

$$
\begin{aligned}
&\Pr\left[|\mathbb{T}_{\sigma_n, n}(F^{\mathbf{a}}) - \mathbb{T}_{\sigma_0}(F^{\mathbf{a}})| > \delta\right] \\
&\leq \Pr\left[|\mathbb{T}_{\sigma_n, n}(F^{\mathbf{a}}) - \mathbb{T}_{\sigma_0, n}(F^{\mathbf{a}})| > \frac{\delta}{2}\right] + \Pr\left[|\mathbb{T}_{\sigma_0, n}(F^{\mathbf{a}}) - \mathbb{T}_{\sigma_0}(F^{\mathbf{a}})| > \frac{\delta}{2}\right].
\end{aligned}
$$

So, in view of Lemma 4.1, it is enough to show that

$$
\sum_{n=1}^{\infty} N(n) \sup_{\mathbf{a}\in\mathbb{R}^d} \Pr\left[|\mathbb{T}_{\sigma_n, n}(F^{\mathbf{a}}) - \mathbb{T}_{\sigma_0, n}(F^{\mathbf{a}})| > \delta/2\right]
$$

is finite. Now, note that

$$
\begin{aligned}
|\mathbb{T}_{\sigma_n, n}(F^{\mathbf{a}}) - \mathbb{T}_{\sigma_0, n}(F^{\mathbf{a}})| &= \left|\frac{\gamma_{K_{\sigma_n}}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_{\sigma_n}}(\mathsf{M}_n, \Pi_n)} - \frac{\gamma_{K_{\sigma_0}}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)}{\gamma_{K_{\sigma_0}}(\mathsf{M}_n, \Pi_n)}\right| \\
&\leq \gamma_{K_{\sigma_n}}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n) \times \left|\frac{1}{\gamma_{K_{\sigma_n}}(\mathsf{M}_n, \Pi_n)} - \frac{1}{\gamma_{K_{\sigma_0}}(\mathsf{M}_n, \Pi_n)}\right| \\
&\qquad + \frac{|\gamma_{K_{\sigma_n}}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n) - \gamma_{K_{\sigma_0}}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)|}{\gamma_{K_{\sigma_0}}(\mathsf{M}_n, \Pi_n)} = A_n + B_n, \text{ (say)}.
\end{aligned}
$$

While $\gamma_{K_{\sigma_n}}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)$ is uniformly bounded, $\left|\frac{1}{\gamma_{K_{\sigma_n}}(\mathsf{M}_n, \Pi_n)} - \frac{1}{\gamma_{K_{\sigma_0}}(\mathsf{M}_n, \Pi_n)}\right|$ is a non-random quantity converging to 0 (follows from Lemma 2.6). So, there exists a natural number n_0 (independent of $\mathbf{a}$) such that for all $n > n_0$, we have $A_n \leq \delta/4$ with probability one.

Now, $\gamma_{K_{\sigma_0}}(\mathsf{M}_n, \Pi_n)$ is a non-random quantity converging to $\gamma_{K_{\sigma_0}}(\mathsf{M}, \Pi)$. Also, from Lemma 2.6, we get a non-random upper bound for the term $|\gamma_{K_{\sigma_n}}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n) - \gamma_{K_{\sigma_0}}(\mathsf{C}_n^{\mathbf{a}}, \Pi_n)|$, which converges to 0. Since this upper bound does not depend on $\mathbf{a}$, we get a natural number n_1 (independent of $\mathbf{a}$) such that for all $n > n_1$, $B_n \leq \delta/4$ with probability one.

So, using these two facts, for all $n > \max\{n_0, n_1\}$, we have $A_n + B_n \leq \delta/2$ with probability one, and hence $\Pr\Big[\,|\mathbb{T}_{\sigma_n,n}(F^{\mathbf{a}}) - \mathbb{T}_{\sigma_0,n}(F^{\mathbf{a}})| > \delta/2\Big] = 0$. This implies the finiteness of $\sum_{n=1}^{\infty} N(n) \sup_{\mathbf{a}\in\mathbb{R}^d} \Pr\Big[\,|\mathbb{T}_{\sigma_n,n}(F^{\mathbf{a}}) - \mathbb{T}_{\sigma_0,n}(F^{\mathbf{a}})| > \delta/2\Big]$. □

Proof of Theorem 4.3. Note that while $\mathbb{T}_{\sigma_0}$ is a bounded functional that satisfies Assumption 4.1, following the proof of Theorem 2.7, we have $\mathbb{T}_{\sigma_n,n} \to \mathbb{T}_{\sigma_0}$ almost surely as $n \to \infty$. From Lemma 4.2, we can see that the sequence of estimators $\{\mathbb{T}_{\sigma_n,n}\}_{n\geq 1}$ also satisfies the conditions of Theorem 4.2. Thus, from Theorem 4.2, we get the almost sure convergence of $\zeta_n(F)$ to $\zeta(F)$. Now, from Theorem 4.1, the quantity $\zeta(F)$ is non-negative, and it takes the value 0 if and only if the sub-vectors are mutually independent. The consistency of the right-tailed test based on ζ_n follows from this fact. □

For the rest of this section, we shall assume that $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are random functions in separable Banach Spaces $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_p$, respectively. As before, the joint distribution of $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ will be denoted by F, and for $\mathbf{a}^{(1)} \in \mathcal{B}_1, \mathbf{a}^{(2)} \in \mathcal{B}_2, \ldots, \mathbf{a}^{(p)} \in \mathcal{B}_p$, the joint distribution of $\|\mathbf{X}^{(1)} - \mathbf{a}^{(1)}\|_{\mathcal{B}_1}, \|\mathbf{X}^{(2)} - \mathbf{a}^{(2)}\|_{\mathcal{B}_2}, \ldots, \|\mathbf{X}^{(p)} - \mathbf{a}^{(p)}\|_{\mathcal{B}_p}$ will be denoted by $F^{\mathbf{a}}$, where $\mathbf{a} = (\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(p)})$. We shall assume that P is a probability measure on the product space $\mathcal{B} = \mathcal{B}_1 \times \mathcal{B}_2 \times \cdots \times \mathcal{B}_p$ and the functional $\mathcal{T}$ satisfies Assumption 4.1. As before, we define $\zeta_{\mathcal{T}}^P(F) = \int_{\mathcal{B}} \mathcal{T}(F^{\mathbf{a}})\, dP(\mathbf{a})$.

Theorem 4.4. *Assume that P is a strictly positive measure on $\mathcal{B}$. If $\mathcal{T}(F^{\mathbf{a}})$ is a continuous function of $\mathbf{a}$, then $\zeta_{\mathcal{T}}^P(F) = 0$ if and only if $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are mutually independent.*

Proof of Theorem 4.4. The if part is trivial. So, we prove the only if part here. We claim that $\zeta_{\mathcal{T}}^P(F) = 0$ implies $\mathcal{T}(F^{\mathbf{a}}) = 0$ for all $\mathbf{a} \in \mathcal{B}$. If this is not true, then there exists $\mathbf{a}_0 \in \mathcal{B}$ for which $\mathcal{T}(F^{\mathbf{a}_0}) > 0$. Now, from the assumption of continuity of $\mathcal{T}(F^{\mathbf{a}})$, we get an open neighborhood $\mathcal{N}_{\mathbf{a}_0}$ of $\mathbf{a}_0$ such that $\mathcal{T}(F^{\mathbf{a}}) > 0$ for all $\mathbf{a} \in \mathcal{N}_{\mathbf{a}_0}$. As P is a strictly positive measure, this leads to a contradiction to the fact that $\zeta_{\mathcal{T}}^P(F) = 0$. So, we have mutual independence of $\|\mathbf{X}^{(1)} - \mathbf{a}^{(1)}\|_{\mathcal{B}_1}, \|\mathbf{X}^{(2)} - \mathbf{a}^{(2)}\|_{\mathcal{B}_2}, \ldots, \|\mathbf{X}^{(p)} - \mathbf{a}^{(p)}\|_{\mathcal{B}_p}$ for all $\mathbf{a}^{(1)} \in \mathcal{B}_1, \mathbf{a}^{(2)} \in \mathcal{B}_2, \ldots, \mathbf{a}^{(p)} \in \mathcal{B}_p$. Hence it follows that $\Pr\big[\|\mathbf{X}^{(i)} - \mathbf{a}^{(i)}\|_{\mathcal{B}_i} < r_i,\ \forall\ 1 \leq i \leq p\big] = \prod_{i=1}^{p} \Pr\big[\|\mathbf{X}^{(i)} - \mathbf{a}^{(i)}\|_{\mathcal{B}_i} < r_i\big]$, and consequently $\mathcal{L}(\mathbf{X}) = \mathcal{L}(\mathbf{X}^{(1)}) \otimes \mathcal{L}(\mathbf{X}^{(2)}) \otimes \cdots \otimes \mathcal{L}(\mathbf{X}^{(p)})$

on $B(\mathbf{a}^{(1)}, r_1) \times B(\mathbf{a}^{(2)}, r_2) \times \cdots \times B(\mathbf{a}^{(p)}, r_p)$ for all $\mathbf{a}^{(1)} \in \mathcal{B}_1, \mathbf{a}^{(2)} \in \mathcal{B}_2, \ldots, \mathbf{a}^{(p)} \in \mathcal{B}_p$ and $r_1, r_2, \ldots, r_p \geq 0$, where $\mathcal{L}$ denotes the law (distribution) of a random function. Now, from separability of the Banach spaces $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_p$, we conclude that $\mathcal{L}(\mathbf{X}) = \mathcal{L}(\mathbf{X}^{(1)}) \otimes \mathcal{L}(\mathbf{X}^{(2)}) \otimes \cdots \otimes \mathcal{L}(\mathbf{X}^{(p)})$ holds for all $\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_p$, where $\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_p$ are measurable sets belonging to $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_p$, respectively. This concludes that $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are mutually independent. □

Theorem 4.5. *Assume that* $\mathbf{X}$ *is a random element in the Banach space* $\mathcal{B}$ *and* F *is a strictly positive probability measure on* $\mathcal{B}$*. Also assume that* $F^{\mathbf{a}}$ *has continuous univariate marginals for all* $\mathbf{a} \in \mathcal{B}$ *and the functional* $\mathcal{T}$ *satisfies the assumption of Theorem 4.4. Then, the power of the test based on* ζ_n*, converges to* 1 *as the sample size* n *tends to infinity.*

Proof of Theorem 4.5. Note that while $\mathbb{T}_{\sigma_0}$ is a bounded functional that satisfies Assumption 4.1, following the proof of Theorem 2.7, we have $\mathbb{T}_{\sigma_n,n} \to \mathbb{T}_{\sigma_0}$ almost surely as $n \to \infty$. Using the same line of reasoning as in Lemma 4.2, we can show that $\sum_{n=1}^{\infty} n \cdot \sup_{\mathbf{a} \in \mathcal{B}} \Pr\left[|\mathbb{T}_{\sigma_n,n}(F^{\mathbf{a}}) - \mathbb{T}_{\sigma_0}(F^{\mathbf{a}})| > \delta\right] < \infty$. With this fact and again using similar arguments as in Theorem 4.2, we can prove that $\zeta_n(F)$ to $\zeta(F)$ almost surely. From Theorem 4.4, we get that the quantity $\zeta(F)$ is non-negative and takes the value 0 if and only if the sub-vectors are mutually independent. The consistency of the right-tailed test based on ζ_n follows from this fact. □

Chapter 5

Test of Independence among Random Vectors: Methods Based on Ranks of Nearest Neighbors

Friedman and Rafsky (1983) developed some graph based methods for testing independence between two random vectors of arbitrary dimensions. Following their ideas, Heller *et al.* (2012) constructed a distribution-free test based on random traversal of the edges of the minimum spanning tree (MST). Instead of random traversal, Biswas *et al.* (2016) used a systematic traversal of the edges of the MST following Prim's algorithm (Prim, 1957) to construct some modified distribution-free tests with better power properties. Later Sarkar and Ghosh (2018) pointed out some limitations of these distribution-free tests. In particular, they showed that in order to possess the distribution-free property, these tests sacrifice a lot of information and use the information contained in only $(n-1)$ edges of the minimum spanning tree out of $\binom{n}{2}$ edges present in the complete graph. In order to take care of this problem, they developed some tests based on ranks of nearest neighbors and demonstrated their superiority over dCov (Székely *et al.*, 2007), HHG (Heller *et al.*, 2013) and HSIC (Gretton *et al.*, 2008) tests in a large class of examples. In this chapter, we discuss about some possible generalizations of their tests for more than two random vectors of arbitrary dimensions.

Before describing our proposed methods, let us first revisit the tests proposed by Sarkar and Ghosh (2018). Let $\mathbf{x}_1 = (\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}), \mathbf{x}_2 = (\mathbf{x}_2^{(1)}, \mathbf{x}_2^{(2)}), \ldots, \mathbf{x}_n = (\mathbf{x}_n^{(1)}, \mathbf{x}_n^{(2)})$ be n independent observations on the random vector $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$, where $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are sub-vectors of dimensions d_1 and d_2, respectively. For testing independence between $\mathbf{X}^{(1)}$

and $\mathbf{X}^{(2)}$, Sarkar and Ghosh (2018) used the ranks of nearest neighbors in the following way. For each $i = 1, 2, \ldots, n$, let $\|\mathbf{x}_i^{(1)} - \mathbf{x}_{i_1}^{(1)}\| \leq \|\mathbf{x}_i^{(1)} - \mathbf{x}_{i_2}^{(1)}\| \leq \cdots \leq \|\mathbf{x}_i^{(1)} - \mathbf{x}_{i_{n-1}}^{(1)}\|$ be the ordered distances of the $\mathbf{x}_j^{(1)}$'s $(j \neq i)$ from $\mathbf{x}_i^{(1)}$. For $k = 1, 2, \ldots, (n-1)$, Sarkar and Ghosh (2018) called $\mathbf{x}_{i_k}$ as the k-th nearest $\mathbf{X}^{(1)}$-neighbor of $\mathbf{x}_i$ and computed the rank of the corresponding $\mathbf{X}^{(2)}$-distance $\|\mathbf{x}_i^{(2)} - \mathbf{x}_{i_k}^{(2)}\|$ among $(n-k)$ distances $\left\{\|\mathbf{x}_i^{(2)} - \mathbf{x}_{i_k}^{(2)}\|, \|\mathbf{x}_i^{(2)} - \mathbf{x}_{i_{k+1}}^{(2)}\|, \ldots, \|\mathbf{x}_i^{(2)} - \mathbf{x}_{i_{n-1}}^{(2)}\|\right\}$. They called it the $\mathbf{X}^{(2)}$-rank of the k-th nearest $\mathbf{X}^{(1)}$-neighbor of $\mathbf{x}_i$. We denote this rank by $R^{(2|1)}(i,k)$. For each $i = 1, 2, \ldots, n$, these ranks $R^{(2|1)}(i,1), R^{(2|1)}(i,2), \ldots, R^{(2|1)}(i,n-1)$ are independent (see, e.g. Heller *et al.*, 2012). Moreover, when $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent, $R^{(2|1)}(i,k)$ follows a discrete uniform distribution with mass points $\{1, 2, \ldots, n-k\}$ (note that $R^{(2|1)}(i,n-1)$ is degenerate at 1). Sarkar and Ghosh (2018) argued that the dependence between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ results in extremely small or large values of these ranks. This type of idea is quite common in the literature (see, e.g., Friedman and Rafsky, 1983; Heller *et al.*, 2012; Biswas *et al.*, 2016). Motivated by this idea, Sarkar and Ghosh (2018) proposed to use the statistic

$$T^{(2|1)} = \max\left\{-2\sum_{i=1}^{n}\sum_{k=1}^{n-2}\varphi\Big(\frac{R^{(2|1)}(i,k)}{n-k}\Big), -2\sum_{i=1}^{n}\sum_{k=1}^{n-2}\varphi\Big(\frac{R_r^{(2|1)}(i,k)}{n-k}\Big)\right\},$$

to measure the deviation from independence. Here, φ is a suitable monotone function on $(0,1]$ and $R_r^{(2|1)}(i,k) = (n-k+1) - R^{(2|1)}(i,k)$ is the reverse rank. They also computed $T^{(1|2)}$, where the roles of $\mathbf{X}^1$ and $\mathbf{X}^{(2)}$ are reversed, and finally they used a symmetric combination (e.g., sum or maximum) of $T^{(1|2)}$ and $T^{(2|1)}$)as the test statistic. They suggested to reject $\mathbb{H}_0$, the null hypothesis of independence, for large values of this statistic.

Here, we have n independent observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ on $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)})$, where $p \geq 2$, and the q-th $(q = 1, 2, \ldots, p)$ sub-vector $\mathbf{X}^{(q)}$ takes values in the d_q-dimensional Euclidean space. To test for mutual independence among more than two sub-vectors, we propose some generalizations of Sarkar and Ghosh (2018), which are described in the following sections.

5.1 Tests based on univariate ranks of a group of sub-vectors

First note that $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are jointly independent if and only if $\mathbf{X}^{(q)}$ and $\mathbf{X}^{(-q)}$ are independent for every $q = 1, 2, \ldots, p$, where $\mathbf{X}^{(-q)} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(q-1)}, \mathbf{X}^{(q+1)}, \ldots, \mathbf{X}^{(p)})$

is the collection of $(p-1)$ sub-vectors excluding $\mathbf{X}^{(q)}$. The proof is easy, but we add the result and the sake for completeness.

