SOME NONPARAMETRIC TESTS FOR HIGH-DIMENSIONAL AND FUNCTIONAL DATA

BILOL BANERJEE

Indian Statistical Institute, Kolkata
July 18, 2025

SOME NONPARAMETRIC TESTS FOR HIGH-DIMENSIONAL AND FUNCTIONAL DATA

BILOL BANERJEE

Thesis submitted to the Indian Statistical Institute
in partial fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy.
July 18, 2025

Indian Statistical Institute
203, B. T. Road, Kolkata, India.

Bilol Banerjee

Author of the Ph.D. thesis
Bilol Banerjee

Anil Kr. Ghosh

Supervisor of the Ph.D. thesis
Prof. Anil K. Ghosh

Certificate of Supervision

I hereby certify that the doctoral thesis titled "Some Nonparametric Tests for High-Dimensional and Functional Data", submitted by Mr. Bilol Banerjee, has been conducted under my supervision. I further declare that the research presented in this thesis is original and has not been submitted elsewhere for the award of any degree.

Prof. Anil K. Ghosh
Theoretical Statistics and Mathematics Unit,
Indian Statistical Institute, Kolkata
Date: July 18, 2025

Acknowledgements

In the early days of my life, I never imagined I would choose the academic research path as my career. Back then, I thought passing my subjects and landing a well-paid job would be enough. However, thanks to the collective efforts of many people and my time at the Indian Statistical Institute, my perspective on life has shifted and brought me to this point. Today, as I prepare to submit my thesis, I want to acknowledge these people's roles in my life and how important they have been to me.

First and foremost, I want to express my heartfelt gratitude to my parents for believing in me even when I was at my lowest. Their unwavering support, love, and sacrifices have shaped who I am today. Words cannot convey how grateful I am to be their son. I hope this thesis makes them proud. I also want to thank my uncle, Lt. Rupak Maitra, for introducing me to statistics. When I decided to leave biology at the 10+2 level, he suggested that I study statistics as a fourth subject in the 11th and 12th standards. He also introduced me to Mr. Sagar Sen, who helped me overcome the initial hurdles in understanding the basics of the subject. Sagar-da has been like an older brother to me, teaching me how to navigate the field of statistics. Due to their collective encouragement, I decided to pursue my Bachelor of Science (Hons.) degree in Statistics at St. Xavier's College, Kolkata. I am grateful to my teachers at St. Xavier's College for fostering my curiosity and enabling me to deepen my knowledge of the subject. However, the best part of joining St. Xavier's College was the group of friends I made, many of whom have now joined top corporate firms, and some have even gotten married. I would specifically like to mention Abhijoy and Ratnadeep. Their company is something I will cherish for life. I would also like to mention Prof. Tulsidas Mukherjee and Prof. Parthasarathi Chakrabarti for their thorough guidance during my time as an undergraduate student.

In 2018, I joined the Master of Statistics Program at the Indian Statistical Institute (ISI). I am thankful to all the professors at ISI Delhi for teaching me numerous theoretical and applied statistical courses and in-depth mathematical courses. I particularly want to thank Prof. Arup Pal for building our mathematical foundation and strengthening it for our future endeavors from the very first day of our postgraduate studies. I am also grateful to Prof. Anil K. Ghosh, who later became my Ph.D. supervisor, for introducing us to modern-day statistics, which opened up new horizons for understanding the topics. I also want to mention Prof. Anish Sarkar, Prof. Rajat

Subhra Hazra, Prof. Soumendu Sundar Mukherjee, and Prof. Ayan Basu for teaching me numerous courses and providing an excellent environment for my studies. Joining ISI as an M.Stat student not only helped me grow academically but also as a person, changing my views of the world. ISI made me realize that I am more academically inclined rather than being a corporate person. Thus, in the second year of my M.Stat course, I decided to pursue a Ph.D. in Statistics as my profession. However, the COVID-19 pandemic hit me hard, causing a six-month delay. Despite this, I managed to secure a research scholar position at the Theoretical Statistics and Mathematics Unit (SMU), ISI, Kolkata, on 2nd December 2020.

Joining the Indian Statistical Institute as a research scholar was a life-changing event. I am deeply grateful to have had the opportunity to study under Prof. Probal Chaudhuri, Prof. Arup Bose, Prof. Arijit Chakrabarty (SMU), Prof. Rudra P. Sarkar, Prof. Parthanil Roy, and Prof. Arijit Chakrabarti (Applied Statistics Unit). Witnessing their in-depth knowledge of their specialties and learning from them was a privilege I enjoyed at ISI as a Ph.D. student. I recall that whenever I got stuck on any problem, I could always approach them without hesitation. I would also like to thank Prof. Bhaswar B. Bhattacharya for guiding me through some advanced concepts in asymptotic statistics.

However, my deepest gratitude is reserved for my supervisor, Prof. Anil K. Ghosh. His unwavering guidance throughout my Ph.D. years taught me what it truly means to be a statistician. Beyond academics, he was also my football coach. Although I was initially hesitant to return to the field, his persistence paid off, and I found immense joy in playing again with him and fellow students at ISI. The blend of academic rigor and personal growth at ISI, marked by a tapestry of emotions, is beyond words.

I am also thankful to the editors, associate editors, and reviewers of various journals whose insightful comments significantly enhanced the quality of my research. Finally, I would like to thank my seniors Angshuman-da, Soham-da, Sukrit-da, Anurag-da, and Priyanka-di for all the different kinds of discussions we had (both academic and non-academic). I thank Deborshi, Bodhisatta, Subhodeep, Sayan, Bishakh-da, Sourav, and Spandan for the fond memories that we share. The difficulties that we went through, the conferences that we attended, and the trips that we canceled have made my journey all the more enjoyable.

Bilol Banerjee

Kolkata, July 18, 2025

Contents

Contents ix

List of Figures xi

List of Tables xv

Notations and Abbreviations xvii

1 Introduction 1

2 Two-Sample Test for High-Dimensional Data 9

2.1 Behaviour of the proposed test in HDLSS setup . . . 11

2.1.1 Test based on the ℓ_2 distance . . . 11

2.1.2 Tests based on generalized distances . . . 13

2.2 What happens in HDHSS regime? . . . 16

2.2.1 Minimax rate optimality . . . 16

2.2.2 Performance under shrinking alternatives . . . 17

2.3 Empirical performance of the proposed tests . . . 19

2.3.1 Analysis of simulated data sets . . . 19

2.3.2 Analysis of benchmark data sets . . . 25

2.4 Proofs and Mathematical details . . . 26

3 Test of Spherical Symmetry for High-Dimensional Data 39

3.1 Estimation of $\zeta(\mathrm{P})$ . . . 40

3.2 Test of Spherical Symmetry . . . 41

3.3 Asymptotic properties of the test . . . 43

3.3.1 Robustness . . . 43

3.3.2 Minimax rate optimality and high-dimensional behaviour . . . 44

3.3.3 Pitmann Efficiency . . . 45

3.4 Numerical studies . . . 47

3.4.1 Analysis of simulated data sets . . . 47

3.4.2 Analysis of a benchmark dataset . . . 50

3.5 Proofs and mathematical details . . . 51

3.5.1 Expression of $\zeta(\mathrm{P})$ for Gaussian distributions . . . 61

4 Distribution-free Tests of Spherical Symmetry **63**
4.1 String signs and string ranks . . . 64
4.2 Tests based on string signs and string ranks . . . 65
4.2.1 Algorithm for finding the shortest covering path . . . 66
4.2.2 Inner products vs. cosine similarities . . . 69
4.3 High dimensional behavior of the proposed tests . . . 70
4.3.1 Behavior of the proposed tests for HDLSS data . . . 70
4.3.2 Behavior of the proposed tests in HDHSS asymptotic regime . . . 72
4.4 Further modifications of the proposed tests . . . 75
4.5 Analysis of simulated and real data sets . . . 81
4.5.1 Analysis of 'Earthquake' data . . . 84
4.6 Tests of spherical symmetry about an unknown center . . . 86
4.7 Proofs and mathematical details . . . 88
4.7.1 Some additional mathematical details . . . 96
4.7.2 Pitman efficiency of the linear rank statistic . . . 98

5 Two-Sample Test for Functional Data **101**
5.1 Estimation of pBF and construction of the two-sample test . . . 103
5.1.1 Two-sample test based on $\hat{\eta}^{\phi}_{n,m}$ . . . 104
5.1.2 Local asymptotic behaviour of the test . . . 105
5.2 Empirical performance of the proposed test . . . 106
5.2.1 Analysis of simulated data sets . . . 107
5.2.2 Analysis of DTI data . . . 113
5.3 Proofs and mathematical details . . . 115

6 Test of Independence for Functional Data **127**
6.1 Projected Hilbert-Schmidt Independence Criterion . . . 128
6.1.1 Estimation of pHSIC . . . 129
6.1.2 Test of Independence based on pHSIC . . . 130
6.2 Results for independence between two random functions . . . 131
6.3 Results for independence among multiple random functions . . . 134
6.4 Proofs and mathematical details . . . 135

7 Concluding Remarks **141**

A Appendix: Brief Descriptions of Competing Tests **149**

Bibliography **153**

List of Figures

2.1 Powers of the permutation test based on $T_{n,m}^{\ell_2}$ in Examples 2.1-2.3. 13
2.2 Powers of BD-ℓ_2, BD-ℓ_1, BD-exp and BD-log tests in Examples 2.3-2.5. 15
2.3 Powers of the BD-ℓ_2 test for different choice of β and γ 19
2.4 Observed levels of the BD-ℓ_2 test for $n = m = 20$, 35 and 50 20
2.5 Powers of different two-sample tests in Examples 2.1-2.4 21
2.6 Powers of different two-sample tests in Examples 2.6-2.8 22
2.7 Powers of different two-sample tests in Examples 2.9-2.11. 23
2.8 Powers of different two-sample tests in Examples 2.12-2.14. 23
2.9 Powers of different two-sample tests in Examples 2.15 and 2.16. 25
2.10 Powers of different two-sample tests in benchmark data sets 26

3.1 Observed levels of the proposed test for observations generated from the standard Gaussian, Cauchy, and t_4 distributions 47
3.2 Powers of the proposed test, OT test, DT test and PP test in Examples 3.2(a)-(c). . 48
3.3 Powers of the proposed test, OT test, DT test and PP test in Examples 3.3-3.6. . . . 49
3.4 Powers of the proposed test, OT test, DT test and PP test in Examples 3.7(a)-(c). . 50
3.5 Powers of the proposed test, OT test, DT test and PP test in the Magic Gamma Telescope data set 50

4.1 Densities of the logarithm of $\|\mathbf{X}_1 - \mathbf{X}_2\|$, $\|\mathbf{X}_1 - \mathbf{X}_2'\|$ and $\|\mathbf{X}_1' - \mathbf{X}_2'\|$ and those of the logarithm of $(\mathbf{X}_1^\top \mathbf{X}_2)^2$, $(\mathbf{X}_1^\top \mathbf{X}_2')^2$ and $(\mathbf{X}_1'^\top \mathbf{X}_2')^2$. 64
4.2 Distributions of T_S and T_R for $d = 10$ and $d = 100$ 66
4.3 Algorithm for constructing the shortest covering path 68
4.4 Barplots of the difference between (a) the sign statistics and (b) the runs statistics constructed based on $\mathcal{P}$ and $\mathcal{P}_0$ for $d = 3$, 30, 300 and 3000 68
4.5 Boxplots showing the distributions of sign and runs statistics 69
4.6 Powers of the sign test and the runs test in Examples 4.1-4.3 when $n = 50$ and $d = 2^i$ for $i = 1, 2, \ldots, 10$. 72
4.7 Powers of the sign test and the runs test in Examples 4.1-4.3 when $n = d + 20$ and $d = 2^i$ for $i = 1, 2, \ldots, 10$. 74

4.8 Powers of the tests based on $\widetilde{T}_S$ and $\widetilde{T}_R$ in Examples 4.1-4.3 when $n = 50$ and $d = 2^i$ for $i = 1, 2, \ldots, 10$. 76
4.9 Powers of the modified sign test and the modified runs test in Examples 4.1-4.3 when $n = 50$ and $d = 2^i$ for $i = 1, 2, \ldots, 10$. 77
4.10 Powers of sign test, runs test, modified sign test and modified runs test in Examples 4.4 and 4.5 for $n = 50$ and $d = 2^i$ for $i = 1, \ldots, 6$. 79
4.11 The density estimates of $\tilde{\theta}(.,.)$ in Examples 4.4 and 4.5. 79
4.12 Powers of sign test, runs test, modified sign test and modified runs test in Examples 4.4 and 4.5 when $n = d^2 + 20$ and $d = 2^i$ for $i = 1, \ldots, 5$. 80
4.13 Powers of sign test, runs test, modified sign test, modified runs test, OT test and DT test in Examples 4.6 and 4.7. 82
4.14 Powers of sign test, runs test, modified sign test, modified runs test, OT test and DT test in Examples 4.8 and 4.9 82
4.15 Powers of sign test, runs test, modified sign test, modified runs test, OT test and DT test in Examples 4.10 and 4.11. 83
4.16 Power of sign test, runs test, modified sign test, modified runs test, OT test and DT test in Examples 4.8 -4.11 when $n = d^2 + 20$ increases with d. 83
4.17 Powers of sign test, runs test. modified sign test, modified runs test, OT test, and DT test based on varying proportions of observations from the positive and the negative cases in the 'Earthquakes' dataset. 85
4.18 Coordinate-wise mean and variance of the feature vector in the 'Earthquakes' data set divided into two groups of positive and negative cases. 85
4.19 Powers of sign test, runs test. modified sign test, modified runs test, OT test, and DT test based on varying proportions of observations from the positive and the negative cases in the truncated 'Earthquakes' dataset. 86
4.20 Type I errors of sign test, runs test, modified sign test, modified runs test, OT test and DT test in Example 4.12 when (a) n increases while d is kept fixed at 20 and (b) d increases while n is kept fixed at 50. 87
4.21 Powers of sign test, runs test, modified sign test, modified runs test, OT test and DT test in Example 4.13 when (a) the samples are centered using the spatial median and (b) we use differences of the observations based on sample splitting. 88