Lemma 5.1. *Let $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ be random vectors. For $q = 1, 2, \ldots, p$, let $\mathbf{X}^{(-q)}$ be the collection of $(p-1)$ random vectors excluding $\mathbf{X}^{(q)}$. Then, $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are mutually independent if and only if $\mathbf{X}^{(q)}$ and $\mathbf{X}^{(-q)}$ are independent for each $q = 1, 2, \ldots, p$.*

This suggests us to perform pairwise tests of independence between $\mathbf{X}^{(q)}$ and $\mathbf{X}^{(-q)}$ for every $q = 1, 2, \ldots, p$ and aggregate these results. For instance, we can use test statistics similar to those proposed by Sarkar and Ghosh (2018). In particular, for each $q = 1, 2, \ldots, p$, and every $i = 1, 2, \ldots, n$, we define $\mathbf{x}_{i_k}$ as the k-th nearest $\mathbf{X}^{(q)}$-neighbor of $\mathbf{x}_i$ if $\|\mathbf{x}_i^{(q)} - \mathbf{x}_{i_1}^{(q)}\| \leq \|\mathbf{x}_i^{(q)} - \mathbf{x}_{i_2}^{(q)}\| \leq \cdots \leq \|\mathbf{x}_i^{(q)} - \mathbf{x}_{i_{n-1}}^{(q)}\|$, for $k = 1, 2, \ldots, (n-1)$. Similarly, we define $R_U^{(q)}(i,k) = R_U^{(-q|q)}(i,k)$ as the rank of $\|\mathbf{x}_i^{(-q)} - \mathbf{x}_{i_k}^{(-q)}\|$ among $(n-k)$ distances $\{\|\mathbf{x}_i^{(-q)} - \mathbf{x}_{i_k}^{(-q)}\|, \|\mathbf{x}_i^{(-q)} - \mathbf{x}_{i_{k+1}}^{(-q)}\|, \ldots, \|\mathbf{x}_i^{(-q)} - \mathbf{x}_{i_{n-1}}^{(-q)}\|\}$. We call it $\mathbf{X}^{(-q)}$-rank of the k-th nearest $\mathbf{X}^{(q)}$-neighbor of $\mathbf{x}_i$. Note that for each q and i, $R_U^{(q)}(i,1), R_U^{(q)}(i,2), \ldots, R_U^{(q)}(i,n-1)$ are independent, while under $\mathbb{H}_0$, $R_U^{(q)}(i,k)$ follows a discrete uniform distribution with mass points $\{1, 2, \ldots, n-k\}$ (see, e.g., Heller *et al.*, 2012; Biswas *et al.*, 2016). On the other hand, following Sarkar and Ghosh (2018), under dependence, we expect $R_U^{(q)}(i,k)$ to be close to its extreme values, or in other words, the magnitude of $\left\{2R_U^{(q)}(i,k) - (n-k+1)\right\}$ is expected to be large. Note that $\left\{2R_U^{(q)}(i,k) - (n-k+1)\right\}/(n-k)$ takes values in $(-1,1)$. So, if its value is close to 1 or -1, that gives us signal against $\mathbb{H}_0$. To magnify this signal, we use a transformation $\varphi : (-1,1) \to (-\infty, \infty)$ which is a strictly increasing, odd function. Partially motivated by the works of Heller *et al.* (2012); Biswas *et al.* (2016) and Sarkar and Ghosh (2018), throughout this chapter, we use the function $\varphi(t) = -\text{sign}(t)\log(1-|t|)$. Note that such a transformation leads to the same magnification for positive and negative

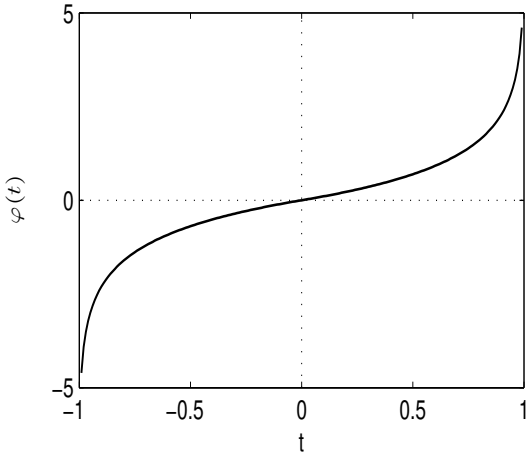


FIGURE 5.1: Transformation function.

values of t, without affecting its sign (see Figure 5.1). So, without using reverse ranks, we can simply consider the statistic

$$\Psi_U^{(q)} = \sum_{i=1}^{n} \sum_{k=1}^{n-2} \varphi\left(\frac{2R_U^{(q)}(i,k) - (n-k+1)}{n-k} \right).$$

Under the null hypothesis of independence, for each $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, (n-2)$, $\left\{ 2R_U^{(q)}(i,k) - (n-k+1) \right\} / (n-k)$ is symmetric around the origin. So, $|\Psi_U^{(q)}|$ is expected to be small. On the other hand, based on our previous discussions, one can expect $|\Psi_U^{(q)}|$ to take large values when the sub-vectors are dependent. However, $|\Psi_U^{(q)}|$ is not symmetric in $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$, and different choices of q may lead to different results. Therefore, we compute $|\Psi_U^{(q)}|$ for each $q = 1, 2, \ldots, p$ and use a symmetric function of $\Psi_U^{(1)}, \Psi_U^{(2)}, \ldots, \Psi_U^{(p)}$ as the test statistic. In particular, we consider the test statistic

$$T_{\text{sum},n}^U := \sum_{1 \le q \le p} |\Psi_U^{(q)}| \quad \text{or} \quad T_{\max,n}^U := \max_{1 \le q \le p} |\Psi_U^{(q)}|$$

and reject $\mathbb{H}_0$ for large values of it. The cut-off is computed using the permutation principle. One can notice that for $p = 2$, this test is almost equivalent to the test proposed in Sarkar and Ghosh (2018).

5.2 Tests based on multivariate ranks

Note that in the previous approach, for testing independence between $\mathbf{X}^{(q)}$ and $\mathbf{X}^{(-q)}$, we considered $\mathbf{X}^{(-q)}$ as a single vector and completely ignored the fact that it consists of several sub-vectors. This may sometimes lead to loss of information regarding the mutual dependence structure and result in loss of power. To take care of this issue, now we adopt another approach. As before, for a fixed q and i, let $\mathbf{x}_{i_k}$ be the k-th nearest $\mathbf{X}^{(q)}$-neighbor of $\mathbf{x}_i$, i.e., $\|\mathbf{x}_i^{(q)} - \mathbf{x}_{i_1}^{(q)}\| \le \|\mathbf{x}_i^{(q)} - \mathbf{x}_{i_2}^{(q)}\| \le \cdots \le \|\mathbf{x}_i^{(q)} - \mathbf{x}_{i_{n-1}}^{(q)}\|$. Now, for each $k = 1, 2, \ldots, (n-1)$, we compute the $(p-1)$-dimensional distance vector

$$\mathbf{D}^{(q)}(i,k) = \left(\|\mathbf{x}_i^{(1)} - \mathbf{x}_{i_k}^{(1)}\|, \ldots, \|\mathbf{x}_i^{(q-1)} - \mathbf{x}_{i_k}^{(q-1)}\|, \|\mathbf{x}_i^{(q+1)} - \mathbf{x}_{i_k}^{(q+1)}\|, \ldots, \|\mathbf{x}_i^{(p)} - \mathbf{x}_{i_k}^{(p)}\| \right)$$

and find its multivariate rank in the set $\left\{ \mathbf{D}^{(q)}(i,k), \mathbf{D}^{(q)}(i,k+1), \ldots, \mathbf{D}^{(q)}(i,n-1) \right\}$ of $(n-k)$ distance vectors. We measure the deviations of these ranks from the ones expected under the null hypothesis. This procedure is repeated for $i = 1, 2, \ldots, n$ to get a statistic indexed by q. We aggregate these statistics over $q = 1, 2, \ldots, p$ to get our final test statistic.

Note that here we are looking for ranks of $(p-1)$-dimensional random vectors. Such ranks are not uniquely defined. For $p > 2$, there are several ways to define multivariate ranks, and the outcome of the proposed test may depend on it. In this article, we use two popular rank functions, viz., coordinate-wise rank (see, e.g., Sen and Puri, 1971) and spatial rank (see, e.g., Taskinen *et al.*, 2005) for the construction of our tests. Detailed description of our tests based on these two choices of multivariate ranks are given below.

5.2.1 Tests based on coordinate-wise ranks

For a collection $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ of d-dimensional observations, for $i = 1, 2, \ldots, n$, the coordinate-wise rank of $\mathbf{z}_i = (z_i^{(1)}, z_i^{(2)}, \ldots, z_i^{(d)})$ is defined as

$$\mathbf{R}_C(i) = \left(r^{(1)}(i), r^{(2)}(i), \ldots, r^{(d)}(i)\right),$$

where $r^{(j)}(i)$ $(j = 1, 2, \ldots, d)$ is the (univariate) rank of $z_i^{(j)}$ among $\{z_1^{(j)}, z_2^{(j)}, \ldots, z_n^{(j)}\}$. In our present context, for a fixed $q = 1, 2, \ldots, p$ and a fixed $i = 1, 2, \ldots, n$, we compute the coordinate-wise rank of the $(p-1)$-dimensional distance vector $\mathbf{D}^{(q)}(i,k)$ with respect to the data cloud consisting of $(n-k)$ vectors $\mathbf{D}^{(q)}(i,k), \mathbf{D}^{(q)}(i,k+1), \ldots, \mathbf{D}^{(q)}(i,n-1)$, and denote it by $\mathbf{R}_C^{(q)}(i,k)$. Note that here $\mathbf{R}_C^{(q)}(i,k)$ is given by

$$\mathbf{R}_C^{(q)}(i,k) = \left(R^{(1|q)}(i,k), \ldots, R^{(q-1|q)}(i,k), R^{(q+1|q)}(i,k), \ldots, R^{(p|q)}(i,k)\right),$$

where, $R^{(s|q)}(i,k)$ is the $\mathbf{X}^{(s)}$-rank $(s = 1, 2, \ldots, p,\ s \neq q)$ of the k-th nearest neighbor of $\mathbf{X}^{(q)}$, i.e., the rank of $\|\mathbf{x}_i^{(s)} - \mathbf{x}_{i_k}^{(s)}\|$ among $(n-k)$ distances $\left\{\|\mathbf{x}_i^{(s)} - \mathbf{x}_{i_k}^{(s)}\|, \|\mathbf{x}_i^{(s)} - \mathbf{x}_{i_{k+1}}^{(s)}\|, \ldots, \|\mathbf{x}_i^{(s)} - \mathbf{x}_{i_{n-1}}^{(s)}\|\right\}$.

From our previous discussion, it follows that if $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are mutually independent, then so are $\mathbf{R}_C^{(q)}(i,1), \mathbf{R}_C^{(q)}(i,2), \ldots, \mathbf{R}_C^{(q)}(i,n-1)$, and for $k = 1, 2, \ldots, (n-2)$, $\mathbf{R}_C^{(q)}(i,k)$ follows a discrete uniform distribution on $\{1, 2, \ldots, n-k\}^{p-1}$, irrespective of the distribution of $\mathbf{X}$. So, under $\mathbb{H}_0$, $\left\{2\mathbf{R}_C^{(q)}(i,k) - (n-k+1)\mathbf{1}_{p-1}\right\}/(n-k)$ follows a discrete uniform distribution with mass points symmetrically distributed around the origin and taking values in the interior of the hypercube $[-1,1]^{p-1}$ (here, $\mathbf{1}_{p-1}$ denotes the $(p-1)$-dimensional vector having all elements equal to 1). On the other hand, under dependence, we expect the magnitude of $\left\{2R^{(s|q)}(i,k) - (n-k+1)\right\}/(n-k)$ to be large. As before, we use the transformation $\varphi(t) = -\text{sign}(t)\log(1-|t|)$ on the coordinates of $\left\{2\mathbf{R}_C^{(q)}(i,k) - (n-k+1)\mathbf{1}_{p-1}\right\}/(n-k)$, and define the transformed coordinate-wise ranks

$$\boldsymbol{\psi}_C^{(q)}(i,k) = \left(\widetilde{R}^{(1|q)}(i,k), \ldots, \widetilde{R}^{(q-1|q)}(i,k), \widetilde{R}^{(q+1|q)}(i,k), \ldots, \widetilde{R}^{(p|q)}(i,k)\right),$$
$$\text{where } \widetilde{R}^{(s|q)}(i,k) = \varphi\left(\tfrac{2R^{(s|q)}(i,k)-(n-k+1)}{n-k}\right) \text{ for } s = 1,2,\ldots,p,\ s \neq q.$$

These transformed ranks contain information regarding departure of the actual ranks from uniformity. To combine the information, we compute $\boldsymbol{\psi}_C^{(q)}(i,k)$ for $i = 1,2,\ldots,n$, $k = 1,2,\ldots,(n-2)$, and define $\boldsymbol{\Psi}_C^{(q)} = \sum_{i=1}^{n}\sum_{k=1}^{n-2}\boldsymbol{\psi}_C^{(q)}(i,k)$. If $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are mutually independent, for each $i = 1,2,\ldots,n$ and $k = 1,2,\ldots,(n-2)$, $\boldsymbol{\psi}_C^{(q)}(i,k)$ is symmetric about the origin. So, under $\mathbb{H}_0$, $\|\boldsymbol{\Psi}_C^{(q)}\|$ is expected to be small. On the other hand, our previous discussions suggest that $\|\boldsymbol{\Psi}_C^{(q)}\|$ should take large values when the sub-vectors are dependent. So, as before, we compute $\|\boldsymbol{\Psi}_C^{(q)}\|$ for $q = 1,2,\ldots,p$ and finally reject $\mathbb{H}_0$ for large values of

$$T^C_{\text{sum},n} := \sum_{1\le q\le p}\|\boldsymbol{\Psi}_C^{(q)}\| \quad \text{or} \quad T^C_{\text{max},n} := \max_{1\le q\le p}\|\boldsymbol{\Psi}_C^{(q)}\|.$$

The cut-off value for this test is computed using the permutation principle as before.

5.2.2 Tests based on spatial ranks

The spatial rank of a d-dimensional observation $\mathbf{z}_i$, for $i = 1,2,\ldots,n$, with respect to the data cloud $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ is defined as

$$\mathbf{R}_S(i) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{Sign}_d(\mathbf{z}_i - \mathbf{z}_j), \text{ where } \mathbf{Sign}_d(\mathbf{t}) = \begin{cases}\mathbf{0} & \text{if } \mathbf{t} = \mathbf{0}\\ \frac{\mathbf{t}}{\|\mathbf{t}\|} & \text{if } \mathbf{t} \in \mathbb{R}^d \setminus \{\mathbf{0}\}\end{cases}.$$

Here, $\mathbf{Sign}_d(\cdot)$ is the d-dimensional spatial sign function (see, e.g., Taskinen *et al.*, 2005), which coincides with the usual sign function in one dimension. To construct our tests based on spatial ranks, for a fixed $q = 1,2,\ldots,p$ and $i = 1,2,\ldots,n$, we compute the spatial rank of the $(p-1)$-dimensional vector $\mathbf{D}^{(q)}(i,k)$ with respect to the data cloud $\left\{\mathbf{D}^{(q)}(i,k), \mathbf{D}^{(q)}(i,k+1), \ldots, \mathbf{D}^{(q)}(i,n-1)\right\}$ consisting of $(n-k)$ observations, and call it $\mathbf{R}_S^{(q)}(i,k)$ for $k = 1,2,\ldots,(n-2)$. Note that the expression for $\mathbf{R}_S^{(q)}(i,k)$ is given by

$$\mathbf{R}_S^{(q)}(i,k) = \frac{1}{n-k}\sum_{k'=k}^{n-1}\mathbf{Sign}_{p-1}\left(\mathbf{D}^{(q)}(i,k) - \mathbf{D}^{(q)}(i,k')\right).$$

Here also, we have non-degenerate random vectors $\mathbf{R}_S^{(q)}(i,1), \mathbf{R}_S^{(q)}(i,2), \ldots, \mathbf{R}_S^{(q)}(i,n-2)$, which take values in $B(\mathbf{0}_{p-1}, 1)$, the $(p-1)$-dimensional open ball of radius one. Unlike coordinate-wise ranks, these spatial ranks do not have the distribution-free property under $\mathbb{H}_0$. But, when $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are independent, their distributions turn out to be

symmetric about the origin. Thus, here also, under the alternative, we expect the magnitudes of the $\mathbf{R}_S^{(q)}(i,k)$'s to be large. Since these magnitudes are bounded by 1, we use a transformation similar to that used before to get

$$\boldsymbol{\psi}_S^{(q)}(i,k) = -\mathbf{Sign}_{p-1}\left(\mathbf{R}_S^{(q)}(i,k)\right)\log\left(1-\|\mathbf{R}_S^{(q)}(i,k)\|\right),$$

for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, (n-2)$. Note that this transformation does not change the direction of the spatial rank function, and under $\mathbb{H}_0$, the distribution of $\boldsymbol{\psi}_S^{(q)}(i,k)$ remains symmetric about the origin. Here also, we define $\boldsymbol{\Psi}_S^{(q)} = \sum_{i=1}^{n}\sum_{k=1}^{n-2}\boldsymbol{\psi}_S^{(q)}(i,i_k)$ and use

$$T^S_{\text{sum},n} := \sum_{1\le q\le p}\|\boldsymbol{\Psi}_S^{(q)}\| \quad \text{or} \quad T^S_{\max,n} := \max_{1\le q\le p}\|\boldsymbol{\Psi}_S^{(q)}\|$$

as the test statistic. The null hypothesis $\mathbb{H}_0$ is rejected when for large values of the test statistic. As before, we determine the cut-off value using the permutation method.

5.3 Tests based on maximum mean discrepancy

In Subsection 5.2.1, we have seen that if $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are independent, for any fixed $q = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$, $\mathbf{R}_C^{(q)}(i,k)$ follows a discrete uniform distribution on $\{1, 2, \ldots, n-k\}^{p-1}$, which we denote by $\mathbb{U}_{n-k}^{p-1}$. Non-uniform distribution of the coordinates of $\mathbf{R}_C^{(q)}(i,k)$ or the dependence among them indicates a dependence among the sub-vectors $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$. In other words, if the distribution of $\mathbf{R}_C^{(q)}(i,k)$ (denote it by $\mathbb{F}_k^{(q)}$) deviates from $\mathbb{U}_{n-k}^{p-1}$, we get a signal against the null hypothesis. To measure this deviation, we considered the discrepancy measure called MMD (see, Chapter 2, for more details). Recall that MMD between two probability distributions P and Q is given by

$$\gamma_{K_\sigma}(P,Q) = [\mathrm{E}K(\mathbf{Y},\mathbf{Y}_*) - 2\mathrm{E}K(\mathbf{Y},\mathbf{Z}) + \mathrm{E}K(\mathbf{Z},\mathbf{Z}_*)]^{\frac{1}{2}},$$

where $\mathbf{Y}, \mathbf{Y}_* \overset{i.i.d.}{\sim} P$, $\mathbf{Z}, \mathbf{Z}_* \overset{i.i.d.}{\sim} Q$ are four independent random vectors, and $K_\sigma(\mathbf{x},\mathbf{y}) = \exp\left(-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2\right)$ is the Gaussian kernel. Since γ_{K_σ} is a metric (see, e.g., Sriperumbudur *et al.*, 2010, for more details), we have $\gamma_{K_\sigma}(P,Q) \ge 0$, where the equality holds if and only if $P = Q$. Like before, here also we use median heuristic to choose the bandwidth σ (see, e.g., Gretton *et al.*, 2008). Putting $P = \mathbb{F}_k^{(q)}$ and $Q = \mathbb{U}_{n-k}^{p-1}$ in the expression of $\gamma_{K_\sigma}(P,Q)$, we get the following expression for MMD given in Lemma 5.2.

Lemma 5.2. *If* $\mathbf{V} = (V^{(1)}, V^{(2)}, \ldots, V^{(p-1)})$ *and* $\mathbf{V}_* = (V_*^{(1)}, V_*^{(2)}, \ldots, V_*^{(p-1)})$ *follow the distribution* $\mathbb{F}_k^{(q)}$*, we have* $\gamma_{K_\sigma}(\mathbb{F}_k^{(q)}, \mathbb{U}_{n-k}^{p-1}) = (S_1 - 2S_2 + S_3)^{1/2}$*, where* $S_1 = \mathrm{E}K_\sigma(\mathbf{V}, \mathbf{V}_*)$*,*$S_2 = \mathrm{E}\left[\prod_{t=1}^{p-1} \frac{1}{n-k} \sum_{\ell=1}^{n-k} e^{-\frac{\left(V^{(t)}-\ell\right)^2}{2\sigma^2}}\right]$ *and* $S_3 = \left[\frac{2}{(n-k)^2} \sum_{\ell=1}^{n-k-1} (n-k-\ell)\, e^{-\frac{\ell^2}{2\sigma^2}} + \frac{1}{n-k}\right]^{p-1}$.

For a fixed q and k, we compute $\mathbf{R}_C^{(q)}(i,k)$ for $i = 1, 2, \ldots, n$. Clearly, $\mathbf{R}_C^{(q)}(1,k)$, $\mathbf{R}_C^{(q)}(2,k), \ldots,$ $\mathbf{R}_C^{(q)}(n,k)$ are identically distributed with the distribution $\mathbb{F}_k^{(q)}$. So, we estimate S_1 and S_2 by

$$\hat{S}_1 = \frac{1}{n^2} \sum_{1 \le i,j \le n} K_\sigma\Big(\mathbf{R}_C^{(q)}(i,k), \mathbf{R}_C^{(q)}(j,k)\Big) \text{ and}$$

$$\hat{S}_2 = \frac{1}{n} \sum_{i=1}^{n} \prod_{\substack{t=1 \\ t \neq q}}^{p} \frac{1}{n-k} \sum_{\ell=1}^{n-k} \exp\left\{-\frac{1}{2\sigma^2}\left(r^{(t|q)}(i,k) - \ell\right)^2\right\},$$

respectively. Thus, we get an estimate of $\gamma_{K_\sigma}(\mathbb{F}_k^{(q)}, \mathbb{U}_{n-k}^{p-1})$, which is given by $\mathbb{M}^{(q)}(k) = (\hat{S}_1 - 2\hat{S}_2 + S_3)^{1/2}$. Therefore, higher values of $\mathbb{M}^{(q)}(k)$ indicate dependence among $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$. We can compute this statistic for all values of $k = 1, 2, \ldots, (n-2)$ (note that $\mathbf{R}_C^{(q)}(i, n-1)$ has a degenerate distribution, and hence we ignore it) to come up with an aggregated measure $\mathbf{\Psi}_M^{(q)} = \sum_{k=1}^{n-2} \mathbb{M}^{(q)}(k)$. One can compute $\mathbf{\Psi}_M^{(q)}$ for $q = 1, 2, \ldots, p$, and use a suitable function of $\mathbf{\Psi}_M^{(1)}, \mathbf{\Psi}_M^{(2)}, \ldots, \mathbf{\Psi}_M^{(p)}$ as the test statistic. We consider two such test statistics $T_{\text{sum},n}^M = \sum_{q=1}^{p} \mathbf{\Psi}_M^{(q)}$ and $T_{\max,n}^M = \max_{1 \le q \le p} \mathbf{\Psi}_M^{(q)}$ in this article and reject the null hypothesis for large values of these statistics. The cut-off is chosen using the permutation method as before.

5.4 Results from the analysis of simulated data sets

We analyzed some simulated data sets to compare the performance of our proposed tests (based on $T_{\text{sum},n}^U$ $T_{\max,n}^U$, $T_{\text{sum},n}^C$, $T_{\max,n}^C$, $T_{\text{sum},n}^S$, $T_{\max,n}^S$, $T_{\text{sum},n}^M$ and $T_{\max,n}^M$) with other state of the art tests. In particular, we used the dHSIC test (Pfister *et al.*, 2018), the JdCov test (Chakraborty and Zhang, 2019) and the rank-JdCov test for comparison. Description of all these tests has already been given in the previous chapters. For our proposed tests, we created two R packages 'INN' and 'INNMMD' containing all necessary codes. These packages are available at `https://github.com/angshumanroycode/INN` and `https://github.com/angshumanroycode/INNMMD` respectively. As before, all tests are considered to have 5% nominal level, and the cut-offs are computed based on 1000 random permutations. We repeated each experiment 1000 times, and the power of a test was

estimated by the proportion of times it rejected the null hypothesis. For $p = 2$, since the tests proposed in Section 5.2.1 and 5.2.2 are similar to those proposed in Sarkar and Ghosh (2018), in this section, we did not use any example involving two sub-vectors.

We considered eight examples, each involving four random vectors $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}, \mathbf{X}^{(4)}$. In all these examples, observations on $\mathbf{X}$ were generated from a 20-dimensional distribution. The first five variables formed the sub-vector $\mathbf{X}^{(1)}$, the next five formed $\mathbf{X}^{(2)}$, the next five formed $\mathbf{X}^{(3)}$, and finally the last five variables formed the sub-vector $\mathbf{X}^{(4)}$. Therefore, we had $p = 4$ and $d_1 = d_2 = d_3 = d_4 = 5$. The first six of these eight examples (Normal, t_5, Mixture Normal-1, Mixture Normal-2, Hypersphere and L_1 Ball) are taken from Section 4.2 (see Figure 4.2).

In the 'Normal' example, all sub-vectors were independent, and this example was used to check the level properties of different tests. Figure 5.2(a) shows that all tests rejected the null hypothesis in nearly 5% of the cases.

Recall that in the example with standard t_5-distribution (t distribution with 5 degrees of freedom) all sub-vectors were uncorrelated but not independent. Our proposed tests could identify this dependency very well. Figure 5.2(b) clearly shows that the performances of all proposed tests were much superior than JdCov, rank-JdCov and dHSIC tests. Among these proposed tests, the one based on $T^{U}_{\text{sum},n}$ had an edge. The JdCov test and its rank version did not have satisfactory performance in this example.

In 'Mixture Normal-1' example, for any $q \neq s$, pairwise $\mathbf{X}^{(q)}$-distances and pairwise $\mathbf{X}^{(s)}$-distances are positively correlated, but in 'Mixture Normal-2' example, they have negative correlation (see Biswas *et al.*, 2016, for details). In both of these examples, powers of our tests were much higher than dHSIC, JdCov and rank-JdCov tests (see, Figure 5.2(c) and 5.2(d)). In 'Mixture Normal-1' example, our proposed tests based on univariate ranks had the best performance closely followed by those based on MMD and coordinate-wise ranks. Compared to them, the tests based on spatial ranks had relatively low powers. But in 'Mixture Normal-2' example, we observed an opposite picture. In that example, the tests based on spatial ranks outperformed others, though those based on coordinate-wise ranks and MMD also had competitive performance. These six tests performed much better than the tests based on univariate ranks. In this example, dHSIC and JdCov and rank-JdCov tests had miserable performance. The JdCov test and its rank version had poor performance in 'Mixture Normal-1' example as well.

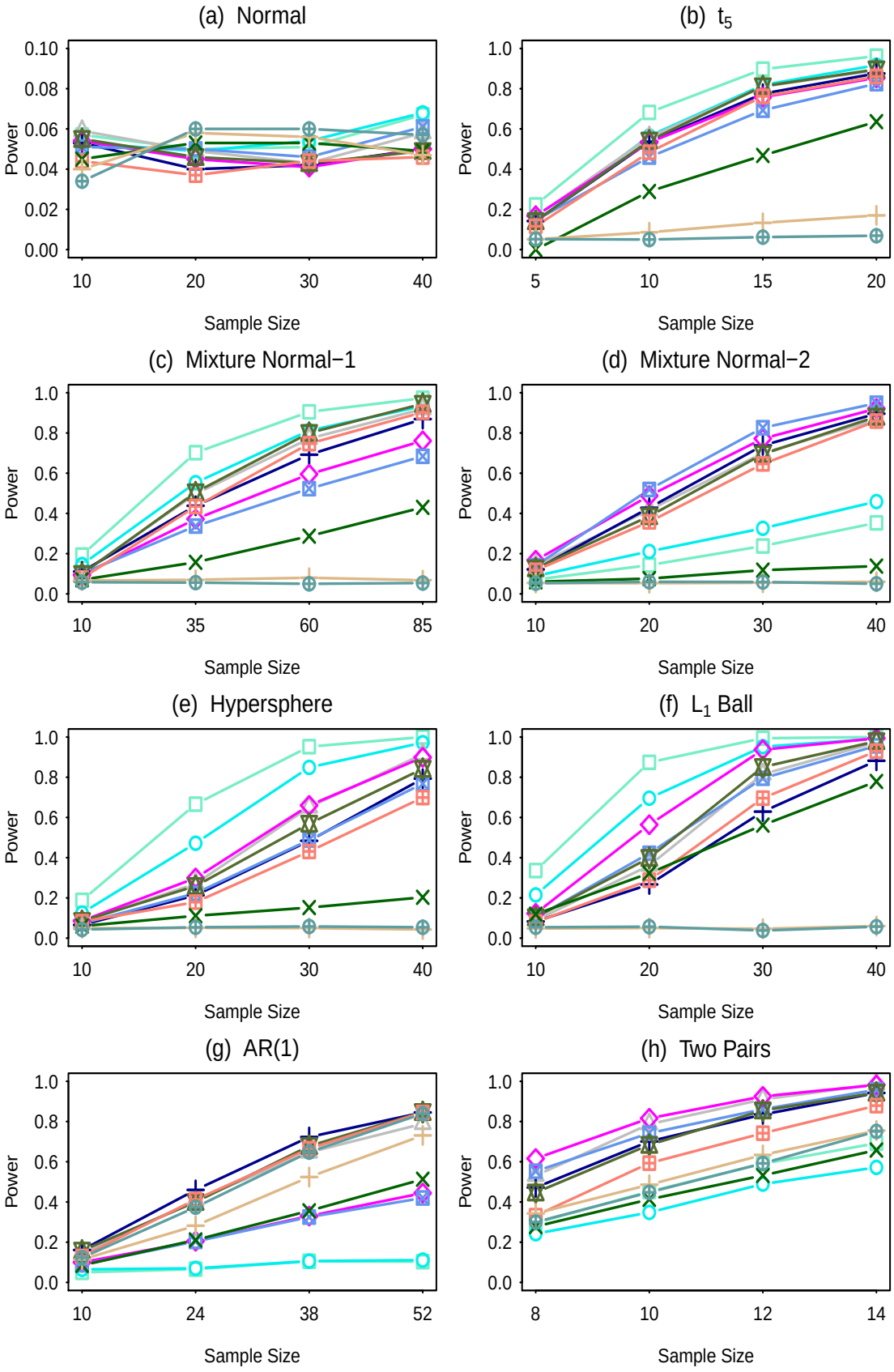


FIGURE 5.2: Powers of dHSIC (×), JdCov (+), rank-JdCov (⊕) tests and the proposed tests based on $T^U_{\text{sum},n}$ (□), $T^U_{\max,n}$ (○), $T^C_{\text{sum},n}$ (△), $T^C_{\max,n}$ (+), $T^S_{\text{sum},n}$ (◇), $T^S_{\max,n}$ (⊠), $T^M_{\text{sum},n}$ (⊠), $T^M_{\max,n}$ (⊞) tests in simulated data sets with four 5-dimensional sub-vectors.

In 'Hypersphere' and 'L_1 Ball' examples also, our proposed methods performed better than their competitors (see Figures 5.2(e) and 5.2(f)). JdCov and rank-JdCov tests had poor performance in both of these examples. The dHSIC test had somewhat competitive power in the 'L_1 Ball' example, but its performance in the 'Hypersphere' example was not satisfactory. Among the proposed tests, the ones based on univariate ranks had better performance in these examples. The test based $T^U_{\text{sum},n}$ had the highest power in both cases.

Next, we considered an example (referred to as the 'AR(1)'), where we generated observations from the 20-dimensional multivariate normal distribution with mean zero and block diagonal covariance matrix $\mathbf{\Sigma}$ of the form $\mathbf{\Sigma} = \text{diag}(\mathbf{S}_{10}, \frac{1}{10}\mathbf{S}_5, 10\mathbf{S}_5)$, where $\mathbf{S}_t$ is the $t \times t$ matrix representing the covariance structure of an AR(1) model (auto-regressive model of order 1). We chose $\mathbf{S}_t = ((s_{ij}))$, where $s_{ij} = \rho^{|i-j|}$ for all $i, j = 1, 2, \ldots, t$, and carried out our experiment for different choices of ρ. Figure 5.2(g) shows the result for $\rho = 0.8$, but for other choices of ρ, the relative performance of different tests were nearly the same. In this example, the tests based on MMD, coordinate-wise ranks and the rank-JdCov test performed better than the rest. Tests based on spatial ranks and the dHSIC test had almost similar performance, but the powers of the tests based univariate ranks were not satisfactory at all. Note that in this example, we have dependence between $X^{(1)}$ and $X^{(2)}$, but the other two sub-vectors $X^{(3)}$ and $X^{(4)}$ are independent and they can be considered as noise. Since $X^{(4)}$ has high stochastic variation compared to other sub-vectors, it played the leading role in determining the $\mathbf{X}^{(-q)}$-ranks of $\mathbf{X}^{(q)}$-neighbors for $q = 1, 2, 3$. This was the main reason behind the poor performance of the tests based on univariate ranks. The tests based on spatial ranks were also somewhat affected by this phenomenon.

In the last example, we generated observations on the two pairs of random vectors $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ and $(\mathbf{X}^{(3)}, \mathbf{X}^{(4)})$ independently from the same distribution. Observations on $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ were generated using the model $\mathbf{X}^{(1)} = \mathbf{W} + \boldsymbol{\epsilon}_1$, $\mathbf{X}^{(2)} = -\mathbf{W} + \boldsymbol{\epsilon}_2$, where $\mathbf{W} \sim N_5(\mathbf{0}, \mathbf{I}_5)$, $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \sim N_5(\mathbf{0}, \delta^2\mathbf{I}_5)$, and they are independent. Observations on $(\mathbf{X}^{(3)}, \mathbf{X}^{(4)})$ were generated similarly. We carried out our experiment for different choices of δ^2 in the range $(0, 1)$. In all cases, our proposed tests based on multivariate ranks of nearest neighbors outperformed their competitors. In Figure 5.2(h), we have reported the result for $\delta^2 = 4/9$. Superiority of the tests based on coordinate-wise ranks, spatial ranks and MMD is quite evident in this figure.