5.1 Powers of pBF-ℓ_2, pBF-exp and pBF-log tests in Examples 5.1-5.3. 107
5.2 Powers of pBF-ℓ_2, pBF-exp, pBF-log, FAD, BD and WD tests in Examples 5.4 (i) and (ii). 108
5.3 Powers of pBF-ℓ_2, pBF-exp, pBF-log, FAD, BD and WD tests in Examples 5.5 (i) and (ii) 109

5.4 Sample paths of $X(t)$ and $Y(t)$ for $0 \leq t \leq 1$ in Examples 5.6 (i) and (ii). 109
5.5 Powers of pBF-ℓ_2, pBF-exp, pBF-log, FAD, BD and WD tests in Examples 5.6 (i) and (ii). 110
5.6 Powers of pBF-ℓ_2, pBF-exp, pBF-log, FAD, BD and WD tests in Example 5.7. . . . 111
5.7 Powers of pBF-ℓ_2, pBF-exp, pBF-log, FAD, BD and WD tests in Examples 5.8 and 5.9. 111
5.8 Powers of pBF-ℓ_2, pBF-exp, pBF-log, FAD, BD and WD tests in Examples 5.10 (i) and (ii). 112
5.9 Powers of pBF-ℓ_2, pBF-exp, pBF-log, FAD, BD and WD tests in Examples 5.11 (i) and (ii). 113
5.10 The fractional anisotropy (FA) tract profiles on the first visit of the patients (divided according to health status and gender) in the DTI dataset 113
5.11 Powers of pBF-ℓ_2, pBF-exp, pBF-log, FAD, BD and WD tests in the DTI data set . 114

6.1 Powers of pHSIC, bCov, dCov, aCov_1 and aCov_2 tests in Examples 6.1 and 6.2. . . . 132
6.2 Powers of pHSIC, bCov, dCov, aCov_1 and aCov_2 tests in Examples 6.3 and 6.4. . . . 132
6.3 Scatter plots of observations from the six unusual bivariate distributions in Newton (2009). 133
6.4 Powers of pHSIC, bCov, dCov, aCov_1 and aCov_2 tests in Examples 6.5 (a)-(f). . . . 133
6.5 Powers of pHSIC and bCov tests in Examples 6.6 and 6.7. 134
6.6 Powers of pHSIC and bCov tests in Examples 6.8 and 6.9. 135

List of Tables

3.1 Powers of the proposed test for the alternative $F_{1-\beta_n n^{-1/2}} = (1-\beta_n n^{-1/2})\mathcal{N}_{10}(\mathbf{0}_{10}, \mathbf{I}_{10}) + \beta_n n^{-1/2}\mathcal{N}_{10}(\mathbf{0}_{10}, 0.5\mathbf{I}_{10} + 0.5\mathbf{J}_{10})$ when $\beta_n = 5n^{\gamma}$. 46

5.1 p-values of pBF-ℓ_2, pB-log, pBF-exp, BD, WD tests, and the Bonferroni corrected p-value of the FAD test for the DTI dataset divided according to health status and gender of the patients. 114

Notations and Abbreviations

$\asymp$	:	Asymptotically of the same order.
$<<$	:	Much smaller than.
$\langle \cdot, \cdot \rangle$	:	Inner product between two vectors or functions.
$\| \cdot \|$	:	Euclidean norm.
$\overset{a.s.}{\to}$	:	Almost sure convergence
$\overset{D}{\to}$	:	Convergence in distribution.
$\overset{\mathrm{P}}{\to}$	:	Convergence in probability.
$\overset{D}{=}$	:	Equality in distribution.
$\overset{iid}{\sim}$	:	Independent ans identically distributed as.
$\bigoplus_{i=1}^{d}$	:	External direct sum of d many Hilbert spaces.
$\otimes_{i=1}^{d}$	:	Product of d many measures.
$\mathbf{0}_d, \mathbf{1}_d$	:	d-dimensional vectors $(0, \ldots, 0)^\top$ and $(1, \ldots, 1)^\top$, respectively.
$\boldsymbol{A}^\top$	:	Transpose of a matrix $\boldsymbol{A}$.
$\big\lvert \boldsymbol{A} \big\rvert$	:	Determinant of the matrix $\boldsymbol{A}$.
$\big\lvert a \big\rvert$	:	Absolute value of a.
$a(\cdot)$	:	A score function used in linear rank statistics.
$a_d = o\big(f(d)\big)$	:	$\frac{a_d}{f(d)}$ converges to zero as $d \to \infty$.
$a_d = O\big(f(d)\big)$	:	$\frac{a_d}{f(d)}$ is a bounded sequence in d.
$a_d = o_P\big(f(d)\big)$	:	$\frac{a_d}{f(d)}$ converges to zero in probability as $d \to \infty$.
$a_d = O_P\big(f(d)\big)$	:	$\frac{a_d}{f(d)}$ is a bounded in probability sequence in d.
α	:	Level of significance of a test.
ANOVA	:	Analysis of Variance.
$\mathbb{B}(\mathbf{x}, \epsilon)$	:	A closed ball of radius ϵ centered at $\mathbf{x}$.
B	:	Number of repetition of the resampling step to approximate the p-value.
$\mathcal{B}(\mathbb{X})$	:	Borel σ-algebra endowed on the metric space $\mathbb{X}$.
χ^2_p	:	Chi-square random variable with p degrees of freedom.
$c_{1-\alpha}, c^{\phi}_{1-\alpha}$	:	Cut-off of the permutation test.
CVM	:	Cramer-von-Mises statistic.
CLT	:	Central Limit Theorem.

$\mathrm{Cov}(X, Y)$	:	Covariance between two random variables X and Y.
d	:	Dimension of the data.
$d_{TV}(F, G)$	:	Total variation distance between two distributions F and G.
$\mathcal{D}$	:	Data set.
$\mathcal{D}^{'}$	:	Collection of spherically symmetric variants of the observations in $\mathcal{D}$.
$\mathcal{D}_A$	:	Augmented data set ($\mathcal{D}_A = \mathcal{D} \cup \mathcal{D}^{'}$)
$\delta(\mathbf{x}, \mathbf{y}, \mathbf{z})$	:	Indicator function $\mathrm{I}[\rho(\mathbf{x}, \mathbf{z}) \leq \rho(\mathbf{y}, \mathbf{z})]$, where ρ is a distance function.
$\mathrm{diag}(\mathbf{\Sigma})$	:	Diagonal matrix constructed with the diagonal elements of the matrix $\mathbf{\Sigma}$.
dHSIC	:	d-variate Hilbert Schmidt Independence Criterion.
$\mathbb{E}(\mathbf{X})$, $\mathbb{E}(X)$	:	Expectation of the random vector $\mathbf{X}$ or the random variable X, respectively.
$\mathbb{E}(X\|Y)$	:	Conditional expectation of the random variable X given the random variable Y.
$\mathcal{E}(\mathrm{F})$	:	Kernel mean embedding function of the distribution F into an RKHS.
$e_{h,\psi}(F, G)$	:	Energy distance between two distributions F and G based on the distance function $\varphi_{h,\psi}(\cdot, \cdot)$.
$\overline{E}$	:	Closure of the set E.
$F\big(\mathbb{B}(\mathbf{x}, \epsilon)\big)$	:	Measure of the closed ball $\mathbb{B}(\mathbf{x}, \epsilon)$ with respect to the distribution F.
$\hat{F}, \hat{F}_n$	:	Empirical distribution based on $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} F$.
$\varphi(\boldsymbol{t})$	:	Characteristic function computed at $\boldsymbol{t}$.
$\varphi_{h,\psi}(\mathbf{x}, \mathbf{y})$	:	Generalized distance between two observations $\mathbf{x}$ and $\mathbf{y}$ (see page 13).
ϕ_0, ϕ, ψ, h	:	Strictly increasing functions from $\mathbb{R}_+$ to $\mathbb{R}_+$ with $\phi_0(0) = \phi(0) = \psi(0) = h(0) = 0$.
$\mathbb{G}_{\mathrm{P}}$	:	The P-Brownian Bridge process.
$\hat{G}, \hat{G}_m$	:	Empirical distribution based on $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \overset{i.i.d.}{\sim} G$.
$\mathbf{H}$	:	Orthogonal matrix.
$\mathcal{H}$	:	Hilbert space.
$\eta_\phi(F, G)$	:	Measure of dissimilarity between two distributions F and G (see page 102).
$\hat{\eta}^{\phi}_{n,m}$	:	Proposed empirical analog of $\eta_\phi(F, G)$ based on samples of size n and m.
H_0, H_1	:	Null hypothesis and alternative hypothesis, respectively.
HDLSS	:	High Dimension, Low Sample Size.
HDHSS	:	High Dimension, High Sample Size.
HSIC	:	Hilbert Schmidt Independence Criterion.
$\mathrm{I}[\cdot], \mathrm{I}\{\cdot\}$	:	Indicator function.
i.i.d.	:	Independent and identically distributed.
$\mathbf{I}_d$	:	$d \times d$ identity matrix.
$\mathbf{J}_d$	:	$d \times d$ matrix with all entries equal to one.
$K(\cdot, \cdot)$	:	Kernel function from $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.
$\mathcal{K}_n$	:	Complete graph on a set of n vertices.
KS	:	Kolmogorov-Smirnov statistic.

$KL(F, G)$: Kullback-Leibler divergence between two distributions F and G.

$L_2(\mathbb{X}, \mathcal{A}, \mathrm{P})$: Space of all square integrable functions on the probability space $(\mathbb{X}, \mathcal{A}, \mathrm{P})$.

$L_2[a, b]$: Hilbert space of all real square integrable functions defined on $[a, b]$.

$\mathrm{Laplace}(\mu, \sigma)$: Laplace distribution with location μ and scale σ.

$\lambda_{\max}(\mathbf{\Sigma})$: Maximum eigenvalue of the matrix $\mathbf{\Sigma}$.

$\liminf$: Limit infimum.

$\limsup$: Limit supremum.

$\mathcal{M}(\mathbb{X})$: Set of all probability distributions on the measurable space $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$.

μ_0 : Uniform measure on $\mathcal{S}^{d-1}$, the surface of the d-dimensional unit sphere.

MMD : Maximum Mean Discrepancy.

n, m : Sample sizes.

$\mathcal{N}_1(\mu, \sigma^2)$: Normal distribution with mean μ and variance σ^2.

$\mathcal{N}_d(\boldsymbol{\mu}, \mathbf{\Sigma})$: d-dimensional normal distribution with mean $\boldsymbol{\mu}$ and dispersion $\mathbf{\Sigma}$.

$\mathcal{N}(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$: Bivariate normal distribution with location (μ_1, μ_2), variances σ_1^2, σ_2^2 and correlation coefficient ρ.

ν_0 : Haar measure over the set of all orthogonal matrices.

$\mathcal{P}$: Actual shortest covering path.

$\mathcal{P}_0$: Covering path obtained using a greedy algorithm as an approximation of $\mathcal{P}$.

P : Probability measure.

P^* : Probability under permutation or random swap resampling distribution.

$\mathbb{P}^f$: Distribution of $\langle X, f \rangle$ where X follows the distribution $\mathbb{P}$.

P_i : Distribution function of the i-th population.

$p_{n,m}$: Permutation p-value for the two-sample problem with sample sizes n and m.

$p_{n,m,B}$: Monte Carlo approximation of $p_{n,m}$ based on B replications.

p_n : Resampling p-value for testing spherical symmetry based on n observations.

$p_{n,B}$: Monte-Carlo approximation of p_n based on B replications.

$\{\pi_1, \pi_2, \ldots, \pi_n\}$: A permutation of the elements of $\{1, 2 \ldots, n\}$.

pBF : Projected Baringhauz-Franz statistic.

pHSIC : Projected Hilbert Schmidt Independence Criterion.

$\Theta_\rho^2(F, G)$: Ball divergence between two distributions F and G based on the metric ρ.

$\mathbb{R}$: Set of real numbers.

$\mathbb{R}^d$: d-dimensional Euclidean space.

$\mathbb{R}_+$: Set of non-negative real numbers $\{x \in \mathbb{R} : x \geq 0\}$.

RKHS : Reproducing Kernel Hilbert Space

$\rho(\mathbf{x}, \mathbf{y})$: Distance between two observations $\mathbf{x}$ and $\mathbf{y}$.

$\boldsymbol{R} = (R_1, R_2, \ldots, R_n)$: Vector of string ranks.

$\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$: Vector of string signs.

$\mathcal{S}_N$: Set of all permutations of $\{1, 2, \ldots, N\}$.

$\mathcal{S}^{d-1}$: Surface of the d-dimensional unit sphere.

$\text{supp}\{\nu\}$: Support of the measure ν.

$\overline{span\{A\}}$: Span closure of a set A.

$\lfloor t \rfloor$: Largest integer $\leq t$.

t_ν : Univariate Student's t distribution with ν degrees of freedom.

$\boldsymbol{t}_\nu$: Multivariate Student's t distribution with ν degrees of freedom.

$T^{\rho}_{n,m}$: Proposed estimator of $\Theta^2_\rho(F, G)$ based on samples of size n and m.

T_{LR} : Linear rank statistic.

T_S : Sign statistic based on string signs.

T_R : Runs statistic based on string ranks.

T^M_S : Modified sign statistic.

T^M_R : Modified runs statistic.

$\text{trace}(\boldsymbol{\Sigma})$: Trace of the matrix $\boldsymbol{\Sigma}$.

$\text{Unif}(a, b)$: Uniform distribution on (a, b).

$\text{Unif}(\mathcal{S}^{d-1})$: Uniform distribution on $\mathcal{S}^{d-1}$.