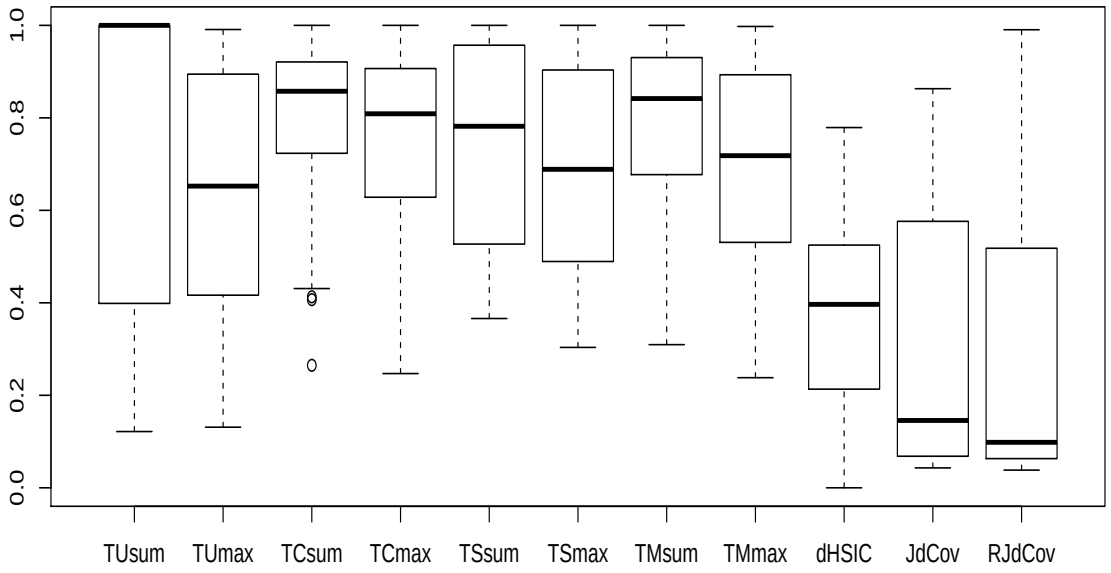


FIGURE 5.3: Boxplots of efficiency scores of dHSIC, JdCov, rank-JdCov (RJdCov) tests and the proposed tests based on $T^U_{\text{sum},n}$ (TUsum), $T^U_{\text{max},n}$ (TUmax), $T^C_{\text{sum},n}$ (TCsum), $T^C_{\text{max},n}$ (TCmax), $T^S_{\text{sum},n}$ (TSsum), $T^S_{\text{max},n}$ (TSmax), $T^M_{\text{sum},n}$ (TMsum), $T^M_{\text{max},n}$ (TMmax) in seven simulated data sets.

To compare the overall performance of different tests in a comprehensive way, we constructed the boxplots of efficiency scores as before. Efficiency scores were computed for different tests on different data sets (barring the 'Normal' example, where $\mathbf{X}$ and $\mathbf{Y}$ are independent), and they are presented using boxplots in Figure 5.3. This figure clearly suggests that the overall performance of our proposed tests was better than dHSIC, JdCov and rank-JdCov tests. It also shows that the tests based on sum (i.e., $T^U_{\text{sum},n}$, $T^C_{\text{sum},n}$, $T^S_{\text{sum},n}$ and $T^M_{\text{sum},n}$) had better overall performance than their corresponding version based on max (i.e., $T^U_{\text{max},n}$, $T^C_{\text{max},n}$, $T^S_{\text{max},n}$ and $T^M_{\text{max},n}$). Among them the test based on $T^U_{\text{sum},n}$ had the highest power in maximum number of cases, but in some cases (e.g., the 'AR(1)' data set), its performance was very poor. On the other hand, the test based on $T^C_{\text{sum},n}$ and $T^M_{\text{sum},n}$ had consistently good performance, and in that sense, they outperformed the corresponding test based on spatial rank. We repeated our experiment with these eight examples for varying choices of d_1, d_2, d_3, d_4, but our basic findings remained almost the same. That is why, we chose not to report those results again.

5.5 Results from the analysis of real data sets

For further evaluation of our proposed tests, we analyzed four real data sets, 'Airfoil Self-noise data', 'Census 1980 data', 'Pollution data' and 'Tecator data'. Airfoil Self-noise data

set is available at the UCI machine learning repository (https://archive.ics.uci.edu/ml/). The other three data sets are taken from the CMU data archive (http://lib.stat.cmu.edu/datasets/). Description of the Airfoil Self-noise data was given in Chapter 2 and that of Tecator data and Pollution data was given in Chapter 4. Brief description of the Census 1980 data is given below.

'Census 1980 data' were collected from 50 states of USA in the year 1980. It contains information on median age, percentage of above 65 years individuals, per capita income, percentage of individuals having education up to 12th standard, and that having education up to the college level. This data set and its description can also be found in Witmer (1997). For our analysis, we divided the variables into 3 groups: age structure, income and education. Naturally, one expects the per capita income to depend on the age structure and the education level of the population. So, we carried out different tests to check whether they can successfully identify the dependence among these three groups of variables.

When we performed our experiment using the full data set, both for Tecator data and Census 1980 data, all tests rejected the null hypothesis. Based on that single experiment, it was not possible to compare among different test procedures. So, we carried out our experiment using randomly chosen subsets of observations from the full data set. For each sub-sample size (reported in Figure 5.4), the experiment was repeated 1000 times to compute the powers of different tests, and they are shown in Figure 5.4. recall that the dHSIC test needs the sample size to be at least twice the number of sub-vectors. So, in the case of Tecator data, this test could not be used for samples of size smaller than 8.

In the case of Airfoil Self-noise data, there are only two sub-vectors. So, the HHG test could be used for this data set. Figure 5.4(a) shows that in this example, all tests had similar powers. In the case of Census 1980 data also, all tests had comparable performance, but the powers of all proposed tests were slightly higher than those of dHSIC, JdCov and rank-JdCov tests (see Figure 5.4(b)). The rank-JdCov test had the best performance in Pollution data (see Figure 5.4(c)). In this data set, the tests based on univariate ranks and spatial ranks did not have satisfactory performance, but those based on coordinate-wise ranks and MMD worked well. These tests and the JdCov test had much higher powers than the dHSIC test. The results on Tecator data shows the superiority of our proposed tests (see Figure 5.4(d)). All of them had much higher powers than JdCov, rank-JdCov and dHSIC tests. The rank-JCdov test had poor performance in this example.

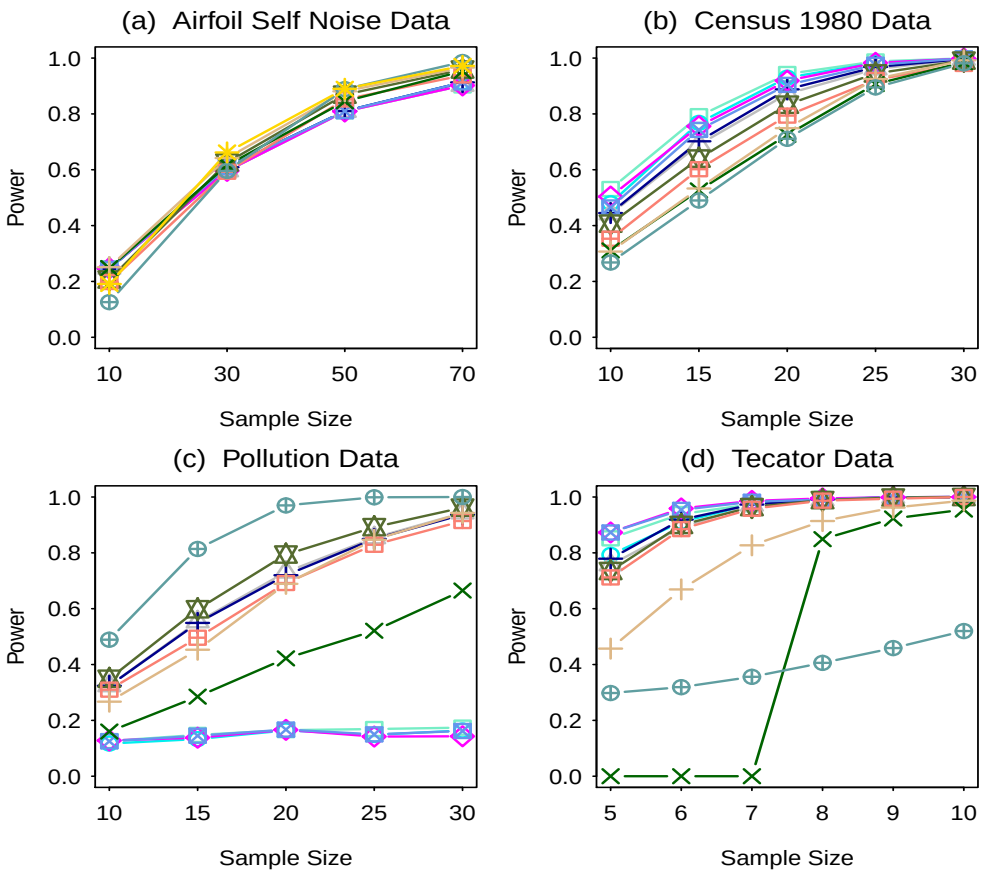


FIGURE 5.4: Powers of dHSIC (×), JdCov (+), rank-JdCov (⊕), HHG (✳) tests and the proposed tests based on $T^{U}_{\text{sum},n}$ (□), $T^{U}_{\text{max},n}$ (○), $T^{C}_{\text{sum},n}$ (△), $T^{C}_{\text{max},n}$ (+), $T^{S}_{\text{sum},n}$ (◇), $T^{S}_{\text{max},n}$ (⊠), $T^{M}_{\text{sum},n}$ (⊠), $T^{M}_{\text{max},n}$ (⊞) in real data sets.

5.6 Analysis of functional data

The tests proposed in this chapter are all based on pairwise Euclidean distances. So, they can also be used for testing independence among several random functions in infinite dimensional functional spaces (e.g., separable Banach spaces). Here, we used the six examples considered in Section 4.6 to investigate their empirical performance for such data sets. For each example, we considered samples of different sizes as before. For each sample size, the experiment was repeated 1000 times to compute the powers of different tests, and they are reported in Figure 5.5. Powers of dHSIC, JdCov and HHG tests are also reported to facilitate the comparison. Note that all these examples deal with only two sub-vectors. In such cases, the tests based on coordinate-wise ranks and spatial ranks coincide with the corresponding tests based univariate ranks discussed in Section 5.1 (clear from the definition of coordinate-wise ranks and spatial ranks). So here we report the results for tests based on univariate ranks only. Of course, the results for the proposed tests based on MMD are also reported.

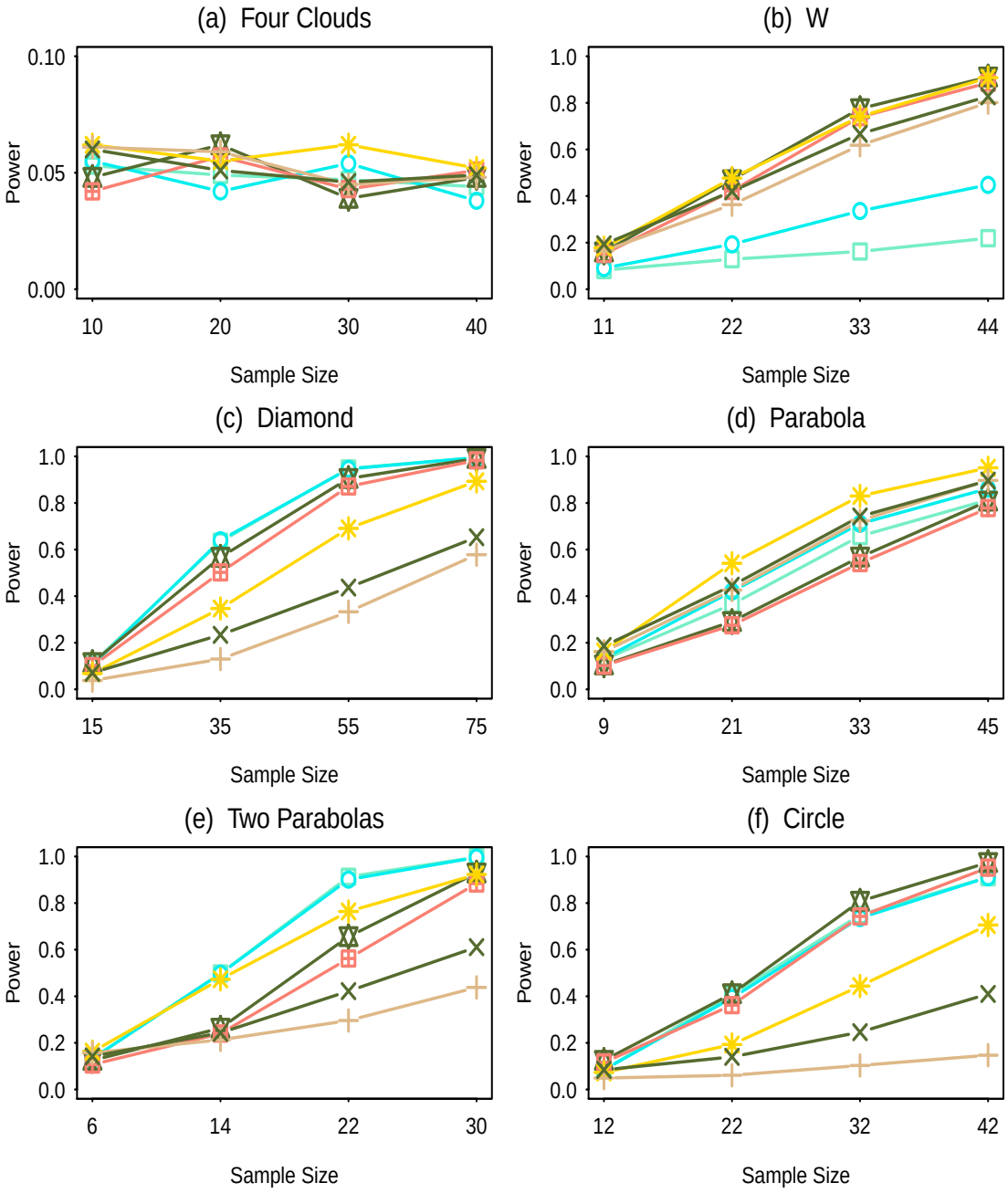


FIGURE 5.5: Powers of JdCov (+), dHSIC (×), HHG(✳) tests and the proposed tests based on $T^U_{\text{sum},n}$ (□), $T^U_{\max,n}$ (○), $T^M_{\text{sum},n}$ (⊠), $T^M_{\max,n}$ (⊞) in functional data sets.

In 'Four Clouds' example, where the two random functions are independent, as expected, all tests had powers close to the nominal level. Our proposed tests based on MMD had satisfactory performance in all examples though in the 'Parabola' example, its power was slightly lower compared to other competitors. In cases of 'W', 'Circle' and 'Diamond' examples, they outperformed most of the tests considered here. Barring the 'W' example, the tests based on univariate ranks performed well. They outperformed all other tests in 'Diamond' and 'Two Parabolas' examples. The HHG test had the highest power in the

'Parabola' example, while in other examples, it had moderate performance. The dHSIC test and the JdCov test had good performance in 'W' and 'Parabola' examples, but in the other three examples ('Diamond;, 'Two Parabolas' and 'Circle'), they had much lower power than their competitors.

5.7 Application in causal discovery

Like Section 4.7, here also we use our proposed tests to unveil the causal relationship among p random vectors $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(p)}$. As before, we consider all possible structural equation models with additive noise, which lead to DAG on p nodes. Recall that such a structural equation model (SEM) has the following form:

$$\mathbf{X}^{(q)} = f_q(\mathbf{PA}^{(q)}) + \boldsymbol{\epsilon}^{(q)}, \text{ for } q = 1, 2, \cdots, p,$$

where $\mathbf{PA}^{(q)}$ denotes the set of all parent nodes of $\mathbf{X}^{(q)}$, and the $\boldsymbol{\epsilon}^{(q)}$'s are independent additive noise vectors. If there are no parent nodes, we take f_q to be zero.

For any such SEM, using the observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ on $\mathbf{X}$, for each $q = 1, 2, \ldots, p$, we construct $\hat{f}_q$, an estimate of f_q, by regressing $\mathbf{X}^{(q)}$ on its parent nodes $\mathbf{PA}^{(q)}$ using a nonparametric method, and compute the corresponding residuals $\widehat{\boldsymbol{\epsilon}}_i^{(q)} := \mathbf{x}_i^{(q)} - \widehat{f}_q(\mathbf{PA}_i^{(q)})$ for $i = 1, 2, \ldots, n$. We perform a test of independence among these residual sub-vectors to compute the corresponding p-value. Among all possible SEMs that can be represented using DAG, the model with the highest p-value is selected. However, if this highest p-value is smaller than 0.05, none of the SEMs is selected.

For our experiment, we considered the same examples used in Section 4.7. We repeated each experiment 100 times as before, and the results are reported in Tables 5.1 and 5.2. Recall that in the examples with two sub-vectors, the proposed tests based on univariate ranks coincide with those based on coordinate-wise ranks or spatial ranks. So, in Tables 5.1, we have reported the results only for the tests based on univariate ranks and MMD. Of course, results for HHG, JdCov, rank-JdCoV and dHSIC tests are also reported for comparison. When the observations on $\mathbf{X}^{(1)} = (U_1, U_2)$ were generated from the standard bivariate normal distribution (recall that those on $\mathbf{X}^{(2)} = (V_1, V_2)$ were obtained from them using the model $V_i = U_i^2 + \varepsilon_i$, where the ε_i's are i.i.d. $N(0, 0.01)$ random variables), the tests based on univariate ranks and the HHG test had better performance than their competitors. Among the rest of the methods, the tests based MMD performed slightly better than

dHSIC and much better than JdCov and rank-JdCov tests. We observed almost the same picture, when instead of standard normal distribution, observations on $\mathbf{X}^{(1)}$ were generated from the standard bivariate t distribution with 2 degrees of freedom.

TABLE 5.1: Proportion of times the correct model was selected by different methods in the example involving two random vectors.