$\text{Unif}(A)$: Uniform distribution over the set A.

$\text{Var}(\mathbf{X})$: Variance-covariance matrix of the random vector $\mathbf{X}$.

$\text{Var}(X)$: Variance of the random variable X.

WLLN : Weak law of large numbers.

$\mathcal{X}$: Collection of independent observations on $\mathbf{X}$.

$(\mathbb{X}, \mathcal{A})$: Measurable space.

$(\mathbb{X}, \mathcal{A}, \text{P})$: Probability space.

$\mathbf{X}, \mathbf{Y}$: Multivariate random vectors or functions.

X, Y : Univariate random variables or functions.

$\mathbf{X} \sim F$: Random vector $\mathbf{X}$ follows the distribution F.

$\mathbf{X}'$: Spherically symmetric variant of $\mathbf{X}$.

$\mathbf{x}, \mathbf{y}$: Observations on the random vectors $\mathbf{X}$ and $\mathbf{Y}$, respectively.

$X^{(q)}$ or $x^{(q)}$: The q-th component of the (random) vector or function $\mathbf{X}$ or $\mathbf{x}$.

$\xi(\mathbb{P})$: Measure of dependence among several random functions having the joint distribution $\mathbb{P}$.

$\hat{\xi}_n$: Proposed empirical analog of $\xi(\mathbb{P})$ based on n observations.

$\mathcal{Y}$: Collection of independent observations on $\mathbf{Y}$.

$\zeta(\text{P})$: Measure of spherical asymmetry of the distribution P.

$\hat{\zeta}_n$: Proposed empirical analog of $\zeta(\text{P})$ based on n observations.

Chapter 1

Introduction

The advancement of information technology and sciences over the last few decades has facilitated the collection, storage, and analysis of huge data sets. Many of these data sets contain observations having a large number of features, and in some cases, this number is comparable to or even much larger than the sample size. For instance, in the fields of medical image analysis (see, e.g., Yushkevich et al., 2001), chemometrics (see, e.g., Schoonover, Marx & Zhang, 2003), genomics (see, e.g., Alon et al., 1999), astronomy (see, e.g., Ratcliffe et al., 2020), and climate forecasting (see, e.g., Christiansen, 2021), we often deal with such high-dimensional data sets, where the data dimension is larger than or comparable to the sample size. A huge pool of such high-dimensional data sets is available at the UCI Machine Learning Repository. Classical statistical methods often fail to analyze these high-dimensional data sets properly. For example, the Hotelling's T^2 test for the two-sample problem, the test of independence based on likelihood ratio or canonical correlation coefficient, and the likelihood ratio test for spherical sphericity (see Mardia, Kent & Bibby, 1979) cannot be used in such cases because of the singularity or near-singularity of the sample variance-covariance matrix. So, meaningful statistical analysis of such high-dimensional data is a challenging problem for the statistics community. We know that the performance of a parametric method depends heavily on the validity of specified model assumptions. However, it is very difficult to test the validity of these parametric model assumptions in high dimensions, and violations of these assumptions often yield misleading inferences. On the other hand, nonparametric methods, which are often preferred because of their robustness and flexibility, suffer from the curse of dimensionality.

In many real-life scenarios, we also encounter situations where the features are not scalars or finite-dimensional vectors but are functions or curves. For instance, we may observe the electrocardiogram (ECG) reading of different individuals over a time interval (see, e.g., Chen et al., 2015), monthly sea surface temperature (see,e.g., Ferraty & Romain, 2011), monthly weather of a country (see, e.g., Ramsay & Silverman, 2005), diffusion tensor imaging of the brain for different individuals (see,e.g., Goldsmith et al., 2011, 2012), growth curves of males and females (see, e.g., Ramsay & Silverman, 2005), or images of handwritten digits (see, e.g., Kussul & Baidyk, 2004). A large collection of such data sets is available at the UCI Time Series Classification Archive. The newly developed branch of statistics that deals with the analysis of such functional data sets is called functional data analysis (see, e.g., Ramsay & Silverman, 2002; Ferraty & Vieu, 2006; Hsing

& Eubank, 2015). To cope with the complex nature of functional data, nonparametric methods are often preferred over parametric ones.

In this thesis, we deal with both high-dimensional and function data. We develop some nonparametric methods for hypothesis testing involving high-dimensional and functional data sets and investigate their theoretical and empirical performance. Our main contributions are given in Chapters 2-6. In Chapters 2-4, we consider some hypothesis testing problems for high-dimensional data, while in Chapters 5 and 6, we deal with problems involving functional data. All proofs and mathematical details related to our contributions are given at the end of respective chapters. Brief descriptions of our contributions are given below.

Two sample tests for high-dimensional data

In Chapter 2, we consider the two-sample problem, where we test the null hypothesis $H_0 : F = G$ against the alternative hypothesis $H_1 : F \neq G$ based on two sets of independent observations $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ and $\mathcal{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m\}$ from two d-dimensional continuous distributions F and G, respectively. Several nonparametric tests are available for this problem, especially for univariate data. While the Wilcoxon-Mann-Whitney test is popular for the univariate two-sample location problem, the Wald-Wolfowitz runs test, the Kolmogorov-Smirnov (KS) test, and the Camer-von-Mises (CVM) test apply to general two-sample problems. We refer the readers to Hollander, Wolfe *&* Chicken (2014) and Gibbons *&* Chakraborti (2011) for an exposition on nonparametric tests for univariate data. Several nonparametric tests have been proposed for multivariate data as well. Friedman *&* Rafsky (1979) used the idea of a minimum spanning tree to generalize the runs test and the KS test to higher dimensions. Baringhaus *&* Franz (2004) proposed a multivariate two-sample test based on inter-point distances, which can be viewed as a generalization of the CVM test. Székely *&* Rizzo (2004) and Aslan *&* Zech (2005) also used inter-point distances to develop tests based on energy statistics. Schilling (1986) and Henze (1988) developed multivariate two-sample tests based on nearest-neighbors. Rosenbaum (2005) proposed a distribution-free test based on optimal non-bipartite matching. Gretton et al. (2012) used the notion of maximum mean discrepancy (MMD) to construct a test based on kernel mean embedding of two probability distributions. These multivariate two-sample tests are consistent in the classical asymptotic regime, i.e., for any fixed d, the powers of these tests converge to one as the sample sizes diverge to infinity. Since these tests are based on pairwise distances among the observations, they can also be used for high-dimensional data even when the dimension is much larger than the combined sample size. But most of them often perform poorly in the high dimension, low sample size (HDLSS) situations, especially when the scale difference between F and G dominates their location difference (see, e.g., Biswas *&* Ghosh, 2014).

Following the seminal paper by Hall, Marron *&* Neeman (2005), the HDLSS regime has recently received increasing attention. Over the last ten years, several two-sample tests have been proposed for HDLSS data. Wei et al. (2016); Ghosh *&* Biswas (2016); Srivastava, Li *&* Ruppert (2016) proposed some tests based on linear projections, mainly useful for two-sample location problems. Biswas *&* Ghosh (2014) and Tsukada (2019) proposed some general two-sample tests based on averages of inter-point distances. Under some appropriate assumptions, these two tests turn out to be consistent in both classical and HDLSS asymptotic regimes, but nothing is known about their asymptotic behavior when the sample sizes increase with the dimension. Moreover, they are not robust against outliers generated from heavy-tailed distributions. Kim, Balakrishnan *&* Wasserman (2020) developed a robust multivariate test based on projection averaging, but it is applicable only when the distances between the observations are measured using the Euclidean metric. Some graph-based high-dimensional two-sample tests have also been proposed in the literature. This includes the test based on nearest neighbors (Mondal, Biswas *&* Ghosh, 2015), multivariate runs test based on the shortest Hamiltonian path (Biswas, Mukhopadhyay *&* Ghosh, 2014), and the test based on triangles (Liu *&* Modarres, 2011). Under appropriate regularity conditions, these graph-based tests are consistent in the HDLSS asymptotic regime. But in classical asymptotic regime, they usually have poor powers against local alternatives (see, Bhattacharya, 2019). Even the large sample consistency of the SHP-based runs test and the triangle test is yet to be established. Also, it is unknown how these tests perform in the high dimension, high sample size (HDHSS) asymptotic regime, where the sample sizes increase with the dimension. This type of asymptotic behavior has been studied for some location (see, e.g., Bai *&* Saranadasa, 1996; Chen *&* Qin, 2010; Aoshima *&* Yata, 2018) and scale (see,e.g., Li *&* Chen, 2012; Cai, Liu *&* Xia, 2013) problems, but for the general two-sample test, the literature is scarce.

In this chapter, we develop some distance-based two-sample tests and investigate their high-dimensional behavior in both HDLSS and HDHSS regimes. Our test statistics can be viewed as empirical analogs of the ball divergence measure proposed by Pan et al. (2018). In that article, the authors studied the large sample behavior of their proposed test. The large sample behavior of our tests can be derived from their results. Therefore, here, we focus only on the high-dimensional behavior of our tests. However, in the HDLSS setup, the test based on the ℓ_2-distance (i.e., the Euclidean distance) may fail to discriminate between two distributions differing outside the first two moments. To take care of this problem, we use the generalized distance functions proposed in Sarkar *&* Ghosh (2018a) to construct our tests and prove their high-dimensional consistency against a larger class of alternatives. We also establish the minimax rate optimality of the proposed tests and prove their consistency against a suitable class of shrinking alternatives. Our empirical study shows the efficacy of our tests against some popular state-of-the-art methods. The contents of this chapter are based on Banerjee *&* Ghosh (2025).

Test of spherical symmetry for high-dimensional data

In Chapters 3 and 4, we consider the problem of testing the spherical symmetry of a multivariate distribution based on a set of independent and identically distributed (i.i,d.) observations $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ from it. A random vector $\mathbf{X}$ is said to follow a spherically symmetric distribution if its distribution is rotation invariant, or in other words, $\mathbf{X}$ and $\mathbf{HX}$ have the same distribution (i.e., $\mathbf{X} \stackrel{D}{=} \mathbf{HX}$) for any orthogonal matrix $\mathbf{H}$. This is an important class of distributions in the statistics literature. Motivated by the spherical symmetry or elliptic symmetry (i.e., spherical symmetry after standardization) of the underlying distributions, several statistical methods have been developed. Robust measures of multivariate location and scale (see, e.g., Van Aelst & Rousseeuw, 2009), tests for multivariate location (see, e.g., Randles, 1989; Chaudhuri & Sengupta, 1993), Stein estimation (see, e.g., Fourdrinier, Strawderman & Wells, 2018), classification (see, e.g., Ghosh & Chaudhuri, 2005; Li, Cuesta-Albertos & Liu, 2012), and clustering (see, e.g., Jörnsten, 2004) are some examples of its widespread applications. Therefore, testing the sphericity of a distribution is an important statistical problem, and several tests have been proposed for it.

Smith (1977) proposed a test for bivariate data that uses the fact that if $\mathbf{X}$ is spherically symmetric, then $\|\mathbf{X}\|$ and $\mathbf{X}/\|\mathbf{X}\|$ are independent while $\mathbf{X}/\|\mathbf{X}\|$ follows a uniform distribution over the perimeter of the unit circle in $\mathbb{R}^2$. Later, Baringhaus (1991) modified the test statistic and generalized the test to any arbitrary dimension. However, this test involves a complex function of dimension d, which makes it difficult to study its high-dimensional behavior. Fang, Zhu & Bentler (1993) proposed an asymptotically distribution-free test for spherical symmetry using the projection pursuit technique. They implemented the Wilcoxon-Mann-Whitney test for several pairs of projection directions and used the minimum over all such projection pairs as the test statistic. However, this is only a necessary test for sphericity and does not have large sample consistency under general alternatives. Koltchinskii & Li (1998) proposed a test based on the difference between the empirical spatial rank function and the theoretical spatial rank function under spherical symmetry, where the unknown components of these theoretical ranks were estimated from the data. The authors proposed a bootstrap method for calibration. Diks & Tong (1999) proposed a Monte Carlo test for multivariate spherical symmetry conditionally on minimal sufficient statistics. Liang, Fang & Hickernell (2008) proposed some necessary tests for spherical symmetry by using the fact that under spherical symmetry $\mathbf{X}/\|\mathbf{X}\|$ is uniformly distributed on $\mathcal{S}^{d-1}$, the surface of the unit sphere in $\mathbb{R}^d$, but they did not consider the independence between $\|\mathbf{X}\|$ and $\mathbf{X}/\|\mathbf{X}\|$. Henze, Hlávka & Meintanis (2014) proposed a test based on characteristic functions and calibrated the test using a bootstrap algorithm. However, this test requires the generation of the uniform grid over the unit sphere, which becomes computationally prohibitive even in moderately large dimensions. Albisetti, Balabdaoui & Holzmann (2020) proposed another test utilizing the fact that $\mathbf{X}$ is spherically symmetric if and only if $\mathbb{E}\{\langle \mathbf{X}, \mathbf{u}\rangle \mid \langle \mathbf{X}, \mathbf{v}\rangle\} = 0$ for all $\mathbf{u}$ and $\mathbf{v}$ with $\langle \mathbf{v}, \mathbf{u}\rangle = 0$. They constructed

a KS-type test statistic based on suitable choices of test functions. Recently, Huang & Sen (2023) proposed some tests for different notions of symmetry for multivariate data using optimal transport.

These above-mentioned tests can be used when the dimension of the data is small compared to the sample size (i.e., $d << n$), and some of them are large-sample consistent for any fixed dimension d. However, the applicability of these tests for high-dimensional data (i.e., when d is comparable to or larger than n) is not clear. For meaningful implementation of Diks & Tong (1999)'s test for high-dimensional data, one needs to go for an appropriate data-driven scale adjustment. The test proposed by Huang & Sen (2023) can also be used for high-dimensional data, but it usually has poor performance in high dimensions unless the location of the distribution significantly differs from the origin. Zou et al. (2014) and Feng & Liu (2017) proposed tests of sphericity for high-dimensional data, where the test statistics were constructed using the multivariate sign function assuming the ellipticity of the underlying distribution. Ding (2020) proposed a test based on the ratio of traces of different powers of the sample variance-covariance matrix and established its high dimensional consistency. However, these tests may fail when the underlying distribution is not spherically symmetric, but $\mathbf{X}/\|\mathbf{X}\|$ is uniformly distributed on $\mathcal{S}^{d-1}$(e.g., angular symmetric) or the variance-covariance matrix is a constant multiple of the identity matrix.

Test based on a maximum mean discrepancy type measure

In Chapter 3, we propose a new measure of the deviation from spherical symmetry by using the fact that $\mathbf{X}$ is spherically symmetric if and only if $\mathbf{X}$ and $\mathbf{X}' = \|\mathbf{X}\|\mathbf{U}$ have the same distribution, where $\mathbf{U} \sim \text{Unif}\,(\mathcal{S}^{d-1})$ and it is independent of $\mathbf{X}$. The proposed measure is based on the maximum mean discrepancy (see, e.g., Gretton et al., 2012) between the distribution of $\mathbf{X}$ and that of its spherically symmetric variant $\mathbf{X}'$. This measure is non-negative and takes the value zero if and only if $\mathbf{X}$ is spherically symmetric. However, it involves some terms that are not estimable from only the observed data $\mathcal{D} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$. To overcome this limitation, we propose a data augmentation approach, where we augment $\mathcal{D}$ with $\mathcal{D}' = \{\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_n\}$ to have an augmented dataset $\mathcal{D}_A = \{(\mathbf{X}_i, \mathbf{X}'_i) : i = 1, 2, \ldots, n\}$. Here $\mathbf{X}'_i = \|\mathbf{X}_i\|\mathbf{U}_i$ for $i = 1, 2, \ldots, n$, where $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n \overset{i.i.d.}{\sim} \text{Unif}(\mathcal{S}^{d-1})$, and they are independent of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. Using the augmented data set $\mathcal{D}_A$, we construct a consistent estimator of the proposed measure and build a nonparametric test based on it. Our test is calibrated using a novel resampling algorithm. We investigate the theoretical properties of our test not only when the dimension remains fixed, and the sample size diverges to infinity but also when both the dimension and sample size diverge simultaneously. Several simulated data sets are analyzed to compare its empirical performance with some state-of-the-art methods. The contents of this chapter are taken from Banerjee & Ghosh (2024a).

Distribution-free sign and runs tests

Though the test proposed in Chapter 3 performs well in classical and HDHSS regimes, like many other existing tests, it has an underwhelming performance in the HDLSS setup for a fairly general class of alternatives. Moreover, the use of the resampling method for calibration leads to higher computing costs. Keeping these in mind, in Chapter 4, we develop some graph-based tests for spherical symmetry, which have the exact distribution-free property. Here also, we use the data augmentation method. We introduce a new notion of signs and ranks that are computed along a path obtained by minimizing an objective function based on pairwise dissimilarities among the observations in the augmented data set $\mathcal{D}_A$. Under spherical symmetry, these sign and rank vectors have the exact distribution-free property, and their joint null distribution does not depend on the dimension of the data. So, any statistic based on these signs and ranks has the same null distribution as the corresponding test statistic constructed based on univariate signs and ranks. Using this new notion of signs and ranks, we construct some exact distribution-free tests for spherical symmetry. These tests can be conveniently used for high-dimensional data sets, even when the dimension is much larger than the sample size. Under appropriate regularity conditions, we study the asymptotic behavior of these tests both in HDLSS and HDHSS asymptotic regimes.

We know that in the case of a spherical distribution, all diagonal elements of the scatter matrix are equal, while all off-diagonal elements are zero. We observe that our proposed tests have excellent performance for alternatives having significant correlations among the variables. However, when they are uncorrelated, and the difference is mainly in the scales of the variables, the performance of these tests may not be satisfactory. Therefore, we also propose some modifications of our tests to make them useful for a wide class of alternatives in the HDLSS asymptotic regime. Our empirical study based on the analysis of simulated and real data sets demonstrates the superiority of the proposed tests over the existing ones for a wide variety of alternatives involving HDLSS and HDHSS data. The contents of this chapter are based on Banerjee & Ghosh (2024b).

Two-sample test for functional data

In Chapter 5, we consider the two-sample problem for functional data sets. In functional data analysis, feature variables are often modeled as elements of a separable Hilbert space (see,e.g., Ramsay & Silverman, 2005; Ferraty & Vieu, 2006; Hsing & Eubank, 2015). For such variables, there are many ANOVA-type tests (see, e.g., Zhang, Peng & Zhang, 2010; Cuesta-Albertos & Febrero-Bande, 2010; Qiu, Chen & Zhang, 2021) that deal with the location problem. In the nonparametric setup, Hall & Van Keilegom (2007) proposed a Cramer-von-Mises type test against general alternatives. Pomann, Staicu & Ghosh (2016) proposed a method that uses the Anderson-Darling test on the first few functional principal components of the mixture distribution and aggregates the results using Bonferroni's correction. Wynne & Duncan (2020) developed a test

based on kernel mean embedding of the distributions. Pan et al. (2018) proposed a test based on ball divergence, which is applicable to Banach-valued random variables. Most of these tests are based on a consistent estimate of a measure of dissimilarity between the two underlying distributions, and they are large sample consistent. However, the exact or limiting null distributions of these test statistics are analytically intractable, and one needs to use the permutation method for calibration. However, the large sample consistency of these permutation tests is somewhat missing from the literature. The existing literature is also somewhat silent about the statistical efficiency of these tests under local contiguous alternatives, partially because of the difficulty in formulating a suitable notion of the density and the likelihood ratio statistic.

In this chapter, we deal with two sets of independent observations $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$ on two functional random variables $X \sim F$ and $Y \sim G$, respectively, which are assumed to lie in an infinite-dimensional separable Hilbert space $\mathcal{H}$ with inner product $\langle ., . \rangle$. We know that X and Y have the same distribution (i.e., $F = G$) if and only if the random variables $\langle X, f \rangle$ and $\langle Y, f \rangle$ are identically distributed for all $f \in \mathcal{H}$ (see Hsing & Eubank, 2015, Theorem 7.1.2). Motivated by this result, we propose a new measure of distributional dissimilarity for functional data. This measure is non-negative, and it takes the value zero if and only if $F = G$. We propose a consistent estimator of this measure and use it as the test statistic to test for the equality of F and G. Large sample distribution of the test statistic is derived both under null and fixed alternative hypotheses. However, it is difficult to use this large sample distribution for calibration. So, we use the conditional test based on the permutation principle and prove its large sample consistency. We also construct a locally asymptotically normal sequence of contiguous alternatives and study the behavior of our test under such alternatives. Our results show that against such contiguous alternatives, our test is Pitman efficient. Extensive empirical studies are carried out to compare the performance of this test with some state-of-the-art methods. The contents of this chapter are taken from Banerjee (2024).

Test of independence for functional data

In Chapter 6, we consider testing independence among d random functions $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$ based on n independent observations $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ on $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})$.

Testing independence or measuring dependence among several random variables or random vectors is a fundamental problem in statistics, and several methods have been proposed for it. For example, Spearman's ρ, Kendall's τ, Bolmqvist's β (Blomqvist, 1950) or Hoeffding's ϕ^2 (Hoeffding, 1948) statistics measure the association between the two random variables. To assess the dependence among multiple random variables, several measures were also constructed (see Nelsen, 1996; Úbeda-Flores, 2005; Gaißer, Ruppert & Schmid, 2010; Póczos, Ghahramani & Schneider, 2012; Roy et al., 2022) in the past few decades. In the multivariate set up, some notable tests of

independence between two random vectors include tests based on interdirections (Gieser & Randles, 1997), spatial signs and ranks (Taskinen, Kankainen & Oja, 2003; Taskinen, Oja & Randles, 2005), distance covariance (dCov) (Székely, Rizzo & Bakirov, 2007), distance-based contingency tables (Heller, Heller & Gorfine, 2013), Hilbert-Smidth Independence Criterion (HSIC) (Gretton et al., 2007; Gretton & Györfi, 2010), projection criterion (Zhu et al., 2017), and those based on graphs (Friedman & Rafsky, 1983; Heller, Gorfine & Heller, 2012; Biswas, Sarkar & Ghosh, 2016; Sarkar & Ghosh, 2018b). Some generalizations of the dCov test (Fan et al., 2017; Jin & Matteson, 2018; Chakraborty & Zhang, 2019) and the HSIC test (Pfister et al., 2018) have also been proposed in the literature for testing mutual independence among more than two random vectors. Other tests for mutual independence among multiple random vectors include those based on half-spaces (Beran, Bilodeau & de Micheaux, 2007), ranks of nearest neighbors (Roy & Ghosh, 2020; Roy et al., 2021), and copula (Roy et al., 2020).

However, most of these existing methods do not have straightforward extensions for functional data. Lyons (2013) studied the applicability of the dCov test for strongly negative type metric spaces. Pan et al. (2020) proposed a test based on ball covariance (bCov) for Banach-valued random variables. These tests are based on pairwise distances. Lai et al. (2021) noted that distance-based methods like dCov and bCov do not consider the geometric structures of the observed data, whereas methods based on the inner-product can reveal more information in this regard. They proposed the angle covariance (aCov) test, which is based on pairwise inner-products. However, this test is applicable only for two random functions.

To overcome this limitation, we propose a general recipe for building a dependency measure for several random functions using pairwise inner-products. One can use a suitable estimator of this measure to construct a test. We particularly focus on one such measure that is based on the d-variate Hilbert Schmidt Independence Criterion (HSIC) (Pfister et al., 2018), and we call it Projected HSIC (pHSIC). Our measure has some nice theoretical properties, and the corresponding test is large sample consistent for fairly general alternatives. However, our measure and the associated test depend on a kernel that comes with an associated smoothing parameter known as the bandwidth. The performance of our test may depend on the kernel and its bandwidth parameter. For our numerical work, we use the Gaussian kernel and choose the bandwidth using the median heuristic approach (Gretton et al., 2007). We also establish the large sample consistency of the resulting tests against fixed alternatives. Extensive simulation studies are carried out to compare the performance of our tests with dCov, aCov, and bCov tests. The contents of this chapter are based on Banerjee & Ghosh (2022).

Chapter 7 of this thesis contains a brief summary of our work, and finally, we conclude with a short discussion on some possible directions for future research.

Chapter 2

Two-Sample Test for High-Dimensional Data

Consider two d-dimensional random vectors $\mathbf{X} \sim F$ and $\mathbf{Y} \sim G$ taking values on a separable metric space $(\mathbb{R}^d, \rho)$, where ρ is a metric on $\mathbb{R}^d$. Let $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\mathcal{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ be two sets of independent observations on $\mathbf{X}$ and $\mathbf{Y}$, respectively. In a two-sample problem, we use these observations to construct a test statistic for testing the null hypothesis $H_0 : F = G$ against the alternative hypothesis $H_1 : F \neq G$. This is a well-studied problem in statistics, and several tests are available in the literature. But as we have mentioned before, many of them either cannot be used or lead to poor performance for high-dimensional data, especially when the dimension is comparable to or larger than the sample size. In this chapter, we develop and investigate some tests that are not only consistent in the classical asymptotic regime, but also perform well in high dimensions.

Our tests are motivated by the well-known fact that two distributions F and G differ if and only if there exists a ball $\mathbb{B}(\mathbf{u}, \epsilon) := \{\mathbf{v} \in \mathbb{R}^d \mid \rho(\mathbf{v}, \mathbf{u}) \leq \epsilon\}$ such that $F(\mathbb{B}(\mathbf{u}, \epsilon)) \neq G(\mathbb{B}(\mathbf{u}, \epsilon))$. So, for any $\epsilon > 0$, $|F(\mathbb{B}(\mathbf{u}, \epsilon)) - G(\mathbb{B}(\mathbf{u}, \epsilon))|$ gives a measure of the difference between F and G in a neighborhood of $\mathbf{u} \in \mathbb{R}^d$. Therefore, we can choose $\mathbf{U}_i$ and $\mathbf{U}_j$ $(i \neq j = 1, \ldots, N)$ from the pooled sample $\mathcal{U} = \{\mathbf{U}_1 = \mathbf{X}_1, \ldots, \mathbf{U}_n = \mathbf{X}_n, \mathbf{U}_{n+1} = \mathbf{Y}_1, \ldots, \mathbf{U}_N = \mathbf{Y}_m\} = \mathcal{X} \cup \mathcal{Y}$ of size $N = m + n$ to construct the balls $\mathbb{B}_{ij} := \mathbb{B}(\mathbf{U}_i, \rho(\mathbf{U}_j, \mathbf{U}_i))$ and compute the differences $D_{ij} = \left|\hat{F}_{ij}(\mathbb{B}_{ij}) - \hat{G}_{ij}(\mathbb{B}_{ij})\right|$. Here $\hat{F}_{ij}$ and $\hat{G}_{ij}$ are the empirical analogs of F and G, obtained from $\mathcal{U}$ after removing $\mathbf{U}_i$ and $\mathbf{U}_j$ from the respective samples. One can use these differences to develop a test. For instance $T = \{N(N-1)\}^{-1} \sum_{i \neq j} D_{ij}^2$ can be used as the test statistics and H_0 can be rejected for large values of T. However, to reduce the computing cost, here we consider only those cases, where $\mathbf{U}_i$ and $\mathbf{U}_j$ come from the same distribution and use the test statistic

$$T_{n,m}^{\rho} = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \left\{ \frac{1}{n-2} \sum_{k=1, k \neq i,j}^{n} \delta(\mathbf{X}_k, \mathbf{X}_j, \mathbf{X}_i) - \frac{1}{m} \sum_{k=1}^{m} \delta(\mathbf{Y}_k, \mathbf{X}_j, \mathbf{X}_i) \right\}^2$$
$$+ \frac{1}{m(m-1)} \sum_{1 \leq i \neq j \leq m} \left\{ \frac{1}{n} \sum_{k=1}^{n} \delta(\mathbf{X}_k, \mathbf{Y}_j, \mathbf{Y}_i) - \frac{1}{m-2} \sum_{k=1, k \neq i,j}^{m} \delta(\mathbf{Y}_k, \mathbf{Y}_j, \mathbf{Y}_i) \right\}^2, \tag{2.1}$$

where $\delta(\mathbf{s}, \mathbf{u}, \mathbf{v}) = \mathrm{I}[\rho(\mathbf{s}, \mathbf{v}) \leq \rho(\mathbf{u}, \mathbf{v})]$, and $\mathrm{I}[\cdot]$ is the indicator function. Pan et al. (2018) proposed a similar test statistic, where they also considered the case $i = j$ while $\mathbf{U}_i$ and $\mathbf{U}_j$ were not removed from $\mathcal{U}$ for computing empirical analogs of $F(\mathbb{B}_{ij})$ and $G(\mathbb{B}_{ij})$. One can show that $T_{n,m}^{\rho}$ converges in probability (follows from Lemmas A2.1 and A2.2) to the ball divergence measure between F and

G, which can be expressed as

$$\Theta_\rho^2(F,G) = \int\int \{F(\mathbb{B}(\mathbf{u},\rho(\mathbf{v},\mathbf{u})) - G(\mathbb{B}(\mathbf{u},\rho(\mathbf{v},\mathbf{u}))\}^2 \{dF(\mathbf{u})dF(\mathbf{v}) + dG(\mathbf{u})dG(\mathbf{v})\}. \tag{2.2}$$

Ball divergence was first defined in Pan et al. (2018), where the authors proposed a two-sample test for Banach-valued random variables and studied its large sample properties. Their test statistic is slightly different from $T^\rho_{n,m}$. But for larger samples, they turn out to be equivalent.