	$T^U_{\text{sum},n}$	$T^U_{\max,n}$	$T^M_{\text{sum},n}$	$T^M_{\max,n}$	dHSIC	JdCov	r-JdCov	HHG
Normal	0.89	0.90	0.75	0.74	0.70	0.47	0.48	0.88
t_2	0.94	0.95	0.88	0.88	0.86	0.61	0.57	0.98

In our second example involving three random variables, we used samples of size 30. As we have mentioned before, in this example, there are two super models (see Figures 4.10(b) and 4.10(c)), which contain the true SEM. So, in this example, we counted the number of times a method selected one of these three models depicted in Figure 4.10, and they are reported in Table 5.2. This table clearly shows that our proposed tests performed much better than dHSIC, JdCov and rank-JdCov tests. Among the proposed tests, the ones based on spatial ranks and MMD had an edge.

TABLE 5.2: Proportion of times the true model and two super models were selected by different methods in the example involving three random variables.

Model	$T^U_{\text{sum},n}$	$T^U_{\max,n}$	$T^C_{\text{sum},n}$	$T^C_{\max,n}$	$T^S_{\text{sum},n}$	$T^S_{\max,n}$	$T^M_{\text{sum},n}$	$T^M_{\max,n}$	dHSIC	JdCov	r-JdCov
True	0.18	0.17	0.14	0.12	0.28	0.18	0.36	0.22	0.06	0.13	0.16
SM-1	0.20	0.18	0.22	0.19	0.18	0.21	0.14	0.18	0.08	0.11	0.16
SM-2	0.15	0.12	0.20	0.21	0.21	0.19	0.15	0.20	0.12	0.10	0.14
Total	0.53	0.47	0.56	0.52	0.67	0.58	0.65	0.60	0.26	0.34	0.46

5.8 Proofs and mathematical details

Proof of Lemma 5.1. If $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are mutually independent, clearly $\mathbf{X}^{(q)}$ and $\mathbf{X}^{(-q)}$ are independent for every $q = 1, 2 \ldots, p$. We now prove that the converse also holds. Let $\chi_{(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\ldots,\mathbf{X}^{(p)})}(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \ldots, \mathbf{t}^{(p)})$ be characteristic function of the joint distribution of $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)})$ for all $\mathbf{t}^{(1)} \in \mathbb{R}^{d_1}, \mathbf{t}^{(2)} \in \mathbb{R}^{d_2}, \ldots, \mathbf{t}^{(p)} \in \mathbb{R}^{d_p}$. Now,

$$\begin{aligned}\chi_{(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\ldots,\mathbf{X}^{(p)})}(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \ldots, \mathbf{t}^{(p)}) &= \chi_{(\mathbf{X}^{(1)},\mathbf{X}^{(-1)})}\left(\mathbf{t}^{(1)}, (\mathbf{t}^{(2)}, \mathbf{t}^{(3)}, \ldots, \mathbf{t}^{(p)})\right) \\ &= \chi_{\mathbf{X}^{(1)}}(\mathbf{t}^{(1)})\chi_{\mathbf{X}^{(-1)}}(\mathbf{t}^{(2)}, \mathbf{t}^{(3)}, \ldots, \mathbf{t}^{(p)}),\end{aligned}$$

where the second line follows because of independence between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(-1)}$. Now,

$$\chi_{\mathbf{X}^{(-1)}}(\mathbf{t}^{(2)},\mathbf{t}^{(3)},\ldots,\mathbf{t}^{(p)}) = \chi_{(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\mathbf{X}^{(3)},\ldots,\mathbf{X}^{(p)})}(\mathbf{0}_{d_1},\mathbf{t}^{(2)},\mathbf{t}^{(3)},\ldots,\mathbf{t}^{(p)})$$
$$= \chi_{(\mathbf{X}^{(2)},\mathbf{X}^{(-2)})}\left(\mathbf{t}^{(2)},(\mathbf{0}_{d_1},\mathbf{t}^{(3)},\ldots,\mathbf{t}^{(p)})\right) = \chi_{\mathbf{X}^{(2)}}(\mathbf{t}^{(2)})\chi_{\mathbf{X}^{(-2)}}(\mathbf{0}_{d_1},\mathbf{t}^{(3)},\ldots,\mathbf{t}^{(p)})$$
$$= \chi_{\mathbf{X}^{(2)}}(\mathbf{t}^{(2)})\chi_{(\mathbf{X}^{(3)},\mathbf{X}^{(4)},\ldots,\mathbf{X}^{(p)})}(\mathbf{t}^{(3)},\mathbf{t}^{(4)}\ldots,\mathbf{t}^{(p)}).$$

Proceeding this way, we get

$$\chi_{(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\ldots,\mathbf{X}^{(p)})}(\mathbf{t}^{(1)},\mathbf{t}^{(2)},\ldots,\mathbf{t}^{(p)}) = \chi_{\mathbf{X}^{(1)}}(\mathbf{t}^{(1)})\chi_{\mathbf{X}^{(2)}}(\mathbf{t}^{(2)})\ldots\chi_{\mathbf{X}^{(p)}}(\mathbf{t}^{(p)}).$$

This completes the proof. □

Proof of Lemma 5.2. Let $\mathbf{W},\mathbf{W}_*$ be two independent random vectors, which follow the distribution $\mathbb{U}_{n-k}^{p-1}$ and independent of $\mathbf{V},\mathbf{V}_*$. So, we have $\gamma_{K_\sigma}\left(\mathbb{F}_k^{(q)},\mathbb{U}_{n-k}^{p-1}\right) = [S_1 - 2S_2 + S_3]^{\frac{1}{2}}$, where $S_1 = \mathrm{E}K_\sigma(\mathbf{V},\mathbf{V}_*)$, $S_2 = \mathrm{E}K_\sigma(\mathbf{V},\mathbf{W})$ and $S_3 = \mathrm{E}K_\sigma(\mathbf{W},\mathbf{W}_*)$. Now, note that $W^{(1)},W^{(2)},\ldots,W^{(p-1)}$ and $W_*^{(1)},W_*^{(2)},\ldots,W_*^{(p-1)}$ are all independent and follow uniform distribution on $\{1,2,\ldots,n-k\}$. So, we get

$$S_2 = \mathrm{E}K_\sigma(\mathbf{V},\mathbf{W}) = \mathrm{E}\left[\prod_{t=1}^{p-1} e^{-\frac{\left(V^{(t)}-W^{(t)}\right)^2}{2\sigma^2}}\right] = \mathrm{E}\left[\prod_{t=1}^{p-1}\mathrm{E}\left\{e^{-\frac{\left(V^{(t)}-W^{(t)}\right)^2}{2\sigma^2}}\middle| V^{(t)}\right\}\right]$$
$$= \mathrm{E}\left[\prod_{t=1}^{p-1}\frac{1}{n-k}\sum_{\ell=1}^{n-k} e^{-\frac{\left(V^{(t)}-\ell\right)^2}{2\sigma^2}}\right] \text{ and}$$
$$S_3 = \mathrm{E}K_\sigma(\mathbf{W},\mathbf{W}_*) = \mathrm{E}\left[\prod_{t=1}^{p-1} e^{-\frac{\left(W^{(t)}-W_*^{(t)}\right)^2}{2\sigma^2}}\right] = \prod_{t=1}^{p-1}\mathrm{E}\left[e^{-\frac{\left(W^{(t)}-W_*^{(t)}\right)^2}{2\sigma^2}}\right]$$
$$= \prod_{t=1}^{p-1}\frac{1}{(n-k)^2}\sum_{\ell=1}^{n-k}\sum_{\ell'=1}^{n-k} e^{-\frac{\left(\ell-\ell'\right)^2}{2\sigma^2}} = \left[\frac{1}{(n-k)^2}\sum_{\ell=1}^{n-k}\sum_{\ell'=1}^{n-k} e^{-\frac{\left(\ell-\ell'\right)^2}{2\sigma^2}}\right]^{p-1}$$
$$= \left[\frac{2}{(n-k)^2}\sum_{\ell=1}^{n-k-1}(n-k-\ell)\,e^{-\frac{\ell^2}{2\sigma^2}} + \frac{1}{n-k}\right]^{p-1}.$$

This completes the proof. □

Chapter 6

Concluding Remarks

In this thesis, we have proposed and investigated some tests for independence among multiple random variables and random vectors of arbitrary dimensions, and we have shown that the tests proposed for random vectors can also be used for testing independence among several random functions.

In Chapter 2, we developed a copula based statistic to measure dependence among several continuous random variables and constructed a distribution-free method to test for the statistical significance of that measure. Unlike most of the existing methods, our proposed measure and the associated tests are invariant under permutations and strictly monotone transformations of the variables. However, they involve a smoothing parameter called bandwidth, which needs to be chosen appropriately. We have seen that though the bandwidth chosen using median heuristic usually performs well, the use of smaller bandwidths sometimes yields better results. While larger bandwidths successfully detect global linear or monotone relationships among the variables, smaller bandwidths are useful for detecting non-monotone or local patterns. In order to capture both types of dependence, we adopted a multi-scale approach, where the results for several choices of the bandwidths were aggregated to arrive at the final decision. This approach also leads to distribution-free tests. Though the computing costs of these multi-scale methods are slightly higher than their single-scale analogs (e.g., the one based on median heuristic), they usually lead to much better performance, especially when the variables have complex non-monotone relationships. We proposed three methods for aggregation. Based on the empirical performance of these three methods, we recommend using the test based on $T_{\max,n}$ (maximum of the test statistics computed for different bandwidths), particularly when one deals with small number of variables. Using the idea of checkerboard copula, in Chapter 3, we generalized our dependency measure and the associated tests so that they can be used for handling

random variables having arbitrary probability distributions, where the observations on a variable may have ties and the ranks cannot be uniquely defined. In this set up also, we observed the multi-scale methods to have an edge over their single-scale analog, especially for detecting complex non-monotone relationships among the variables. Among the multi-scale methods, the one based on $T^{\maltese}_{\max,n}$, the maximum of the test statistics, usually yields better results.

In Chapter 4, we proposed multivariate generalizations of our tests so that they can be used for testing independence among several random vectors of arbitrary dimensions. We proposed two general recipes for this purpose. One of them was based on pairwise distances (or distances of the observations from specified points) and the other one was based on linear projections. We carried out several experiments to compare between these two methods of generalization. When the norms of the sub-vectors carry significant information about dependence, which is often the case, the method based on pairwise distances are preferred. Otherwise, the method based on linear projection may yield better results. Again, we have single-scale and multi-scale versions of these two methods, and the multi-scale versions usually perform better in complex examples. We have seen that when the sub-vectors are one-dimensional, the single-scale method based on linear projections is almost equivalent to the copula based test proposed in Chapter 2. So, among the multi-scale versions of this method, the one based on maximum of the test statistics ($\widetilde{\zeta}_{\max,n}$) usually performs better. However, the one-dimensional version of the single-scale test based on pairwise distances differ from the copula based test proposed in Chapter 2. Among the multi-scale versions of this test, the one based on sum of the test statistics ($\zeta_{\text{sum},n}$) generally yields better results, and we have seen that in most of our experiments.

In Chapter 5, we proposed some tests of independence based on ranks of nearest neighbors. Sarkar and Ghosh (2018) used the idea of nearest neighbors to construct some tests of independence between two random vectors. We proposed several ways for generalizing these tests for more than two random vectors of arbitrary dimensions. A comparison among these different methods was carried out in Section 5.4, which shows that the tests based on $T^{C}_{\text{sum},n}$ (coordinate-wise rank) and $T^{M}_{\text{sum},n}$ (MMD) are probably the best ones in this lot.

Among the tests based on one-dimensional projections (i.e. the tests based on pairwise distances or those based on linear projections) and those based on ranks of nearest neighbors, these is no clear winner. Depending on nature of the problem, one of these two types

of tests come up with better performance. Using several simulated and real data sets, in Chapters 4 and 5, we have amply demonstrated that these proposed tests can outperform the state of the art tests like dHSIC (Pfister *et al.*, 2018), JdCov, rank-JdCov (Chakraborty and Zhang, 2019) and HHG (Heller *et al.*, 2013) tests in a wide variety of examples. We have also observed the same picture when dealing with functional data. Note that unlike the dHSIC test, our proposed tests can be conveniently used even when the sample size is smaller than the number of sub-vectors.

However, these proposed methods are not above all limitations. The choice of the bandwidth is still an issue to be resolved. Though the bandwidth chosen using median heuristic usually works well, in many examples, smaller bandwidths lead to better results. In order to take care of this problem, we adopted a multi-scale approach, where the results for different bandwidths were aggregated. However, our choice of the number of bandwidths for aggregation and the choices of upper and lower bounds of those bandwidths were somewhat adhoc. The resulting tests worked well in all simulated and real data sets considered in this thesis, but a more judicious choice of these parameters may lead to further improvement. One can use the Bayesian multi-scale approach (Erästö and Holmström, 2005; Mukhopadhyay and Ghosh, 2011; Dutta *et al.*, 2016) for this purpose, but this method needs the prior distribution to be chosen and its computing cost is usually higher. Also, our empirical experience suggests that instead of using a multi-scale approach, sometimes it is better to choose a suitable data driven estimate of the bandwidth, both in terms of power of the resulting test and the computing time. But we are yet to develop an automatic data driven method in this regard. Note that the dHSIC test (Pfister *et al.*, 2018) also uses a kernel (usually Gaussian kernel) with a bandwidth chosen using median heuristic. The use of the multi-scale approach or a data driven choice of bandwidth may improve the performance of that test as well.

For the tests based on ranks of nearest neighbors, Sarkar and Ghosh (2018) used another type of multi-scale approach, where instead of taking the sum over all $k = 1, 2, \ldots, n-2$ (see the equation in the second page of Chapter 5), they took the sum over $k = 1, 2, \ldots, k_0$ for some $k_0 \leq n-2$. They looked at the results for several choices of $k_0 \leq n-2$ and then aggregated them to come up with the final decision. Similar aggregation techniques can be used for the methods proposed in Chapter 5. However, to reduce the computing cost, we did not consider that approach in this thesis. This can be investigated in a future work.

In Chapter 2, Theorem 2.5 shows that for any positive constant Δ, it is possible to test a null hypothesis of the form $\mathbb{H}_0^{'} : I_\sigma(\mathbf{X}) \geq \Delta$ against the alternative $\mathbb{H}_1^{'} : I_\sigma(\mathbf{X}) < \Delta$. But for constructing such a test, one needs to come up with a consistent estimator of δ^2 (see Theorem 2.5). Finding such an estimator of δ^2 will also enable us to find the asymptotic power of the proposed test of independence under suitable shrinking alternatives. This can be considered as an interesting problem for future research.

In order to prove the consistency of our proposed test based on pairwise distances, in Theorem 4.3, we assumed the underlying joint distribution F to be absolutely continuous. But, from the proof of this theorem, it is clear that it is enough to have continuity of the marginal distributions $F_j^{\mathbf{a}}$'s and non-singularity of F with respect to the Lebesgue measure. For the implementation of our tests based on one-dimensional projections, throughout Chapter 4, we used $P = F$. Our empirical experience suggests that it is a reasonably good choice. However, one can consider other choices of P as well. It would have been ideal to choose the probability distribution P in such a way that the power of the resulting test gets maximized in a given problem. But, it is extremely difficult to find out such a probability distribution. This can be considered as a problem for future investigation. If we do not use $P = F$, one also needs to decide how many observations are to generated from P for constructing the estimate of the dependency measure in Subsection 4.1.2. In this thesis, we proposed two general recipes for multivariate generalization of the tests used in Chapter 2. One of them is based on pairwise distances and the other one is based on linear projections. From our numerical results, it seems to be a better idea to use the method based on pairwise distances, but the method based on linear projections sometimes lead to higher power. Our discussion in Section 4.3 provides some insight in this regard. But it will be advantageous if one can develop an algorithm that can automatically decide on the method to be used for a given data set.

Throughout this thesis, for computing MMD, we used the Gaussian kernel. Other characteristic kernel functions (e.g., exponential kernel) may also be used for this purpose, but we have not investigated the empirical performance of the resulting tests for those choices of the kernel functions. Similarly, in this thesis, we used the Euclidean distance as the distance function for the implementation of our tests in Chapter 4 and 5. But from the description of the proposed tests it is clear that other appropriate distance functions can also be used. For instance, if the measurement variables are not of comparable units

and scales, one can use the tests based on Mahalanobis distance. But we have not studied the performance of the tests based on those distance functions. For the construction of our tests based on ranks of nearest neighbors, in Sections 5.1 and 5.2, we made a transformation $t \mapsto -\text{sign}(t)\log(1-|t|)$ (respectively, $\mathbf{t} \mapsto -\mathbf{Sign}_{p-1}(\mathbf{t})\log(1-\|\mathbf{t}\|)$ in the case of tests based on spatial rank). The choice of this transformation was motivated by the work of Heller *et al.* (2012); Biswas *et al.* (2016); Sarkar and Ghosh (2018). It magnifies the signal against $\mathbb{H}_0$ for extreme values of rank without affecting its sign (respectively, direction in the case of spatial rank). But there are several other transformations, which also have this property. At this moment, it is not clear to us how to find the optimal transformation (the one leading to the maximum power) for a given data set. This can be investigated in future. The tests based on ranks of nearest neighbors proposed in Chapter 5 performed reasonably well in all simulated and real data sets analyzed in this thesis. In many cases, they outperformed the popular tests available in the literature. In all of our examples, we observed the powers of these tests to increase with the sample size. But, at this moment, we do not have any theoretical result related to the consistency of these tests. On the other hand, we could not construct a single counter-example to show that the power of these tests may not converge to unity with increasing sample size. So, most probably, these tests are consistent, but we are yet to prove it. This is an interesting and challenging theoretical problem, which we would like to investigate in near future.

In this thesis, we used the HHG test (Heller *et al.*, 2013) when there were two sub-vectors. In many examples, this test performed well. But, unfortunately, we do not have meaningful generalization of this test for the $p > 2$ case. Using the idea given Section 5.1, one can go for its generalization, and the consistency of the resulting test can also be proved using Lemma 5.1. But, as we have seen, this method of generalization has some limitations. On the other hand, generalization using multi-way contingency table does not lead to good results unless the number of sub-vectors is very small. So, successful generalization of this test still remains a challenging problem.