Clearly, a large value of $T^\rho_{n,m}$ gives an evidence against $H_0 : F = G$, and therefore, we reject H_0 when $T^\rho_{n,m}$ exceeds a threshold. For a given level of significance α $(0 < \alpha < 1)$, this threshold (or cut-off) is computed using the permutation method. The algorithm is described below.

- Consider a permutation π of $\{1,\ldots,N\}$ and the corresponding permutation $\mathcal{U}_\pi = \{\mathbf{U}_{\pi(1)},\ldots,\mathbf{U}_{\pi(N)}\}$ of the pooled data $\mathcal{U}$.
- Use $\mathcal{X}_{n,\pi} = \{\mathbf{U}_{\pi(1)},\ldots,\mathbf{U}_{\pi(n)}\}$ and $\mathcal{Y}_{m,\pi} = \{\mathbf{U}_{\pi(n+1)},\ldots,\mathbf{U}_{\pi(n+m)}\}$ as the two samples to calculate $T^\rho_{n,m,\pi}$, the permutation analog of $T^\rho_{n,m}$.
- Repeat this method for all possible permutations. If $\mathcal{S}_N$ denotes the set of all permutations of $\{1,\ldots,N\}$, the critical value is given by

$$c_{1-\alpha} = \inf\{t \in \mathbb{R} : \frac{1}{N!}\sum_{\pi\in\mathcal{S}_N} \mathrm{I}[T^\rho_{n,m,\pi} \le t] \ge 1-\alpha\}.$$

We reject H_0 if $T^\rho_{n,m} > c_{1-\alpha}$ or the corresponding p-value $p_{n,m} = \frac{1}{N!}\sum_{\pi\in\mathcal{S}_N} \mathrm{I}[T^\rho_{n,m,\pi} \ge T^\rho_{n,m}]$ is smaller than α. The following lemma shows that this permutation method leads to a valid level α test irrespective of the sample sizes n, m and the dimension d.

Lemma 2.1. *Let $T^\rho_{n,m}$ be a two-sample test statistic computed based on n and m independent observations from d-dimensional distributions F and G, respectively. If $p_{n,m}$ denotes the corresponding permutation p-value, then under $H_0 : F = G$, we have $\mathbb{P}[p_{n,m} < \alpha] \le \alpha$ irrespective of the values of n, m and d.*

Note that Lemma 2.1 holds for any permutation test. Interestingly, for our test, the cut-off $c_{1-\alpha}$ can be upper bounded by a deterministic function of n and m that converges to zero as n and m diverge to infinity. This is asserted by the following lemma.

Lemma 2.2. *Let $\mathcal{X} = \{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$ and $\mathcal{Y} = \{\mathbf{Y}_1,\ldots,\mathbf{Y}_m\}$ be two sets of independent random vectors from two d-dimensional distributions F and G, respectively. For any α $(0 < \alpha < 1)$, the inequality $0 < c_{1-\alpha} \le 2/\big(3\alpha(\min\{n,m\} - 2)\big)$ holds with probability one.*

Note that this upper bound on $c_{1-\alpha}$ is of the order $O((\min\{n,m\})^{-1})$, and it does not depend on d. So, $c_{1-\alpha}$ converges to 0 as $\min\{n,m\}$ diverges to infinity. Therefore, under the alternative hypothesis $H_1 : F \ne G$, if $T^\rho_{m,n}$ converges to a positive constant, the power of the test converges to one. Theorem 2.1 shows this large sample consistency of the permutation test.

Theorem 2.1. *If $\Theta^2_\rho(F,G) > 0$, the power of the level α $(0 < \alpha < 1)$ test based on $T^\rho_{m,n}$ converges to 1 as $\min\{n,m\}$ increases to infinity.*

If $(\mathbb{R}^d, \rho)$ is a finite dimensional separable metric space, $\Theta^2_\rho(F,G) = 0$ if and if only $F = G$. So, under $H_1 : F \neq G$, we have $\Theta^2_\rho(F,G) > 0$, and this proves the large sample consistency of the proposed test under general alternatives. For computing $c_{1-\alpha}$, instead of considering all $N!$ permutations of $\{1,\ldots,N\}$, it is enough to consider all possible subsets of $\mathcal{U}$ of size n. But, if n and m are moderately large, it may not be computationally feasible to consider all $\binom{N}{n}$ subsets or all $N!$ permutations. In such cases, we generate B random permutations $\pi_1,\ldots,\pi_B$ and reject H_0 if the corresponding p-value $p_{n,m,B} = \frac{1}{B+1}\Big\{\sum_{i=1}^B \mathrm{I}[T^\rho_{n,m,\pi_i} \geq T^\rho_{n,m}] + 1\Big\}$ is smaller than α. Recall that the use of all $N!$ permutations leads to the p-value $p_{n,m} = \frac{1}{N!}\Big\{\sum_{\pi\in\mathcal{S}_N} \mathrm{I}[T^\rho_{n,m,\pi} \geq T^\rho_{n,m}]\Big\}$. As B increases $p_{n,m,B} - p_{n,m}$ converges to 0 almost surely (see Lemma 2.3). This justifies the implementation of the test based on random permutations.

Lemma 2.3. *Given the pooled sample $\mathcal{U}$, $|p_{n,m,B} - p_{n,m}| \xrightarrow{a.s.} 0$ as $B \to \infty$.*

Though Pan et al. (2018) also suggested implementing their test using the permutation method, they proved the consistency of their test based on the large sample distribution of the test statistic. The consistency of the permutation test was missing. Moreover, they did not investigate the high-dimensional behavior of their test, which is the main focus of this chapter.

2.1 Behaviour of the proposed test in HDLSS setup

In this section, we study the high-dimensional behaviour of the proposed test when the dimension of the data grows to infinity while the sample sizes remain fixed. Note that the behaviour of the proposed test may depend on the metric ρ. Since the ℓ_2 distance is arguably the most popular choice as the distance function on $\mathbb{R}^d$, we first consider the test based on this distance.

2.1.1 Test based on the ℓ_2 distance

The test statistic based on the ℓ_2 distance, denoted by $T^{\ell_2}_{n,m}$, is obtained replacing $\delta(\mathbf{s},\mathbf{u},\mathbf{v})$ in $T^\rho_{n,m}$ (see Equation (2.1)) by $\mathrm{I}[\|\mathbf{s}-\mathbf{v}\| \leq \|\mathbf{u}-\mathbf{v}\|]$. To investigate the behaviour of the resulting test, we consider the following assumptions.

(A2.1) If $\mathbf{X}_1, \mathbf{X}_2 \overset{iid}{\sim} F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{iid}{\sim} G$ are independent, for $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2$, $\mathbf{X}_1 - \mathbf{Y}_1$ and $\mathbf{Y}_1 - \mathbf{Y}_2$, $\left|d^{-1}\|\mathbf{W}\|^2 - d^{-1}\mathbb{E}(\|\mathbf{W}\|^2)\right| \xrightarrow{P} 0$ as $d \to \infty$.

(A2.2) There exist constants ν^2, σ_F^2, and σ_G^2 such that $d^{-1}\|\boldsymbol{\mu}_F - \boldsymbol{\mu}_G\|^2 \to \nu^2$, $d^{-1}\mathrm{trace}(\boldsymbol{\Sigma}_F) \to \sigma_F^2$, and $d^{-1}\mathrm{trace}(\boldsymbol{\Sigma}_G) \to \sigma_G^2$ as $d \to \infty$. (Here $\boldsymbol{\mu}_F = \mathbb{E}(\mathbf{X})$, $\boldsymbol{\mu}_G = \mathbb{E}(\mathbf{Y})$, $\boldsymbol{\Sigma}_F = \mathrm{Var}(\mathbf{X})$ and $\boldsymbol{\Sigma}_G = \mathrm{Var}(\mathbf{Y})$.)

These two assumptions are quite common in the HDLSS literature. While (A2.1) gives the Weak Law of Large Numbers (WLLN) for the sequence of possibly dependent and non-identically distributed random variables $\{(\mathbf{W}^{(q)})^2 : q \geq 1\}$ (i.e., $\left|\frac{1}{d}\sum_{q=1}^d (W^{(q)})^2 - \mathbb{E}\frac{1}{d}\sum_{q=1}^d (W^{(q)})^2\right| \xrightarrow{P} 0$

as $d \to \infty$), (A2.2) gives the limiting value of $d^{-1}\mathbb{E}\|\mathbf{W}\|^2$ and hence that of $d^{-1}\|\mathbf{W}\|^2$ depending on whether $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2, \mathbf{Y}_1 - \mathbf{Y}_2$ or $\mathbf{X}_1 - \mathbf{Y}_1$. In addition to (A2.2), Hall, Marron & Neeman (2005) assumed uniformly bounded fourth moments and a ρ-mixing property for the coordinate variables to investigate the high-dimensional behavior of some popular classifiers. The weak law (A2.1) holds under those conditions. However, instead of ρ-mixing, it is enough to assume $\sum_{i\neq j}\mathrm{Cov}(W^{(i)}, W^{(j)}) = o(d^2)$ for WLLN (see Sarkar, Biswas & Ghosh, 2020). One can also assume (A2.2) and sufficient moment conditions like $\mathrm{Var}(\|\mathbf{X} - \boldsymbol{\mu}_F\|^2) = o(d^2)$, $\mathrm{Var}(\|\mathbf{Y} - \boldsymbol{\mu}_G\|^2) = o(d^2)$, $\mathrm{trace}(\boldsymbol{\Sigma}_F^2) = o(d^2)$ and $\mathrm{trace}(\boldsymbol{\Sigma}_G^2) = o(d^2)$. In addition to assuming uniformly bounded fourth moments and a ρ-mixing condition on the standardized variables, following Ahn et al. (2007) and Jung & Marron (2009) one can also assume a sphericity condition (i.e. $\mathrm{trace}(\boldsymbol{\Sigma}^2)/(\mathrm{trace}(\boldsymbol{\Sigma}))^2 \to 0$ as $d \to \infty$ both for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_F$ and $\boldsymbol{\Sigma}_G$) for WLLN. The non-strongly spiked eigenvalue (NSSE) model (see Aoshima & Yata, 2018) (where $\lambda_{max}^2(\boldsymbol{\Sigma})/\mathrm{trace}(\boldsymbol{\Sigma}^2) \to 0$ as $d \to \infty$, for $\lambda_{max}(\boldsymbol{\Sigma})$ being the largest eigenvalue of $\boldsymbol{\Sigma}$) satisfies the sphericity condition. Yata & Aoshima (2012, 2020) also assumed similar conditions. Under those conditions, (A2.1) holds for $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2$ and $\mathbf{W} = \mathbf{Y}_1 - \mathbf{Y}_2$. Under these assumptions, we have the following lemma.

Lemma 2.4. *Suppose that* $\mathbf{X}_1, \mathbf{X}_2 \overset{iid}{\sim} F$ *and* $\mathbf{Y}_1, \mathbf{Y}_2 \overset{iid}{\sim} G$ *are independent. If* F *and* G *satisfy assumptions (A2.1) and (A2.2), then* $d^{-1/2}\|\mathbf{X}_1 - \mathbf{X}_2\| \xrightarrow{P} \sigma_F\sqrt{2}$, $d^{-1/2}\|\mathbf{Y}_1 - \mathbf{Y}_2\| \xrightarrow{P} \sigma_G\sqrt{2}$ *and* $d^{-1/2}\|\mathbf{X}_1 - \mathbf{Y}_1\| \xrightarrow{P} \sqrt{\sigma_G^2 + \sigma_F^2 + \nu^2}$ *as* $d \to \infty$.

These distance convergence results can be used to show that if $\nu^2 > 0$ or $\sigma_F^2 \neq \sigma_G^2$, then $\mathbb{P}(T_{n,m}^{\ell_2} > 1/3)$ converges to 1 as d grows to infinity.