In Chapters 4 and 5, we used different tests for discovering causal relationships among the sub-vectors. For this purpose, we considered all structural equation models that can be represented using DAG and found the best model in that class. However, this method seems to be computationally feasible only when the number of sub-vectors is small. Note that there are only three competing models for $p = 2$, but for $p = 3$, it increases to 25. So,

finding the best structural equation model in this way becomes computationally prohibitive when the number of sub-vectors is moderately large. One needs to properly address this computational issue to come up with a scalable algorithm.

In order to prove the consistency result for functional data (see Theorem 4.5), we assumed all functions to be fully observed. However, in practice, each function is usually observed only on some grid points, from which one needs to estimate the functions (or the pairwise L_2 distances or the inner products). We did not prove large sample consistency for those practical versions of the tests. However we implemented these versions of the tests in Section 4.6, where we considered densely observed equispaced data on the domain of such function. In the case of sparsely observed data, these methods may not have satisfactory performance, and one may need to construct different test procedures to cope with such situations. This can be considered as another interesting problem for future investigation.

Appendix A

Exact and Asymptotic Means and Variances of $n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)$

Proposition A.1. *Under the null hypothesis of mutual independence, we have*

$$\begin{aligned}
\mathrm{E}\left[n\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi_n)\right] &= 1+(n-1)\left\{\frac{n}{(n)_2}\left(nu_3-1\right)\right\}^p - nu_3^p,\\
\mathrm{Var}\left[n\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi_n)\right] &= \frac{1}{n^2}\left[2(n)_2c_1+4(n)_3c_2+(n)_4c_3\right]+4\left[nu_2^p+(n)_2\left\{\frac{n}{(n)_2}(nu_3^2-u_2)\right\}^p\right]\\
&\quad -\frac{4}{n}\left[2(n)_2c_4+(n)_3c_5\right]-\left[(n-1)\left\{\frac{n}{(n)_2}(nu_3-1)\right\}^p-2nu_3^p\right]^2,
\end{aligned}$$

$$\begin{aligned}
\textit{where}\quad u_1 &= \frac{2}{n^2}\sum_{i=1}^{n-1}(n-i)e^{-\left(\frac{i}{n\sigma}\right)^2}+\frac{1}{n},\\
u_2 &= \frac{1}{n^3}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}e^{-\frac{1}{2}\left(\frac{i-j}{n\sigma}\right)^2}\right]^2,\\
u_3 &= \frac{2}{n^2}\sum_{i=1}^{n-1}(n-i)e^{-\frac{1}{2}\left(\frac{i}{n\sigma}\right)^2}+\frac{1}{n},\\
c_1 &= \left\{\frac{n}{(n)_2}\left(nu_1-1\right)\right\}^p,\ c_2=\left[\frac{n}{(n)_3}\left\{n^2u_2-n(2u_3+u_1)+2\right\}\right]^p,\\
c_3 &= \left[\frac{n}{(n)_4}\left\{n^3u_3^2-2n^2(u_3+2u_2)+n(8u_3+2u_1+1)-6\right\}\right]^p,\\
c_4 &= \left\{\frac{n}{(n)_2}\left(nu_2-u_3\right)\right\}^p,\ c_5=\left[\frac{n}{(n)_3}\left\{n^2u_3^2-n(2u_2+u_3)+2u_3\right\}\right]^p.
\end{aligned}$$

Proof. First of all, from the definitions of u_1, u_2 and u_3, one can easily verify that $u_1 = \frac{1}{n^2}\sum_{1\le i,j\le n} K_{\frac{\sigma}{\sqrt{2}}}\left(\frac{i}{n},\frac{j}{n}\right), u_2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}K_\sigma\left(\frac{i}{n},\frac{j}{n}\right)\right]^2$ and $u_3 = \frac{1}{n^2}\sum_{1\le i,j\le n}K_\sigma\left(\frac{i}{n},\frac{j}{n}\right)$. Next observe that $n\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi_n) = ns_1 - 2ns_2 + nv_3$, where s_1, s_2 and v_3 are as defined in Equation (2.2). So, for deriving the expectation of $n\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi_n)$, we need to find $\mathrm{E}(ns_1)$

and $\mathrm{E}(ns_2)$ first. Under the null hypothesis of mutual independence,

$$\begin{aligned}
\mathrm{E}(ns_1) &= \frac{1}{n}\sum_{1\le i\ne j\le n}\mathrm{E}K_\sigma(\mathbf{y}_i,\mathbf{y}_j)+1 = \frac{1}{n}n(n-1)\mathrm{E}K_\sigma(\mathbf{y}_1,\mathbf{y}_2)+1\\
&= (n-1)\left[\mathrm{E}K_\sigma(y_1^{(1)},y_2^{(1)})\right]^p+1 = (n-1)\left[\frac{1}{(n)_2}\sum_{1\le i\ne j\le n}K_\sigma\left(\frac{i}{n},\frac{j}{n}\right)\right]^p+1\\
&= (n-1)\left\{\frac{n}{(n)_2}\left(nu_3-1\right)\right\}^p+1, \qquad \text{(A.1)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(ns_2) &= \mathrm{E}\left[\sum_{i=1}^n\prod_{j=1}^p\frac{1}{n}\sum_{l=1}^n K_\sigma\left(y_i^{(j)},\frac{l}{n}\right)\right] = \sum_{i=1}^n\prod_{j=1}^p\frac{1}{n}\sum_{l=1}^n\mathrm{E}K_\sigma\left(y_i^{(j)},\frac{l}{n}\right)\\
&= \sum_{i=1}^n\prod_{j=1}^p\frac{1}{n^2}\sum_{l_1=1}^n\sum_{l_2=1}^n K_\sigma\left(\frac{l_1}{n},\frac{l_2}{n}\right) = nu_3^p. \qquad \text{(A.2)}
\end{aligned}$$

So, from Equations (A.1) and (A.2), we get

$$\begin{aligned}
&\mathrm{E}[n\gamma^2_{K_\sigma}(\mathsf{C}_n,\Pi_n)] = n\mathrm{E}[s_1]-2n\mathrm{E}[s_2]+nv_3 = (n-1)\left\{\frac{n}{(n)_2}\left(nu_3-1\right)\right\}^p+1-2nu_3^p+nu_3^p\\
&= 1+(n-1)\left\{\frac{n}{(n)_2}\left(nu_3-1\right)\right\}^p-nu_3^p.
\end{aligned}$$

Similarly, under the null hypothesis, we have

$$\begin{aligned}
\mathrm{Var}(ns_1) &= n^2\mathrm{Var}\left(\frac{1}{n^2}\sum_{1\le i,j\le n}K_\sigma(\mathbf{y}_i,\mathbf{y}_j)\right) = n^2\mathrm{Var}\left(\frac{1}{n^2}\sum_{1\le i\ne j\le n}K_\sigma(\mathbf{y}_i,\mathbf{y}_j)\right)\\
&= \frac{1}{n^2}\sum_{1\le i\ne j\le n}\sum_{1\le i'\ne j'\le n}\mathrm{Cov}\left(K_\sigma(\mathbf{y}_i,\mathbf{y}_j),K_\sigma(\mathbf{y}_{i'},\mathbf{y}_{j'})\right)\\
&= \frac{1}{n^2}\sum_{1\le i\ne j\le n}\sum_{1\le i'\ne j'\le n}\mathrm{E}\left[K_\sigma(\mathbf{y}_i,\mathbf{y}_j)K_\sigma(\mathbf{y}_{i'},\mathbf{y}_{j'})\right]-(n-1)^2\left[\mathrm{E}K_\sigma(\mathbf{y}_1,\mathbf{y}_2)\right]^2\\
&= \frac{1}{n^2}\Bigg[2(n)_2\underbrace{\mathrm{E}\left\{K_\sigma(\mathbf{y}_1,\mathbf{y}_2)K_\sigma(\mathbf{y}_1,\mathbf{y}_2)\right\}}_{c_1}+4(n)_3\underbrace{\mathrm{E}\left\{K_\sigma(\mathbf{y}_1,\mathbf{y}_2)K_\sigma(\mathbf{y}_2,\mathbf{y}_3)\right\}}_{c_2}\\
&\qquad +(n)_4\underbrace{\mathrm{E}\left\{K_\sigma(\mathbf{y}_1,\mathbf{y}_2)K_\sigma(\mathbf{y}_3,\mathbf{y}_4)\right\}}_{c_3}\Bigg]-(n-1)^2\left\{\frac{n}{(n)_2}\left(nu_3-1\right)\right\}^{2p}, \qquad \text{(A.3)}
\end{aligned}$$

where c_1, c_2 and c_3 are further evaluated in Equations (A.4),(A.5) and (A.6) below.

$$\begin{aligned}
c_1 &= \mathrm{E}\left\{K_\sigma(\mathbf{y}_1,\mathbf{y}_2)K_\sigma(\mathbf{y}_1,\mathbf{y}_2)\right\} = \mathrm{E}\left\{K^2_\sigma(\mathbf{y}_1,\mathbf{y}_2)\right\} = \left[\mathrm{E}\left\{K_{\frac{\sigma}{\sqrt{2}}}(y_1^{(1)},y_2^{(1)})\right\}\right]^p\\
&= \left[\frac{1}{(n)_2}\sum_{1\le i\ne j\le n}K_{\frac{\sigma}{\sqrt{2}}}\left(\frac{i}{n},\frac{j}{n}\right)\right]^p = \left\{\frac{n}{(n)_2}\left(nu_1-1\right)\right\}^p \qquad \text{(A.4)}
\end{aligned}$$

$$
\begin{aligned}
c_2 &= \mathrm{E}\left\{K_\sigma(\mathbf{y}_1,\mathbf{y}_2)K_\sigma(\mathbf{y}_2,\mathbf{y}_3)\right\} = \left[\mathrm{E}\left\{K_\sigma(y_1^{(1)},y_2^{(1)})K_\sigma(y_2^{(1)},y_3^{(1)})\right\}\right]^p \\
&= \left[\frac{1}{(n)_3}\sum_{\substack{1\le i_1,i_2,i_3\le n\\ i_1\neq i_2\neq i_3\neq i_1}} K_\sigma\left(\frac{i_1}{n},\frac{i_2}{n}\right)K_\sigma\left(\frac{i_2}{n},\frac{i_3}{n}\right)\right]^p \\
&= \left[\frac{1}{(n)_3}\left\{\sum_{1\le i_1,i_2,i_3\le n} - \sum_{i_1=i_2} - \sum_{i_2=i_3} - \sum_{i_1=i_3} + \sum_{i_1=i_2,i_2=i_3} + \sum_{i_2=i_3,i_3=i_1} + \sum_{i_3=i_1,i_1=i_2}\right.\right. \\
&\qquad \left.\left. - \sum_{i_1=i_2,i_2=i_3,i_3=i_1}\right\}K_\sigma\left(\frac{i_1}{n},\frac{i_2}{n}\right)K_\sigma\left(\frac{i_2}{n},\frac{i_3}{n}\right)\right]^p \\
&= \left[\frac{1}{(n)_3}\left\{n^3u_2 - n^2u_3 - n^2u_3 - n^2u_1 + n + n + n - n\right\}\right]^p \\
&= \left[\frac{n}{(n)_3}\left\{n^2u_2 - n(2u_3+u_1) + 2\right\}\right]^p
\end{aligned} \tag{A.5}
$$

$$
\begin{aligned}
c_3 &= \mathrm{E}\left\{K_\sigma(\mathbf{y}_1,\mathbf{y}_2)K_\sigma(\mathbf{y}_3,\mathbf{y}_4)\right\} = \left[\mathrm{E}\left\{K_\sigma(y_1^{(1)},y_1^{(2)})K_\sigma(y_1^{(3)},y_1^{(4)})\right\}\right]^p \\
&= \left[\frac{1}{(n)_4}\sum_{\substack{1\le i_1,i_2,i_3,i_4\le n\\ \text{all } i_j\text{'s are distinct}}} K_\sigma\left(\frac{i_1}{n},\frac{i_2}{n}\right)K_\sigma\left(\frac{i_3}{n},\frac{i_4}{n}\right)\right]^p \\
&= \left[\frac{1}{(n)_4}\left\{\sum_{1\le i_1,i_2,i_3,i_4\le n} -2\sum_{i_1=i_2} -4\sum_{i_1=i_3} +12\sum_{i_1=i_2,i_2=i_3} + \sum_{i_1=i_2,i_3=i_4} +2\sum_{i_1=i_3,i_2=i_4}\right.\right. \\
&\qquad \left.\left. - 4\sum_{\substack{i_1=i_2,i_2=i_3\\ i_3=i_1}} -(16-\binom{6}{4}+\binom{6}{5}-\binom{6}{6})\sum_{\text{all } i_j\text{'s are equal}}\right\}K_\sigma\left(\frac{i_1}{n},\frac{i_2}{n}\right)K_\sigma\left(\frac{i_3}{n},\frac{i_4}{n}\right)\right]^p \\
&= \left[\frac{1}{(n)_4}\left\{n^4u_3^2 - 2n^3u_3 - 4n^3u_2 + 12n^2u_3 + n^2 + 2n^2u_1 - 4n^2u_3 - 6n\right\}\right]^p \\
&= \left[\frac{n}{(n)_4}\left\{n^3u_3^2 - 2n^2(u_3+2u_2) + n(8u_3+2u_1+1) - 6\right\}\right]^p
\end{aligned} \tag{A.6}
$$

Also we have,

$$
\begin{aligned}
&\mathrm{Cov}(ns_1, ns_2) = n^2\mathrm{Cov}\left(\frac{1}{n^2}\sum_{1\le i,j\le n}K_\sigma(\mathbf{y}_i,\mathbf{y}_j), \frac{1}{n}\sum_{i=1}^n\prod_{j=1}^p\frac{1}{n}\sum_{l=1}^n K_\sigma\left(y_i^{(j)},\frac{l}{n}\right)\right) \\
&= n^2\mathrm{Cov}\left(\frac{1}{n^2}\sum_{1\le i\neq j\le n}K_\sigma(\mathbf{y}_i,\mathbf{y}_j), \frac{1}{n}\sum_{i=1}^n\prod_{j=1}^p\frac{1}{n}\sum_{l=1}^n K_\sigma\left(y_i^{(j)},\frac{l}{n}\right)\right) \\
&= n^2\mathrm{E}\left[\frac{1}{n^2}\sum_{1\le i\neq j\le n}K_\sigma(\mathbf{y}_i,\mathbf{y}_j)\times\frac{1}{n}\sum_{i=1}^n\prod_{j=1}^p\frac{1}{n}\sum_{l=1}^n K_\sigma\left(y_i^{(j)},\frac{l}{n}\right)\right] \\
&\quad - n^2\mathrm{E}\left[\frac{1}{n^2}\sum_{1\le i\neq j\le n}K_\sigma(\mathbf{y}_i,\mathbf{y}_j)\right]\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\prod_{j=1}^p\frac{1}{n}\sum_{l=1}^n K_\sigma\left(y_i^{(j)},\frac{l}{n}\right)\right]
\end{aligned}
$$

$$= \frac{1}{n}\Bigg[2(n)_2 \underbrace{\mathrm{E}\Big\{K_\sigma(\mathbf{y}_1,\mathbf{y}_2)\prod_{j=1}^{p}\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_2^{(j)},\frac{l}{n}\right)\Big\}}_{c_4}$$
$$+ (n)_3 \underbrace{\mathrm{E}\Big\{K_\sigma(\mathbf{y}_1,\mathbf{y}_2)\prod_{j=1}^{p}\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_3^{(j)},\frac{l}{n}\right)\Big\}}_{c_5}\Bigg] - n(n-1)\left\{\frac{n}{(n)_2}\left(nu_3-1\right)\right\}^p u_3^p, \quad \text{(A.7)}$$

where c_4 and c_5 are further evaluated in Equations (A.8) and (A.9) below.

$$c_4 = \mathrm{E}\left\{K_\sigma(\mathbf{y}_1,\mathbf{y}_2)\prod_{j=1}^{p}\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_2^{(j)},\frac{l}{n}\right)\right\} = \left[\mathrm{E}\left\{K_\sigma(y_1^{(1)},y_2^{(1)})\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_2^{(1)},\frac{l}{n}\right)\right\}\right]^p$$
$$= \left[\frac{1}{n(n)_2}\sum_{1\le i\ne j\le n}\sum_{l=1}^{n}K_\sigma\left(\frac{i}{n},\frac{j}{n}\right)K_\sigma\left(\frac{j}{n},\frac{l}{n}\right)\right]^p$$
$$= \left[\frac{1}{n(n)_2}\left\{\sum_{1\le i,j\le n}\sum_{l=1}^{n} - \sum_{1\le i=j\le n}\sum_{l=1}^{n}\right\}K_\sigma\left(\frac{i}{n},\frac{j}{n}\right)K_\sigma\left(\frac{j}{n},\frac{l}{n}\right)\right]^p$$
$$= \left\{\frac{1}{n(n)_2}(n^3u_2 - n^2u_3)\right\}^p = \left\{\frac{n}{(n)_2}(nu_2-2u_3)\right\}^p \quad \text{(A.8)}$$

$$c_5 = \mathrm{E}\left\{K_\sigma(\mathbf{y}_1,\mathbf{y}_2)\prod_{j=1}^{p}\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_3^{(j)},\frac{l}{n}\right)\right\} = \left[\mathrm{E}\left\{K_\sigma(y_1^{(1)},y_1^{(2)})\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_3^{(1)},\frac{l}{n}\right)\right\}\right]^p$$
$$= \left[\frac{1}{(n)_3}\sum_{1\le i_1\ne i_2\ne i_3\le n}\sum_{l=1}^{n}K_\sigma\left(\frac{i_1}{n},\frac{i_2}{n}\right)K_\sigma\left(\frac{i_3}{n},\frac{l}{n}\right)\right]^p$$
$$= \Bigg[\frac{1}{(n)_3}\Bigg\{\sum_{1\le i_1,i_2,i_3\le n}\sum_{l=1}^{n} - \sum_{i_1=i_2}\sum_{l=1}^{n} - 2\sum_{i_1=i_3}\sum_{l=1}^{n}$$
$$+\left(\binom{3}{2}-\binom{3}{3}\right)\sum_{\text{all } i_j\text{'s are equal}}\sum_{l=1}^{n}\Bigg\}K_\sigma\left(\frac{i_1}{n},\frac{i_2}{n}\right)K_\sigma\left(\frac{i_3}{n},\frac{l}{n}\right)\Bigg]^p$$
$$= \left\{\frac{1}{(n)_3}\left(n^4u_3^2 - n^2u_3 - 2n^2u_2 + 2nu_3\right)\right\}^p = \left\{\frac{n}{(n)_3}\left(n^3u_3^2 - n^2(2u_2+u_3) + 2u_3\right)\right\}^p \quad \text{(A.9)}$$