Lemma 2.5. *Assume that the two distributions* F *and* G *satisfy assumptions (A2.1) and (A2.2). If* $\nu^2 + (\sigma_F - \sigma_G)^2 > 0$, *we have* $\lim_{d\to\infty}\mathbb{P}\{T_{n,m}^{\ell_2} > 1/3\} = 1$.

In Lemma 2.2, we have already seen that the critical value of the permutation test $c_{1-\alpha}$ is smaller than $\frac{2}{3\alpha}\big(\frac{1}{\min\{n,m\}-2}\big)$ with probability one, which in turn is smaller than $1/3$ if $\min\{n, m\} > 2 + 2/\alpha$. So, in view of Lemma 2.5, the resulting test has the high-dimensional consistency if $\min\{n, m\} - 2 > 2/\alpha$. This result is stated as Theorem 2.2 below.

Theorem 2.2. *Assume that* F *and* G *satisfy (A2.1)-(A2.2). If* $\nu^2 + (\sigma_F - \sigma_G)^2 > 0$ *and* $\min\{n, m\} \geq 2 + 2/\alpha$, *the power of the level* α $(0 < \alpha < 1)$ *test based on* $T_{n,m}^{\ell_2}$ *converges to one as the dimension* d *increases to infinity.*

Theorem 2.2 gives a sufficient condition for the consistency of the proposed test in HDLSS asymptotic regime. It shows that if F and G differ in their locations ($\nu^2 > 0$) or scales ($\sigma_F^2 \neq \sigma_G^2$), the test based on $T_{n,m}^{\ell_2}$ turns out to be consistent.

Now, to study the empirical performance of the proposed test, let us consider three simple examples, each involving two multivariate normal distributions.

Example 2.1. *Two distributions* $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ *and* $G = \mathcal{N}_d(0.15\,\mathbf{1}_d, \mathbf{I}_d)$ *differ only in their means. Here* $\mathbf{0}_d = (0, \ldots, 0)^\top, \mathbf{1}_d = (1, \ldots, 1)^\top \in \mathbb{R}^d$, $\mathbf{I}_d$ *is the* $d \times d$ *identity matrix and* $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *denotes the* d*-variate normal distribution with the mean vector* $\boldsymbol{\mu}$ *and the dispersion matrix* $\boldsymbol{\Sigma}$.

Example 2.2. *Two distributions* $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ *and* $G = \mathcal{N}_d(\mathbf{0}_d, 1.1\mathbf{I}_d)$ *have the same location, but they differ in their scales.*

Example 2.3. *Both* $F = \mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma}_{1,d})$ *and* $G = \mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma}_{2,d})$ *have diagonal dispersion matrices. The first* $d/2$ *diagonal elements of* $\boldsymbol{\Sigma}_{1,d}$ *are* 1 *and the rest are* 2. *On the contrary,* $\boldsymbol{\Sigma}_{2,d}$ *has the first* $d/2$ *diagonal elements equal to* 2 *and the rest equal to* 1.

For each example, we considered 10 different choices of d ($d = 2^i$ for $i = 1, \ldots, 10$), and in each case, we used the test based on 100 observations (50 from each distribution). This process was repeated 500 times to estimate the power of the test by the proportion of times it rejected H_0. In Examples 2.1 and 2.2, we have $\nu^2 + (\sigma_F - \sigma_G)^2 > 0$. So, as one would expect in view of Theorem 2.2, the power of the proposed test increased with the dimension (see Figure 2.1). However, in Example 2.3, we have $\nu^2 = 0$ and $\sigma_F = \sigma_G$. So, the sufficient conditions for consistency (see Theorem 2.2) do not hold. In this example, the proposed test had a poor performance. To overcome this limitation of the test based on the ℓ_2 distance, in the next subsection, we use different distance functions to construct the test statistic and study the high-dimensional performance of the resulting tests.

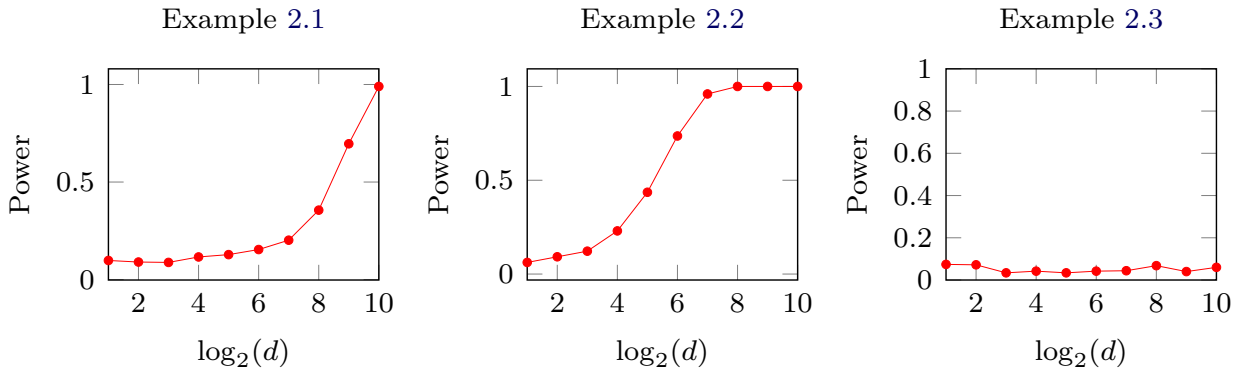


Fig. 2.1 *Powers of the permutation test based on* $T^{\ell_2}_{n,m}$ *in Examples 2.1-2.3.*

2.1.2 Tests based on generalized distances

Instead of ℓ_2 distance, we can consider the generalized distance function proposed by Sarkar & Ghosh (2018a). The generalized distance between two d-dimensional observations $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})^\top$ and $\mathbf{y} = (y^{(1)}, \ldots, y^{(d)})^\top$ is given by

$$\varphi_{h,\psi}(\mathbf{x}, \mathbf{y}) = h\Big\{\frac{1}{d}\sum_{q=1}^{d} \psi(|x^{(q)} - y^{(q)}|^2)\Big\},$$

where $h, \psi : \mathbb{R}_+ \to \mathbb{R}_+$ are continuous, monotonically increasing functions with $h(0) = \psi(0) = 0$. Note that all ℓ_p distances (with $p \geq 1$) are special cases of $\varphi_{h,\psi}$ (up to a multiplicative constant).

Using $\varphi_{h,\psi}$, we construct the generalized test statistic $T^{h,\psi}_{n,m}$ (replace $\delta(\mathbf{s},\mathbf{u},\mathbf{v})$ by $\mathrm{I}[\varphi_{h,\psi}(\mathbf{s},\mathbf{v}) \leq \varphi_{h,\psi}(\mathbf{u},\mathbf{v})]$ in the expression of $T^{\rho}_{n,m}$ in (2.1)) and reject H_0 for large values of it. The cut-off is chosen using the permutation method as before.

High-dimensional behavior of this test can be studied under an assumption similar to (A2.1). Recall that (A2.1) gives WLLN for $\{(W^{(q)})^2 : \ q \geq 1\}$ with $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2, \mathbf{X}_1 - \mathbf{Y}_1$ and $\mathbf{Y}_1 - \mathbf{Y}_2$. Here we consider a similar assumption for the random variables $\{\psi(W^{(q)})^2 : \ q \geq 1\}$.

(A2.3) If $\mathbf{X}_1, \mathbf{X}_2 \overset{iid}{\sim} F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{iid}{\sim} G$ are independent, for $W = \mathbf{X}_1 - \mathbf{X}_2$, $\mathbf{X}_1 - \mathbf{Y}_1$ and $\mathbf{Y}_1 - \mathbf{Y}_2$, (i) $\limsup_{d\to\infty} d^{-1} \sum_{q=1}^{d} \mathbb{E}\psi(|W^{(q)}|^2) < \infty$ and (ii) $d^{-1} \sum_{q=1}^{d} \{\psi(|W^{(q)}|^2) - \mathbb{E}\psi(|W^{(q)}|^2)\} \xrightarrow{P} 0$ as $d \to \infty$.

Note that if the $W^{(q)}$'s are independent or m-dependent or the $\psi(|W^{(q)}|^2)$'s satisfy the ρ-mixing property, (A2.3) holds if the $\psi(|W^{(q)}|^2)$'s have uniformly bounded second moments. If ψ is bounded, the moment condition gets automatically satisfied. Now, define

$$\varphi^*_{h,\psi}(F,F) = h\{d^{-1}\sum_{q=1}^{d} \mathbb{E}\psi(|X_1^{(q)} - X_2^{(q)}|^2)\},$$

$$\varphi^*_{h,\psi}(G,G) = h\{d^{-1}\sum_{q=1}^{d} \mathbb{E}\psi(|Y_1^{(q)} - Y_2^{(q)}|^2)\},$$

$$\varphi^*_{h,\psi}(F,G) = h\{d^{-1}\sum_{q=1}^{d} \mathbb{E}\psi(|X_1^{(q)} - Y_1^{(q)}|^2)\}.$$

There is an interesting lemma due to Sarkar & Ghosh (2018a) (see Lemma 1 in that article) involving these three quantities. The lemma is stated below.

Lemma 2.6. *Suppose that h is concave and ψ has a non-constant completely monotone derivative. Then for any fixed d, we have $e_{h,\psi}(F,G) := 2\varphi^*_{h,\psi}(F,G) - \varphi^*_{h,\psi}(F,F) - \varphi^*_{h,\psi}(G,G) \geq 0$, where the equality holds if and only if F and G have the same one-dimensional marginal distributions.*

Note that $e_{h,\psi}(F,G)$ can be viewed as an energy distance (see, e.g., Aslan & Zech, 2005) between F and G. In view of Lemma 2.6, for appropriate choices of h and ψ, it is somewhat reasonable to assume that under $H_1 : F \neq G$, the limiting energy distance between F and G remains bounded away from 0 (i.e., $\liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$). Under this assumption, we can establish the consistency of the test based on $T^{h,\psi}_{m,n}$ in the HDLSS asymptotic regime. This result is given by the following theorem.

Theorem 2.3. *Suppose that F and G satisfy (A2.3) and $\liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$. If $\min\{n,m\} \geq 2 + 2/\alpha$, the power of the level α $(0 < \alpha < 1)$ test based on $T^{h,\psi}_{n,m}$ converges to one as the dimension d increases to infinity.*

For $h(t) = \sqrt{t}$ and $\psi(t) = t$, $\varphi_{h,\psi}$ turns out to be the ℓ_2 metric (up to a multiplicative constant), but this choice of ψ does not have a non-constant completely monotone derivative as

mentioned in Lemma 2.6. But there are other choices of ψ that satisfy this property. For instance, one can use $\psi_1(t) = \sqrt{t}$, $\psi_2(t) = 1 - e^{-t/2}$ or $\psi_3(t) = \log(1+t)$ with $h(t) = t$ in all these cases. For ψ_1 and ψ_2, $\varphi_{h,\psi}$ turns out to be a distance function (ψ_1 leads to a scalar multiple of ℓ_1 distance), but that is not the case for ψ_3. In that case, we can call it a dissimilarity measure. Note that for the ℓ_2 distance, under (A2.2), we have $\liminf_{d\to\infty} e_{h,\psi}(F,G) = 2\sqrt{\nu^2 + \sigma_F^2 + \sigma_G^2} - \sigma_F\sqrt{2} - \sigma_G\sqrt{2}$, which is positive if and only if either $\nu^2 > 0$ or $\sigma_F \neq \sigma_G$, i.e. $\nu^2 + (\sigma_F - \sigma_G)^2 > 0$. In Example 2.3, we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$, but $\liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$ for $\psi = \psi_1, \psi_2, \psi_3$ and $h(t) = t$. So, while the test based on $T^{\ell_2}_{n,m}$ had poor performance, those based on $T^{h,\psi}_{n,m}$ with $\psi = \psi_1, \psi_2$ and ψ_3 (henceforth referred to as $T^{\ell_1}_{n,m}$, $T^{\exp}_{n,m}$ and $T^{\log}_{n,m}$, respectively) performed well (see Figure 2.2). For further study on the high-dimensional behavior of these tests, we consider two other examples.

Example 2.4. *Both F and G have i.i.d. coordinate variables. While in F, they follow the standard univariate Cauchy distribution, in G, each has a location shift of one unit.*

Example 2.5. *Two distributions $F = \mathcal{N}_d(\boldsymbol{\mu}_d, \mathbf{I}_d)$ and $G = \mathcal{N}_d(-\boldsymbol{\mu}_d, \mathbf{I}_d)$ differ only in their means. Here $\|\boldsymbol{\mu}_d\| = \|d^{-1/2}\mathbf{1}_d\| = 1$ for all values of d.*

We considered 10 different choices of d, and in each case, we generated 50 observations from each distribution. The process was repeated 500 times as before to estimate the powers of the tests.

In Example 2.4, the ball divergence tests based on the test statistics $T^{\ell_1}_{n,m}$ and $T^{\ell_2}_{n,m}$ (henceforth referred to as BD-ℓ_1 and BD-ℓ_2 tests) did not work well, but those based on $T^{\exp}_{n,m}$ and $T^{\log}_{n,m}$ (henceforth referred to as BD-exp and BD-log tests) had excellent performance (see Figure 2.2). Among them, BD-exp had an edge. Note that in this example, the coordinate variables in F and G do not have finite moments. That affected the performance of BD-ℓ_1 and BD-ℓ_2. Since $\psi_2(t) = 1 - e^{-t/2}$ is a bounded function, we do not have such problems for BD-exp. Assumption (A2.3) holds for this choice of ψ. It holds for $\psi_3(t) = \log(1+t)$ as well. This was the reason behind the good performance of the tests based on these two choices of ψ.

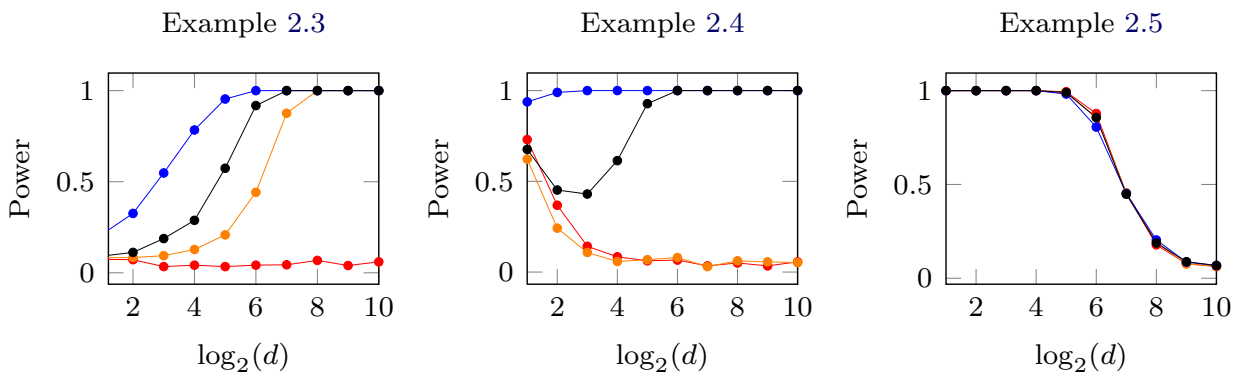


Fig. 2.2 *Powers of BD-ℓ_2(●) BD-ℓ_1(●) BD-exp(●) and BD-log(●) tests in Examples 2.3-2.5.*

In Example 2.5, the Mahalanobis distance between F and G remains the same for all values of d, and we have $\lim_{d\to\infty} e_{h,\psi}(F,G) = 0$ for all choices of h and ψ considered here. So, as expected, the powers of all tests gradually dropped as d increased. In such situations, for good performance of the proposed tests, we need to increase the sample sizes appropriately with the dimensions.

2.2 What happens in HDHSS regime?

In this section, we deal with the cases where F and G gradually become close as d increases. For such shrinking alternatives, the power of any test based on fixed sample sizes is expected to go down as d increases. Therefore, to achieve better performance in high-dimension, one needs to increase the sample sizes as well. Now, one may be curious to know whether it is possible to increase n and m with d at an appropriate rate such that one can construct a valid level α $(0 < \alpha < 1)$ test with power converging to 1 as the dimension increases. We show that this is possible for our tests as long as $\Theta^2_\rho(F,G)$ shirks to 0 at an appropriately slower rate. This is obtained by establishing the minimax rate optimality of our tests.

2.2.1 Minimax rate optimality

Consider the hypotheses $H_0' : \Theta^2_{\ell_2}(F,G) = 0$ and $H_1' : \Theta^2_{\ell_2}(F,G) > \epsilon$ for some $\epsilon > 0$. Define $\mathbb{P}^{(n,m)}_{F,G}$ as the joint distribution of $\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{Y_1}, \ldots, \mathbf{Y}_m$, where $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} F$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \overset{iid}{\sim} G$. Let $\mathcal{F}(\epsilon) := \{(F,G) \mid \Theta^2_{\ell_2}(F,G) > \epsilon\}$ denote the class of alternatives, and for a given $\alpha \in (0,1)$, $\mathbb{T}_{n,m,d}(\alpha)$ denote the class of all level α test functions $\phi : \mathcal{U} \to \{0,1\}$. The minimax type II error rate for this class is defined as

$$R_{n,m,d}(\epsilon) = \inf_{\phi \in \mathbb{T}_{n,m,d}(\alpha)} \sup_{(F,G)\in\mathcal{F}(\epsilon)} \mathbb{P}^{(n,m)}_{F,G}(\phi = 0).$$

Now, we want to find an $\epsilon_0 = \epsilon_0(n,m,d)$, for which the following conditions hold.

(a) For any $0 < \zeta < 1-\alpha$, there exists a constant $c(\alpha,\zeta) > 0$ such that for all $0 < c < c(\alpha,\zeta)$, we have $\liminf_{n,m,d\to\infty} R_{n,m,d}(c\ \epsilon_0(n,m,d)) \geq \zeta$.

(b) There exists a level α test ϕ_0 such that for any $0 < \zeta < 1-\alpha$, we can find $C(\alpha,\zeta) > 0$ for which $\limsup_{n,m,d\to\infty} \sup_{(F,G)\in\mathcal{F}(c\ \epsilon_0(n,m,d))} \mathbb{P}^{(n,m)}_{F,G}(\phi_0 = 0) \leq \zeta\ \forall c > C(\alpha,\zeta)$, i.e.,

$$\limsup_{n,m,d\to\infty} R_{n,m,d}(c\ \epsilon_0(n,m,d)) \leq \zeta\ \forall\ c > C(\alpha,\zeta).$$

The rate $\epsilon_0(n,m,d)$ (which is unique up to a constant) is called the minimax rate of separation, and ϕ_0 is called the minimax rate optimal test. Theorem 2.4 shows that if ϵ is of smaller order than $(1/\sqrt{n} + 1/\sqrt{m})^2$, for all level α tests, the maximum type II error is bounded away from 0.

Theorem 2.4. *For $0 < \zeta < 1-\alpha$, there exists a constant $c_0(\alpha,\zeta)$ such that for $\lambda(n,m) = (1/\sqrt{n} + 1/\sqrt{m})^2$, the minimax type II error $R_{n,m,d}(c\lambda(n,m))$ is lower bounded by ζ for all $0 < c < c_0(\alpha,\zeta)$.*

This shows that the minimax rate of separation cannot be smaller than $O\big((1/\sqrt{n}+1/\sqrt{m})^2\big)$. Now, we show that for $\epsilon_0(n,m,d) = \lambda(n,m)$, the test based on $T^{\ell_2}_{n,m}$ satisfies condition (b).

Theorem 2.5. *For $0 < \zeta < 1-\alpha$, there exists a constant $C_0(\alpha,\zeta)$ such that asymptotically the maximum type II error of the test based on $T^{\ell_2}_{n,m}$ over $\mathcal{F}(c\lambda(n,m))$ is uniformly bounded above by ζ for all $c > C_0(\alpha,\zeta)$, i.e.,*

$$\limsup_{n,m,d\to\infty} \sup_{(F,G)\in\mathcal{F}(c\lambda(n,m))} P^{(n,m)}_{F,G}(T_{n,m} \le c_{1-\alpha}) \le \zeta \quad \text{for all } c > C_0(\alpha,\zeta).$$

Theorems 2.4 and 2.5 together show that the minimax rate of separation $\epsilon_0(n,m,d) = (1/\sqrt{n}+1/\sqrt{m})^2$ does not depend on the dimension, and they also establish the minimax rate optimality of the permutation test based on $T^{\ell_2}_{n,m}$ for the class of alternatives $\mathcal{F}(\epsilon)$.

2.2.2 Performance under shrinking alternatives

Theorem 2.5 gives us a lower bound $\lambda(n,m)$ on the rate of $\Theta^2_{\ell_2}(F,G)$ that enables us to detect the difference between F and G using the permutation test based on $T^{\ell_2}_{n,m}$. If we increase the sample sizes with the dimension such that $\Theta^2_{\ell_2}(F,G)$ converges to zero slower than $\lambda(n,m)$ (i.e., $\Theta^2_{\ell_2}(F,G)/\lambda(n,m) \to \infty$ as $d\to\infty$), the test based on $T^{\ell_2}_{n,m}$ turns out to be consistent. This result is asserted by the following theorem.

Theorem 2.6. *If n and m, the sample sizes from F and G, grow as a function of the dimension d in such a way that $\lim_{d\to\infty} \Theta^2_{\ell_2}(F,G)/\lambda(n,m) = \infty$, then, the power of the level α $(0<\alpha<1)$ test based on $T^{\ell_2}_{n,m}$ converges to one as dimension increases to infinity.*

If $\liminf_{d\to\infty} \Theta^2_{\ell_2}(F,G) > 0$, the assumption in Theorem 2.6 holds even if n and m grow very slowly with d. In such cases, one can expect good results even in the HDLSS setup, and we have observed the same in our numerical studies. It is easy to show that if F and G satisfy the conditions of Theorem 2.2, we have $\liminf_{d\to\infty} \Theta^2_{\ell_2}(F,G) \ge 1/3$. As expected, in such cases, we also have the consistency of the test in the HDHSS regime, where m and n also increase with d.

So far, we have discussed the minimax rate optimality of the test based on $T^{\ell_2}_{n,m}$ and established its consistency for shrinking alternatives. The results similar to Theorems 2.4-2.6 hold even when the test is constructed based on other distance functions considered in Section 2.1.2. We state the result below as Theorem 2.7, but we skip the details of the proof since they are exactly the same as in the case of the test based on the ℓ_2 distance.

Theorem 2.7. *Let $h,\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous, monotonically increasing functions with $h(0) = \psi(0) = 0$. Assume that n and m, the sample sizes from F and G, grow as functions of the dimension d in such a way that $\lim_{d\to\infty} \Theta^2_{\varphi_{h,\psi}}(F,G)/\lambda(n,m) = \infty$. Then the power of the level α $(0<\alpha<1)$ test based on $T^{h,\psi}_{n,m}$ converges to one as d increases to infinity.*

Remark 2.1. *In the HDLSS setup, we need Assumptions (A2.1)-(A2.2) for the consistency of the tests. But in the HDHSS setup, when n and m grow with d, such assumptions are not needed.*

Also, unlike the HDLSS setup, in the HDHSS setup, we have consistency for the test based on the ℓ_p-distance for all $p \geq 1$.

Remark 2.2. *Theorems 2.6 and 2.7 remain silent about the asymptotic behaviour of the proposed test when $\lim_{d\to\infty} \Theta^2_{\ell_2}(F,G)/\lambda(n,m) = c$ (or $\lim_{d\to\infty} \Theta^2_{h,\psi}(F,G)/\lambda(n,m) = c$) for some $c \in (0,\infty)$. However, in such cases, one can show that the asymptotic power of the test has a lower bound $1 - (C_1 c + C_2)/\big(c - \frac{1}{3\alpha}\big)^2$, where C_1 and C_2 are two universal constants (see the proof of Theorem 2.5).*

Now, consider a simple example involving two multivariate normal distributions $F = \otimes_{i=1}^d \mathcal{N}_1(1/d^\beta, 1)$ and $G = \otimes_{i=1}^d \mathcal{N}_1(-1/d^\beta, 1)$, where β is a positive constant. Note that as d grows to infinity, here we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$ and $\lim_{d\to\infty} e_{h,\psi}(F,G) = 0$ for all h and ψ considered in this chapter. So, the conditions for the HDLSS consistency of the tests are not satisfied. Now, we study the behaviour of the BD-ℓ_2 test when the sample sizes increase with the dimension at the rate $O(d^\gamma)$ for some $\gamma > 0$. We find out the relation between γ and β that leads to the consistency of the test in the HDHSS setup. Our findings are summarized in the following proposition.

Proposition 2.1. *Suppose that n and m are the sample sizes from $F = \otimes_{i=1}^d \mathcal{N}_1(1/d^\beta, 1)$ and $G = \otimes_{i=1}^d \mathcal{N}_1(-1/d^\beta, 1)$, respectively. If $\beta > 0$ and $n \asymp m \asymp d^\gamma$ for some $\gamma > 0$, then, for the ball divergence test based on $T^{\ell_2}_{n,m}$, we have the following results.*

(a) If $\beta \leq 1/4$, for any $\gamma > 0$, the test is consistent.

(b) If $1/4 < \beta \leq 1/2$, the test is consistent if $\gamma > 4\beta - 1$.

(c) If $\beta > 1/2$, the test is consistent if $\gamma > 2\beta$. If $\gamma < 2\beta - 1$, there exist no level α $(0 < \alpha < 1)$ tests with asymptotic power more than α.

Proposition 2.1(c) says that if $\beta > 1/2$, for the HDHSS consistency of any test, one needs to increase the sample size at a rate faster than $O(d^{2\beta-1})$. So, the HDLSS consistency is not possible in this case. Recall that in Example 2.5, we have $\beta = 1/2$. Therefore, if we increase n and m at a rate faster than $O(d)$, our test will be consistent. We confirm it in our numerical study.

For this study, we used three different choices of β $(0.2, 0.3$ and $0.5)$, and in each case, seven different choices of γ $(0, 0.4, 0.5, 0.6, 0.9.1$ and $1.1)$ and 10 different values of d (2^i for $i = 1, \ldots, 10$) were considered. We took $n = m = 5 + \lfloor d^\gamma \rfloor$ to ensure $n, m \geq 5$, and each experiment was repeated 500 times to compute the power of the test based on $T^{\ell_2}_{n,m}$. The results are reported in Figure 2.3. In this example, for higher values of β, $\Theta^2_{\ell_2}(F,G)$ converges to zero at a faster rate. Therefore, to discriminate between F and G, we need to increase the sample sizes at a higher rate as well. Figure 2.3 shows that for higher values of β, the tests corresponding to lower values of γ performed poorly. Note that here $\gamma = 0$ represents the HDLSS scenario. We can see that for $\beta = 0.2$, even for $m = n = 6$ (i.e., $\gamma = 0$), the power of our test converged to 1 in high dimensions. This was expected in view of Proposition 2.1(a). As expected, the test had higher power for larger values of γ. For $\beta = 0.3$ and 0.5, it did not work well in the HDLSS setup, but when m and n increased

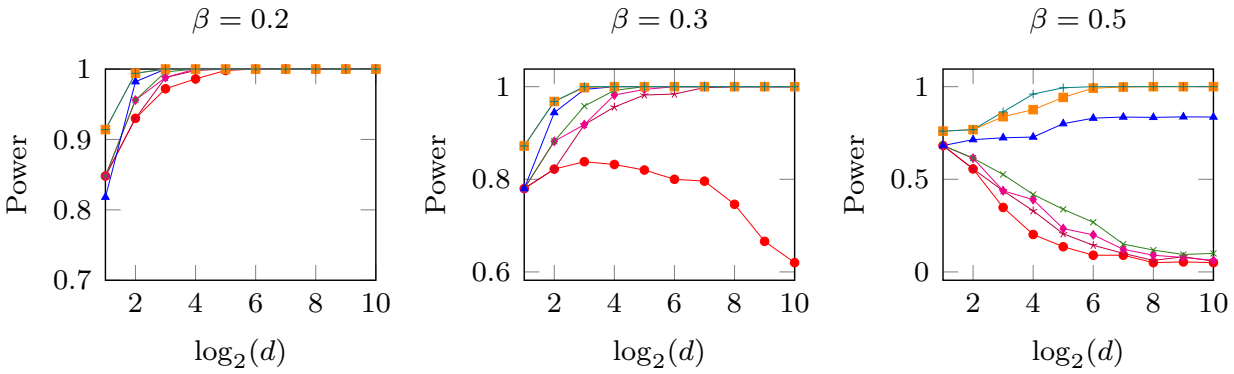


Fig. 2.3 *Powers of the BD-ℓ_2 test for different choice of β (0.2, 0.3 and 0.5) and γ (0 (●), 0.4 (⋆), 0.5 (♦), 0.6 (×), 0.9 (▲), 1 (■), 1.1 (+)).*

with d at an appropriate rate, the power of the test converged to unity as we expect in view of Proposition 2.1(b)-(c).