Again, under the null hypothesis of mutual independence,

$$\mathrm{Var}(ns_2) = \mathrm{E}\left[\sum_{i=1}^{n}\prod_{j=1}^{p}\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_i^{(j)},\frac{l}{n}\right)\right]^2 - \mathrm{E}^2\left[\sum_{i=1}^{n}\prod_{j=1}^{p}\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_i^{(j)},\frac{l}{n}\right)\right]$$
$$= n\mathrm{E}\left[\prod_{j=1}^{p}\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_j^{(1)},\frac{l}{n}\right)\right]^2 + (n)_2\mathrm{E}\left[\prod_{j=1}^{p}\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_1^{(j)},\frac{l}{n}\right)\prod_{j=1}^{p}\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_2^{(j)},\frac{l}{n}\right)\right]$$
$$- n^2\left[\mathrm{E}\left\{\frac{1}{n}\sum_{l=1}^{n}K_\sigma\left(y_1^{(1)},\frac{l}{n}\right)\right\}\right]^{2p}$$

$$
\begin{aligned}
&= n\mathrm{E}^p\left\{\frac{1}{n}\sum_{l=1}^{n} K_\sigma\left(y_1^{(1)}, \frac{l}{n}\right)\right\}^2 + (n)_2\mathrm{E}^p\left\{\frac{1}{n}\sum_{l=1}^{n} K_\sigma\left(y_1^{(j)}, \frac{l}{n}\right)\frac{1}{n}\sum_{l=1}^{n} K_\sigma\left(y_2^{(j)}, \frac{l}{n}\right)\right\} - n^2u_3^{2p} \\
&= nu_2^p + (n)_2\left[\frac{1}{n^2(n)_2}\sum_{1\le i_1,i_2\le n}\sum_{1\le j_1\ne j_2\le n} K_\sigma\left(\frac{i_1}{n}, \frac{j_1}{n}\right)K_\sigma\left(\frac{i_2}{n}, \frac{j_2}{n}\right)\right]^p - n^2u_3^{2p} \\
&= nu_2^p + (n)_2\left[\frac{1}{n^2(n)_2}\left\{\sum_{i_1,i_2,j_1,j_2} - \sum_{i_1,i_2,j_1=j_2}\right\} K_\sigma\left(\frac{i_1}{n}, \frac{j_1}{n}\right)K_\sigma\left(\frac{i_2}{n}, \frac{j_2}{n}\right)\right]^p - n^2u_3^{2p} \\
&= nu_2^p + (n)_2\left\{\frac{1}{n^2(n)_2}(n^4u_3^2 - n^3u_2)\right\}^p - n^2u_3^{2p} \\
&= nu_2^p + (n)_2\left\{\frac{n}{(n)_2}(nu_3^2 - u_2)\right\}^p - n^2u_3^{2p} \qquad \text{(A.10)}
\end{aligned}
$$

From (A.3), (A.7) and (A.10), we get

$$
\begin{aligned}
&\mathrm{Var}(n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)) = \mathrm{Var}(ns_1 - 2ns_2 + nv_3) = \mathrm{Var}(ns_1) - 4\mathrm{Cov}(ns_1, ns_2) + 4\mathrm{Var}(ns_2) \\
&= \frac{1}{n^2}\left[2(n)_2c_1 + 4(n)_3c_2 + (n)_4c_3\right] - (n-1)^2\left\{\frac{n}{(n)_2}(nu_3 - 1)\right\}^{2p} - \frac{4}{n}\left[2(n)_2c_4 + (n)_3c_5\right] \\
&\quad + 4n(n-1)\left\{\frac{n}{(n)_2}(nu_3 - 1)\right\}^d u_3^p + 4\left[nu_2^p + (n)_2\left\{\frac{n}{(n)_2}(nu_3^2 - u_2)\right\}^p\right] - 4n^2u_3^{2p} \\
&= \frac{1}{n^2}\left[2(n)_2c_1 + 4(n)_3c_2 + (n)_4c_3\right] - \frac{4}{n}\left[2(n)_2c_4 + (n)_3c_5\right] \\
&\quad + 4\left[nu_2^p + (n)_2\left\{\frac{n}{(n)_2}(nu_3^2 - u_2)\right\}^p\right] - \left[(n-1)\left\{\frac{n}{(n)_2}(nu_3 - 1)\right\}^p - 2nu_3^p\right]^2. \qquad \square
\end{aligned}
$$

Proposition A.2. *Under the null hypothesis of mutual independence,*

$$
\begin{aligned}
&\lim_{n\to\infty}\mathrm{E}[n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)] = 1 + (p-1)w_3^p - dw_3^{p-1}, \\
&\lim_{n\to\infty}\mathrm{Var}[n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)] = 2\Big[w_1^p + 2(p-1)w_2^p - 2pw_2^{p-1}w_1 + pw_3^{2p-2}w_1 \\
&\qquad\qquad -(p-1)w_3^{2p} + p(p-1)w_3^{2p-4}(w_3^2 - w_2)^2\Big],
\end{aligned}
$$

where $w_1 = \kappa\left(\frac{\sigma}{\sqrt{2}}\right)$, $w_2 = \int_0^1 \lambda^2(u, \sigma)\,du$, *and* $w_3 = \kappa(\sigma)$, *for* $\kappa(\cdot)$ *and* $\lambda(\cdot)$ *being defined in Theorem 2.1.*

Proof. From definition of w_1, w_2 and w_3, it is easy to check that $w_1 = \lim_{n\to\infty} u_1$, $w_2 = \lim_{n\to\infty} u_2$ and $w_3 = \lim_{n\to\infty} u_3$. Then observe that for a natural number $k \le n$ and real numbers $a_0, a_1, \cdots, a_{k-1}$, we can define function $g : \mathbb{R}^k \mapsto \mathbb{R}$ such that

$$
g(a_0, a_1, \cdots, a_{k-1}) := (n)_k\left\{\frac{n}{(n)_k}\left(a_0n^{k-1} + a_1n^{k-2} + \cdots + a_{k-1}\right)\right\}^p.
$$

Using basic algebra, it can be verified that for $k = 2$,

$$
g(a_0, a_1) = a_0^pn^2 + \left\{(p-1)a_0^p + pa_0^{p-1}a_1\right\}n + \frac{p(p-1)}{2}a_0^{p-2}(a_0 + a_1)^2 + \mathcal{O}\left(\frac{1}{n}\right).
$$

Similarly, for $k = 3$ and $k = 4$, we have

$$g(a_0, a_1, a_2) = a_0^p n^3 + \left\{3(p-1)a_0^p + pa_0^{p-1}a_1\right\} n^2 + \left\{pa_0^{p-1}a_2 - 2(p-1)a_0^p + \frac{p(p-1)}{2}a_0^{p-2}(3a_0 + a_1)^2\right\} n + \mathcal{O}(1), \text{ and}$$

$$g(a_0, a_1, a_2, a_4) = a_0^p n^4 + \left\{6(p-1)a_0^p + pa_0^{p-1}a_1\right\} n^3 + \left\{pa_0^{p-1}a_2 - 11(p-1)a_0^p + \frac{p(p-1)}{2}a_0^{p-2}(6a_0 + a_1)^2\right\} n^2 + \mathcal{O}(n).$$

Now, $\mathrm{E}[n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)] = 1 + (n-1)\left\{\frac{n}{(n)_2}(nu_3 - 1)\right\}^p - nu_3^p = 1 + \frac{1}{n}g(u_3, -1) - nu_3^p.$

Taking limit at both sides of this equation, we get

$$\lim_{n\to\infty} \mathrm{E}[n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)] = 1 + (p-1)w_3^p - dw_3^{p-1}.$$

Again, from Proposition A.1, we have

$$\begin{aligned}
&\mathrm{Var}[n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)] = \frac{1}{n^2}\left[2(n)_2c_1 + 4(n)_3c_2 + (n)_4c_3\right] - \frac{4}{n}\left[2(n)_2c_4 + (n)_3c_5\right] \\
&+ 4\left[nu_2^p + (n)_2\left\{\frac{n}{(n)_2}(nu_3^2 - u_2)\right\}^p\right] - \left[(n-1)\left\{\frac{n}{(n)_2}(nu_3 - 1)\right\}^p - 2nu_3^p\right]^2 \\
&= \frac{1}{n^2}\left[2g(u_1, -1) + 4g(u_2, -(2u_3 + u_1), 2) + g(u_3^2, -2(u_3 + 2u_2), (8u_3 + 2u_1 + 1), -6)\right] \\
&- \frac{4}{n}\left[2g(u_2, -u_3) + g(u_3^2, -(2u_2 + u_3), 2u_3)\right] + 4\left[nu_2^p + g(u_3^2, -u_2)\right] - \left[\frac{g(u_3, -1)}{n} - 2nu_3^p\right]^2 \\
&= \frac{1}{n^2}\Bigg[2n^2u_1^p + 4n^3u_2 - 4n^2\left\{3(p-1)u_2^p - pu_2^{p-1}(2u_3 + u_1)\right\} + n^4u_3^{2p} + n^3\left\{6(p-1)u_3^{2p}\right. \\
&\left.-2pu_3^{2p-2}(u_3 + 2u_2)\right\} + n^2\left\{pu_3^{2p-2}(8u_3 + 2u_1 + 1) - 11(p-1)u_3^{2p}\right. \\
&\left.+ \frac{p(p-1)}{2}u_3^{2p-4}\left(6u_3^2 + 6(p-1)u_3^{2p} - 2pu_3^{2p-2}(u_3 + 2u_2)\right)^2\right\}\Bigg] \\
&- \frac{4}{n}\left[2n^2u_2^p + n\left\{(p-1)u_2^p - pu_2^{p-1}u_3\right\}\right] + 4\Bigg[nu_2^p + n^2u_3^{2p} + +n\left\{(p-1)u_3^{2p} - pu_3^{2p-2}u_2\right\} \\
&+ \frac{p(p-1)}{2}u_3^{2p-4}(u_3^2 - u_2)^2\Bigg] - \left[nu_3^p + \left\{(p-1)u_3^p - pu_3^{p-1}\right\} - 2nu_3^p\right]^2 + \mathcal{O}\left(\frac{1}{n}\right) \\
&= 2\left[u_1^p + 2(p-1)u_2^p - 2pu_2^{p-1}u_1 + pu_3^{2p-2}u_1 - (p-1)u_3^{2p} + p(p-1)u_3^{2p-4}(u_3^2 - u_2)^2\right] + \mathcal{O}\left(\frac{1}{n}\right)
\end{aligned}$$

Taking limit at both sides of the above equation, we get

$$\lim_{n\to\infty} \mathrm{Var}[n\gamma^2_{K_\sigma}(\mathsf{C}_n, \Pi_n)] = 2\left[w_1^p + 2(p-1)w_2^p - 2pw_2^{p-1}w_1 + pw_3^{2p-2}w_1 - (p-1)w_3^{2p} + p(p-1)w_3^{2p-4}(w_3^2 - w_2)^2\right]. \qquad \square$$

Appendix B

Brief Descriptions of the Existing Tests Used in Different Chapters

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be n independent observations on a d-dimensional random vector $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)})$ with sub-vectors $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ of dimensions $d_1, d_2, \ldots, d_p$, respectively $(d_1 + d_2 + \cdots + d_p = d)$. Based on these observations, we want to test the null hypothesis $\mathbb{H}_0$, which states that $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(p)}$ are mutually independent. Here we describe some of the existing methods that we have used for this purpose in different chapters of this thesis.

Tests based on generalized versions of Spearman's ρ, Kendall's τ, Blomqvist's β and Hoeffding's ϕ statistics

Here we consider all sub-vectors to be continuous and one-dimensional (i.e., $d_1 = d_2 = \ldots = d_p = 1$). For any fixed $j = 1, 2, \ldots, p$ and for $i = 1, 2, \ldots, n$, define $r_i^{(j)}$ as the rank of $x_i^{(j)}$ (the j-th component of $\mathbf{x}_i$) in the set $\{x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}\}$ to get $\mathbf{r}_i = (r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(p)})$, the coordinate-wise rank of $\mathbf{x}_i$. Now, consider the normalized rank vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, where $\mathbf{y}_i = \mathbf{r}_i/n$ for $i = 1, 2, \ldots, n$.

The generalized versions of Spearman's ρ (Schmid and Schmidt, 2007), Kendall's τ (Nelsen, 2002) and Blomqvist's β (Úbeda-Flores, 2005) statistics are given by

$$T_{\text{Spearman}} = \frac{p+1}{2^p - p - 1} \left\{ \frac{2^p}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} (1 - y_i^{(j)}) - 1 \right\},$$

$$T_{\text{Kendall}} = \frac{1}{2^{p-1} - 1} \left\{ \frac{2^p}{n^2} \sum_{i,k=1}^{n} \prod_{j=1}^{p} \mathbb{I}[x_i^{(j)} \leq x_k^{(j)}] - 1 \right\} \text{ and}$$

$$T_{\text{Blomqvist}} = \frac{1}{2^{p-1}-1}\left\{\frac{2^{p-1}}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}\mathbb{I}\left[y_i^{(j)} \le \frac{1}{2}\right] + \frac{2^{p-1}}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}\mathbb{I}\left[y_i^{(j)} > \frac{1}{2}\right] - 1\right\},$$

respectively. For each of the above tests, we reject $\mathbb{H}_0$ if the observed values of the test statistic is too large or too small.

The generalized versions of Hoeffding's ϕ statistic (Gaißer *et al.*, 2010) is given by

$$T^2_{\text{Hoeffding}} = h(p,n)\left[\frac{1}{n^2}\sum_{i,k=1}^{n}\prod_{j=1}^{p}(1-\max\{y_i^{(j)}, y_k^{(j)}\}) + \left\{\frac{(n-1)(2n-1)}{6n^2}\right\}^p - \frac{2}{2^p n}\sum_{i=1}^{n}\prod_{j=1}^{p}\left\{1-(y_i^{(j)})^2 - \frac{1-y_i^{(j)}}{n}\right\}\right],$$

$$\text{where } h(p,n)^{-1} = \frac{1}{n^2}\sum_{i,k=1}^{n}\prod_{j=1}^{p}\left(1-\max\left\{\frac{i}{n},\frac{k}{n}\right\}\right) + \left\{\frac{(n-1)(2n-1)}{6n^2}\right\}^p - \frac{2}{n}\sum_{i=1}^{n}\left\{\frac{n(n-1)-i(i-1)}{2n^2}\right\}^p.$$

The null hypothesis is rejected for large values of the test statistic $T^2_{\text{Hoeffding}}$.

These four tests have the distribution-free property. If the sample size is large, cut-offs can also be computed based on the large sample distributions of the test statistics.

Genest test (Genest *et al.*, 2019)

This test also deals with one-dimensional sub-vectors (variables), but it does not need them to be continuous. For $i,k \in \{1,2,\ldots,n\}$ and for $j = 1,2,\ldots,p$, define

$$I_{i,k}^{(j)} = \frac{1}{6}\sum_{l=1}^{n}\Big\{2\mathbb{I}[x_i^{(j)} \le x_l^{(j)}]\mathbb{I}[x_k^{(j)} \le x_l^{(j)}] + \mathbb{I}[x_i^{(j)} \le x_l^{(j)}]\mathbb{I}[x_k^{(j)} < x_l^{(j)}] + \mathbb{I}[x_i^{(j)} < x_l^{(j)}]\mathbb{I}[x_k^{(j)} \le x_l^{(j)}] + 2\mathbb{I}[x_i^{(j)} < x_l^{(j)}]\mathbb{I}[x_k^{(j)} < x_l^{(j)}]\Big\}.$$

The null hypothesis $\mathbb{H}_0$ is rejected for large values of

$$T_{\text{Genest}} = \frac{1}{n}\sum_{i,k=1}^{n}\prod_{j=1}^{p} I_{i,k}^{(j)} + \frac{n}{3^p} - \frac{2}{n^p}\sum_{i=1}^{n}\prod_{j=1}^{p}\sum_{k=1}^{n} I_{i,k}^{(j)}.$$

dHSIC test (Pfister *et al.*, 2018)

This test can be used for testing independence among p ($p \ge 2$) random vectors of arbitrary dimensions when the sample size is greater than or equal to $2p$. For $j = 1,2,\ldots,p$, let

$k^{(j)} : \mathbb{R}^{d_j} \times \mathbb{R}^{d_j} \mapsto \mathbb{R}$ be a continuous, bounded, positive semi-definite kernel. Assume that the tensor products of these kernels $k^{(1)} \otimes k^{(2)} \otimes \cdots \otimes k^{(p)}$ is a characteristic kernel (for more details, see Sriperumbudur *et al.*, 2010). In our experiments, we used the Gaussian kernel $k^{(j)}(\mathbf{u}^{(j)}, \mathbf{v}^{(j)}) = \exp\left(-\frac{\|\mathbf{u}^{(j)} - \mathbf{v}^{(j)}\|^2}{2\sigma_j^2}\right)$, for $2\sigma_j^2$ being the median of all pairwise distances of the form $\|\mathbf{x}_s^{(j)} - \mathbf{x}_t^{(j)}\|^2$, where $s \neq t \in \{1, 2, \ldots, n\}$. The null hypothesis is rejected for large values of the test statistic

$$T_{\text{dHSIC}} = \frac{1}{n^2} \sum_{1 \leq i_1, i_2 \leq n} \prod_{j=1}^{p} k^{(j)}(x_{i_1}^{(j)}, x_{i_2}^{(j)}) + \frac{1}{n^{2p}} \sum_{1 \leq i_1, i_2, \ldots, i_{2p} \leq n} \prod_{j=1}^{p} k^{(j)}(x_{i_{2j-1}}^{(j)}, x_{i_{2j}}^{(j)}) - \frac{2}{n^{p+1}} \sum_{1 \leq i_1, i_2, \ldots, i_{p+1} \leq n} \prod_{j=1}^{p} k^{(j)}(x_{i_1}^{(j)}, x_{i_{j+1}}^{(j)}).$$

Here also, the cut-off can be chosen based on the asymptotic null distribution of the test statistic. In the case of small sample size, conditional test based on the permutation principle can be used.

JdCov and rank-JdCov tests (Chakraborty and Zhang, 2019)

These tests can be used for testing independence among several random vectors of arbitrary dimensions. For $j = 1, 2, \ldots, p$ and $k, l \in \{1, 2, \ldots, n\}$, define $U_{k,l}^{(j)} = \frac{1}{n} \sum_{s=1}^{n} \|\mathbf{x}_k^{(j)} - \mathbf{x}_s^{(j)}\| + \frac{1}{n} \sum_{t=1}^{n} \|\mathbf{x}_l^{(j)} - \mathbf{x}_t^{(j)}\| - \|\mathbf{x}_k^{(j)} - \mathbf{x}_l^{(j)}\| - \frac{1}{n^2} \sum_{s,t=1}^{n} \|\mathbf{x}_s^{(j)} - \mathbf{x}_t^{(j)}\|$. For any constant $c > 0$ (we used $c = 1$), the JdCov statistic and its scaled version are given by

$$T_{\text{JdCov}} = \frac{1}{n^2} \sum_{k,l=1}^{n} \prod_{j=1}^{p} \left(U_{k,l}^{(j)} + c\right) - c^p \quad \text{and}$$

$$T_{\text{JdCov}_s} = \frac{1}{n^2} \sum_{k,l=1}^{n} \prod_{j=1}^{p} \left(\frac{n^2 U_{k,l}^{(j)}}{\sum_{s,t=1}^{n} \left(U_{s,t}^{(j)}\right)^2} + c \right) - c^p,$$

respectively. In this thesis, we used the scaled version of the statistic for the JdCov test.

Replacing $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ by their corresponding normalized coordinate-wise rank vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ (defined earlier in this Chapter), one gets rank versions of the $U_{k,l}^{(j)}$s. Using them in the definition of T_{JdCov}, one gets the test statistic for the rank-JdCov test.

Both, JdCov and rank-JdCov tests, reject $\mathbb{H}_0$ for large values of the test statistics. Cut-offs can be computed either using the large sample distributions of the test statistics or using the permutation principle.

HHG test (Heller *et al.*, 2013)

This test can be used for testing independence between two random vectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ of arbitrary dimensions. For each $i \neq j \in \{1, 2, \ldots, n\}$ and each $k \in \{1, 2, \ldots, n\} \setminus \{i, j\}$, depending whether $\|\mathbf{x}_i^{(1)} - \mathbf{x}_k^{(1)}\| \leq \|\mathbf{x}_i^{(1)} - \mathbf{x}_j^{(1)}\|$ and $\|\mathbf{x}_i^{(2)} - \mathbf{x}_k^{(2)}\| \leq \|\mathbf{x}_i^{(2)} - \mathbf{x}_j^{(2)}\|$, put $\mathbf{x}_k$ in one of the four cells of a 2×2 contingency table and compute the Pearson's Chi Squared statistics $S_{i,j}$ based on those cell frequencies. This test rejects $\mathbb{H}_0$ for large values of the test statistic

$$T_{\text{HHG}} = \sum_{1 \leq i \neq j \leq n} S_{i,j}.$$

The cut-off can be computed using permutation method.

Bibliography

Anderson, T. W. (2003) *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New York.

Bartlett, P. L. and Mendelson, S. (2003) Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, **3**, 463–482.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165–1188.

Beran, R., Bilodeau, M. and Lafaye de Micheaux, P. (2007) Nonparametric tests of independence between random vectors. *Journal of Multivariate Analysis*, **98**, 1805–1824.

Bilodeau, M. and Lafaye de Micheaux, P. (2005) A multivariate empirical characteristic function test of independence with normal marginals. *Journal of Multivariate Analysis*, **95**, 345–369.

Bilodeau, M. and Nangue, A. G. (2017) Tests of mutual or serial independence of random vectors with applications. *Journal of Machine Learning Research*, **18**, 2518–2557.

Biswas, M., Sarkar, S. and Ghosh, A. K. (2016) On some exact distribution-free tests of independence between two random vectors of arbitrary dimensions. *Journal of Statistical Planning and Inference*, **175**, 78–86.

Blomqvist, N. (1950) On a measure of dependence between two random variables. *The Annals of Mathematical Statistics*, **21**, 593–600.

Böttcher, B., Keller-Ressel, M., Schilling, R. L. *et al.* (2019) Distance multivariance: new dependence measures for random vectors. *The Annals of Statistics*, **47**, 2757–2789.

Breitenberger, E. (1963) Analogues of the normal distribution on the circle and the sphere. *Biometrika*, **50**, 81–88.

Brill, B. and Kaufman, S. (2019) *HHG: Heller-Heller-Gorfine Tests of Independence and Equality of Distributions*. URL https://CRAN.R-project.org/package=HHG. R package version 2.3.2.

Brooks, T. F., Pope, D. S. and Marcolini, M. A. (1989) Airfoil self-noise and prediction. Tech. rep., NASA Langley Research Center.

Cameron, A. C., Trivedi, P. K., Milne, F. and Piggott, J. (1988) A microeconometric model of the demand for health care and health insurance in Australia. *The Review of Economic Studies*, **55**, 85–106.

Chakraborty, S. and Zhang, X. (2019) Distance metrics for measuring joint dependence with application to causal inference. *Journal of the American Statistical Association*, **114**, 1638–1650.

Cuesta-Albertos, J. A. and Febrero-Bande, M. (2010) A simple multiway ANOVA for functional data. *Test*, **19**, 537–557.

De Jonge, J., Dormann, C., Janssen, P. P., Dollard, M. F., Landeweerd, J. A. and Nijhuis, F. J. (2001) Testing reciprocal relationships between job characteristics and psychological well-being: a cross-lagged structural equation model. *Journal of Occupational and Organizational Psychology*, **74**, 29–46.

Draper, D. (1995) Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B*, **57**, 45–70.

Dutta, S., Sarkar, S. and Ghosh, A. K. (2016) Multi-scale classification using localized spatial depth. *Journal of Machine Learning Research*, **17**, 7657–7686.

Erästö, P. and Holmström, L. (2005) Bayesian multiscale smoothing for making inferences about features in scatterplots. *Journal of Computational and Graphical Statistics*, **14**, 569–589.

Fan, Y., Lafaye de Micheaux, P., Penev, S. and Salopek, D. (2017) Multivariate nonparametric test of independence. *Journal of Multivariate Analysis*, **153**, 189–210.

Fang, K. T., Kotz, S. and Ng., K. W. (1990) *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC Press, New York.

Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis: Theory and Practice*. Springer-Verlag, New York.

Ferreira, J. C. and Menegatto, V. A. (2012) An extension of Mercer's theory to L^p. *Positivity*, **16**, 197–212.

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

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2009a) Kernel dimension reduction in regression. *The Annals of Statistics*, **37**, 1871–1905.

Fukumizu, K., Gretton, A., Lanckriet, G. R., Schölkopf, B. and Sriperumbudur, B. K. (2009b) Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22* (eds. Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta), 1750–1758.

Gaißer, S., Ruppert, M. and Schmid, F. (2010) A multivariate version of Hoeffding's phi-square. *Journal of Multivariate Analysis*, **101**, 2571–2586.

Genest, C., Nešlehová, J. G. and Rémillard, B. (2017) Asymptotic behavior of the empirical multilinear copula process under broad conditions. *Journal of Multivariate Analysis*, **159**, 82–110.

Genest, C., Nešlehová, J. G., Rémillard, B. and Murphy, O. A. (2019) Testing for independence in arbitrary distributions. *Biometrika*, **106**, 47–68.

Ghosh, A. K., Chaudhuri, P. and Sengupta, D. (2006) Classification using kernel density estimates: multiscale analysis and visualization. *Technometrics*, **48**, 120–132.

Ghoudi, K., Kulperger, R. J. and Rémillard, B. (2001) A nonparametric test of serial independence for time series and residuals. *Journal of Multivariate Analysis*, **79**, 191–218.

Gieser, P. W. and Randles, R. H. (1997) A nonparametric test of independence between two vectors. *Journal of the American Statistical Association*, **92**, 561–567.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012) A kernel two-sample test. *Journal of Machine Learning Research*, **13**, 723–773.

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 20* (eds. J. C. Platt, D. Koller, Y. Singer and S. T. Roweis), 585–592.

Haberman, S. J. (1976) Generalized residuals for log-linear models. In *Proceedings 9th International Biometric Conference, Boston*, 104–122.

Heller, R., Gorfine, M. and Heller, Y. (2012) A class of multivariate distribution-free tests of independence based on graphs. *Journal of Statistical Planning and Inference*, **142**, 3097–3106.

Heller, R. and Heller, Y. (2016) Multivariate tests of association based on univariate tests. In *Advances in Neural Information Processing Systems 29* (eds. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett), 208–216.

Heller, R., Heller, Y. and Gorfine, M. (2013) A consistent multivariate test of association based on ranks of distances. *Biometrika*, **100**, 503–510.

Hochberg, Y. (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–802.

Hoeffding, W. (1948) A non-parametric test of independence. *The Annals of Mathematical Statistics*, **19**, 546–557.

Hsieh, T.-J., Chang, S.-H. and Tai, J. J. (2014) A family-based robust multivariate association test using maximum statistic. *Annals of Human Genetics*, **78**, 117–128.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001) *Independent Component Analysis*. John Wiley & Sons, New York.

Jin, Z. and Matteson, D. S. (2018) Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete V-statistics. *Journal of Multivariate Analysis*, **168**, 304–322.

Joe, H. (1990) Multivariate concordance. *Journal of Multivariate Analysis*, **35**, 12–30.

Kendall, M. G. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–93.

Kibble, W. F. (1945) An extension of a theorem of Mehler's on Hermite polynomials. In *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 41, 12–15. Cambridge University Press, Cambridge.

Li, R., Zhong, W. and Zhu, L. (2012) Feature screening via distance correlation learning. *Journal of the American Statistical Association*, **107**, 1129–1139.

Lindeberg, T. (2013) *Scale-Space Theory in Computer Vision*. Springer Science & Business Media, Berlin.

Lopez-Paz, D., Hennig, P. and Schölkopf, B. (2013) The randomized dependence coefficient. In *Advances in Neural Information Processing Systems 26* (eds. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger), 1–9.

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

Mardia, K. V. and Jupp, P. E. (2009) *Directional Statistics*. John Wiley & Sons, New York.

Massart, P. (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, **18**, 1269–1283.

McDiarmid, C. (1989) On the method of bounded differences. *Surveys in Combinatorics*, **141**, 148–188.

McDonald, G. C. and Schwing, R. C. (1973) Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, **15**, 463–481.

Mukhopadhyay, S. and Ghosh, A. K. (2011) Bayesian multiscale smoothing in supervised and semi-supervised kernel discriminant analysis. *Computational Statistics and Data Analysis*, **55**, 2344–2353.

Nelsen, R. B. (1996) Nonparametric measures of multivariate association. In *Distributions with Fixed Marginals and Related Topics* (eds. L. Rüschendorf, B. Schweizer and M. D. Taylor), vol. 28 of *Lecture Notes–Monograph Series*, 223–232. Hayward, C A: Institute of Mathematical Statistics.

Nelsen, R. B. (2002) Concordance and copulas: a survey. In *Distributions with Given Marginals and Statistical Modelling* (eds. C. M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena), 169–177. Dordrecht: Springer Netherlands.

Nelsen, R. B. (2007) *An Introduction to Copulas*. Springer Science & Business Media, New York.

Newton, M. A. (2009) Introducing the discussion paper by Székely and Rizzo. *The Annals of Applied Statistics*, **3**, 1233–1235.

Peters, J., Mooij, J. M., Janzing, D. and Schölkopf, B. (2014) Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, **15**, 2009–2053.

Pfister, N., Bühlmann, P., Schölkopf, B. and Peters, J. (2018) Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B*, **80**, 5–31.

Pfister, N. and Peters, J. (2019) *dHSIC: Independence Testing via Hilbert Schmidt Independence Criterion*. URL https://CRAN.R-project.org/package=dHSIC. R package version 2.1.

Póczos, B., Ghahramani, Z. and Schneider, J. (2012) Copula-based kernel dependency measures. In *Proceedings of the 29th International Conference on Machine Learning* (eds. J. Langford and J. Pineau), 775–782. Edinburgh, Scotland.

Prim, R. C. (1957) Shortest connection networks and some generalizations. *The Bell System Technical Journal*, **36**, 1389–1401.

Ramsay, J. and Silverman, B. W. (2005) *Functional Data Analysis*. Springer, New York.

Rawat, R. and Sitaram, A. (2000) Injectivity sets for spherical means on $\mathbb{R}^n$ and on symmetric spaces. *Journal of Fourier Analysis and Applications*, **6**, 343–348.

Rényi, A. (1959) On measures of dependence. *Acta Mathematica Hungarica*, **10**, 441–451.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. and Sabeti, P. C. (2011) Detecting novel associations in large data sets. *Science*, **334**, 1518–1524.

Roy, A. (2020) Some copula-based tests of independence among several random variables having arbitrary probability distributions. *Stat*, To appear.

Roy, A., Das, K., Sarkar, S. and Ghosh, A. K. (2020a) Some tests of mutual independence among several random vectors using ranks of nearest neighbors. *Journal of Statistical Computation and Simulation*, Under revision.

Roy, A. and Ghosh, A. K. (2020) Some tests of independence baed on maximum mean discrepancy and ranks of nearest neighbors. *Statistics and Probability Letters*, To appear.

Roy, A., Ghosh, A. K., Goswami, A. and Murthy, C. A. (2020b) Some new copula based distribution-free tests of independence among several random variables. *Sankhya: Series A*, To appear.

Roy, A., Sarkar, S., Ghosh, A. K. and Goswami, A. (2020c) On some consistent tests of mutual independence among several random vectors of arbitrary dimensions. *Statistics and Computing*, Under revision.

Sarkar, S. and Ghosh, A. K. (2018) Some multivariate tests of independence based on ranks of nearest neighbors. *Technometrics*, **60**, 101–111.

Sarkar, S., Ghosh, A. K. and Goswami, A. (2020) A consistent test of independence between two random vectors of arbitrary dimensions. Tech. rep., Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata.

Schmid, F. and Schmidt, R. (2007) Multivariate extensions of Spearman's rho and related statistics. *Statistics and Probability Letters*, **77**, 407–416.

Schmid, F., Schmidt, R., Blumentritt, T., Gaißer, S. and Ruppert, M. (2010) Copula-based measures of multivariate association. In *Copula Theory and Its Applications* (eds. P. Jaworski, F. Durante, W. K. Hardle and T. Rychlik), 209–236. Berlin: Springer.

Sen, P. K. and Puri, M. L. (1971) *Nonparametric Methods in Multivariate Analysis*. John Wiley & Sons, New York.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

Shen, H., Jegelka, S. and Gretton, A. (2009) Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, **57**, 3498–3511.

Simon, N. and Tibshirani, R. (2014) Comment on 'detecting novel associations in large data sets' by Reshef et al, Science, 2011. *arXiv preprint arXiv:1401.7645*.

Spearman, C. (1904) The proof and measurement of association between two things. *The American Journal of Psychology*, **15**, 72–101.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B. and Lanckriet, G. R. G. (2010) Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, **11**, 1517–1561.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, **35**, 2769–2794.

Taskinen, S., Kankainen, A. and Oja, H. (2003) Sign test of independence between two random vectors. *Statistics and Probability Letters*, **62**, 9–21.

Taskinen, S., Oja, H. and Randles, R. H. (2005) Multivariate nonparametric tests of independence. *Journal of the American Statistical Association*, **100**, 916–925.

Tsukahara, H. (2005) Semiparametric estimation in copula models. *Canadian Journal of Statistics*, **33**, 357–375.

Úbeda-Flores, M. (2005) Multivariate versions of Blomqvist's beta and Spearman's footrule. *Annals of the Institute of Statistical Mathematics*, **57**, 781–788.

Um, Y. and Randles, R. H. (2001) A multivariate nonparametric test of independence among many vectors. *Journal of Nonparametric Statistics*, **13**, 699–708.

Witmer, J. (1997) *Data Analysis: An Introduction.* Prentice Hall, New Jersey.