2.3 Empirical performance of the proposed tests

In this section, we compare the empirical performance of our tests with some popular tests. For this comparison, we consider the multivariate run tests based on minimum spanning tree (Friedman & Rafsky, 1979) and shortest Hamiltonian path (Biswas, Mukhopadhyay & Ghosh, 2014), the tests based on averages of inter-point distances proposed by Baringhaus & Franz (2004) and Biswas & Ghosh (2014), the nearest neighbor test (Schilling, 1986; Henze, 1988), and the test based on maximum mean discrepancy (Gretton et al., 2012). Henceforth, we shall refer to them as the FR test, the SHP test, the BF test, the BG test, the NN test, and the MMD test, respectively. For the NN test, we consider the test based on 3 neighbors, which has been reported to perform well (see, e.g., Schilling, 1986). Throughout this chapter, all tests are considered to have the 5% nominal level. The SHP test has the distribution-free property. For all other tests, the cut-off is computed based on 500 random permutations.

2.3.1 Analysis of simulated data sets

First, we study the level properties of our tests. We generated two sets of independent observations from $\mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and used them as observations from F and G, respectively. This experiment was repeated 500 times, and for each test, we computed the proportion of times it rejected H_0. We carried out our experiment for different sample sizes ($n = m = 20$, 35 and 50) and dimension ($d = 2^i$ for $i = 1, 2, \ldots, 10$). Figure 2.4 shows that on all occasions, the BD-ℓ_2 test rejected H_0 in nearly 5% of the cases. BD-ℓ_1, BD-exp, BD-log, and other competing tests also exhibited similar level properties. But, to avoid repetition, we decided not to report them.

Next, we investigate the power properties of the proposed tests. We consider two types of examples. First, we deal with examples with fixed n and m and study the performance of

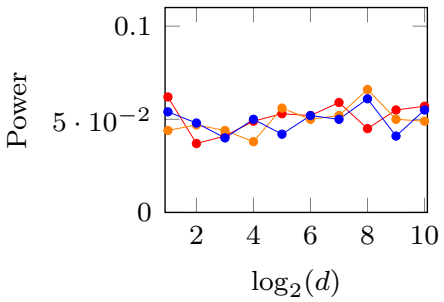


Fig. 2.4 *Observed levels of the BD-ℓ_2 test for $n = m = 20$ (●), 35 (●) and 50 (●).*

different tests as the d increases. Next, we consider the situations where the conditions for HDLSS consistency of the proposed tests do not hold (i.e., we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$ and $\lim e_{h,\psi}(F, G) = 0$). In such cases, we investigate the performance of different tests when n and m grow with d.

Dimension increases when the sample sizes remain fixed

We begin with Examples 2.1-2.4 discussed in Sections 2.1.1 and 2.1.2. For each example, the powers of the proposed tests and other competing tests are computed based on 500 repetitions of the experiment, and they are reported in Figure 2.5.

In the location problem in Example 2.1, BF and MMD tests outperformed all other tests considered here. However, the powers of the proposed tests were comparable to the rest of the competing tests (BG, NN, FR, and SHP tests).

In Example 2.2, all tests based on ball divergence and the BG test had similar performance, and they performed much better than their competitors. Among the rest, the SHP test had a relatively higher power. In high dimensions, FR and NN tests had powers close to 0. Biswas, Mukhopadhyay & Ghosh (2014) and Mondal, Biswas & Ghosh (2015) explained the reasons for such poor performance of FR and NN tests in high-dimensional scale problems.

In Example 2.3, we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$, but $\liminf_{d\to\infty} e_{h,\psi}(F, G) > 0$ for $\psi = \psi_1, \psi_2, \psi_3$ with $h(t) = t$. So, as expected, BD-ℓ_2 did not have satisfactory performance, but BD-ℓ_1, BD-exp and BD-log performed well in high-dimensions. Among them, BD-exp had a clear edge. Unlike these three tests, the powers of other competing methods did not increase with the dimension. Note that these competing methods are based on ℓ_2 distances. The use of a different distance function may improve their performance.

In the presence of heavy-tailed distributions in Example 2.4, all tests except BD-exp and BD-log had poor performance in high dimensions. Among these two tests, the one based on bounded ψ-function (i.e., $\psi_2(t) = 1 - e^{-t/2}$) performed better. We also observed this in Section 2.1.2.

As we have discussed before, in Example 2.5, the power of any test based on fixed sample sizes is expected to decrease as the dimension increases. We also observed the same for all tests

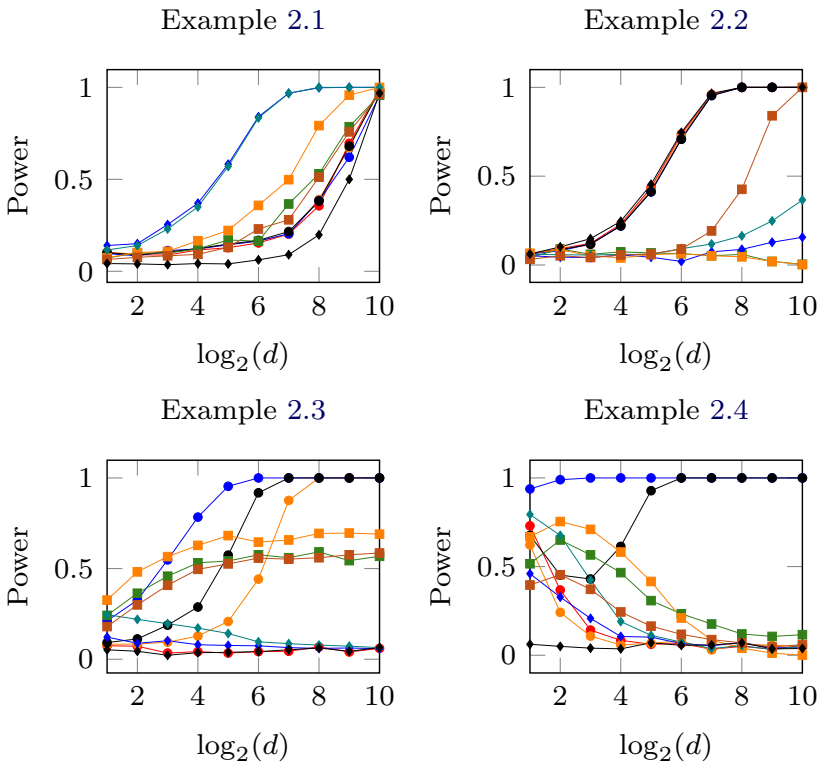


Fig. 2.5 *Powers of BD-ℓ_2 (●), BD-ℓ_1 (●), BD-*exp *(●), BD-*log *(●), FR (■), BF (◆), NN (■), MMD (◆), SHP (■) and BG (◆) tests in Examples 2.1-2.4.*

considered here. So, we do not report those results here. Instead, we consider three other examples (Examples 2.6-2.8) as mentioned below.

Example 2.6. *Here F is the d-variate standard normal distribution while G is an equal mixture of $\mathcal{N}_d(0.5\,\mathbf{1}_d, \mathbf{I}_d)$ and $\mathcal{N}_d(-0.5\,\mathbf{1}_d, \mathbf{I}_d)$.*

Example 2.7. *F is same as in Example 2.6, but G is mixture of $\mathcal{N}_d(\mathbf{1}_d, \mathbf{I}_d)$ and $\mathcal{N}_d(-0.25\,\mathbf{1}_d, \mathbf{I}_d)$ with mixing proportions 0.2 and 0.8, respectively.*

Example 2.8. *Both F and G have i.i.d. coordinate variables. The coordinate variables in F follow $\mathcal{N}_1(0, 2)$ distribution, but in G, they follow the standard t distribution with 4 degrees of freedom.*

For each example, we generated 50 observations from each distribution and considered 10 different choices of d (2^i for $i = 1, \ldots, 10$) as before. Each experiment was repeated 500 times to estimate the power of different tests, and they are shown in Figure 2.6. This figure clearly shows that in the examples involving mixture normal distributions, the BG test and the ball divergence tests performed better than their competitors. In Example 2.8, the ball divergence tests outperformed all other competing tests considered here. Like Example 2.3, here also, we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$, but $\liminf_{d\to\infty} e_{h,\psi}(F, G) > 0$ for other three choices of ψ. So, as expected, BD-ℓ_1, BD-exp and BD-log performed much better than BD-ℓ_2.

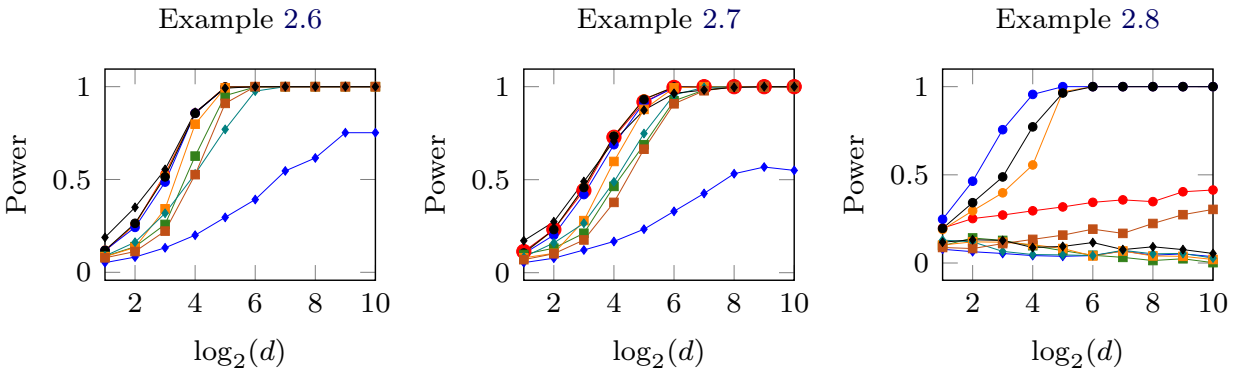


Fig. 2.6 *Powers of BD-ℓ_2 (●), BD-ℓ_1 (●), BD-*exp *(●), BD-*log *(●), FR (■), BF (◆), NN (■), MMD (◆), SHP (■), BG (◆) tests in Examples 2.6-2.8.*

Sample sizes grow with the dimension

Now, we deal with some examples where we do not have a theoretical guarantee for the consistency of the ball divergence tests in the HDLSS regime, and we investigate how the proposed tests and their competitors perform when the sample sizes also grow with the dimension. As before, the powers of all tests were computed based on 500 replications. We consider six examples. Descriptions of the first three examples are given below. In Example 2.9, we consider the sample sizes $n = m = 5 + \lfloor \sqrt{d} \rfloor$ while in Examples 2.10 and 2.11, we have $n = m = d + 5$.

Example 2.9. *The coordinate variables in F and G are i.i.d. $\mathcal{N}_1(d^{-0.3}, 1)$ and $\mathcal{N}_1(-d^{-0.3}, 1)$, respectively.*

Example 2.10. *Both $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}^{\circ}_{1,d})$ and $G = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}^{\circ}_{2,d})$ have the same mean $\mathbf{0}_d$, but different diagonal dispersion matrices. The first $d/2$ diagonal elements of $\mathbf{\Sigma}^{\circ}_{1,d}$ are 1, and the rest are 5. On the contrary, $\mathbf{\Sigma}^{\circ}_{2,d}$ has the first $d/2$ diagonal elements equal to 5 and rest equal to 1.*

Example 2.11. *Here $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}^{*}_{1,d})$ and $G = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}^{*}_{2,d})$ have the same mean, but different dispersion matrices $\mathbf{\Sigma}^{*}_{1,d} = ((0.1^{|i-j|}))$ and $\mathbf{\Sigma}^{*}_{2,d} = ((0.5^{|i-j|}))$.*

In Example 2.9, BF and MMD tests had the best performance (see Figure 2.7), closely followed by NN, BG, and ball divergence tests. The SHP test had relatively low power. Interestingly, the powers of all competing tests converged to one as the dimension increased.

Example 2.10 is similar to Example 2.3 though the parameters are different. In this example, NN, FR, and BD-exp tests had better performance than their competitors, with the BD-exp test having an edge (see Figure 2.7). Interestingly, the powers of all tests, barring the BG test, showed a tendency to converge to unity as the dimension increased. Note that this was not the case in Example 2.3 when samples of fixed sizes were used.

In Example 2.11, two distributions have the same mean and marginal variances, but they differ in their correlation structures. In this example, we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$ and

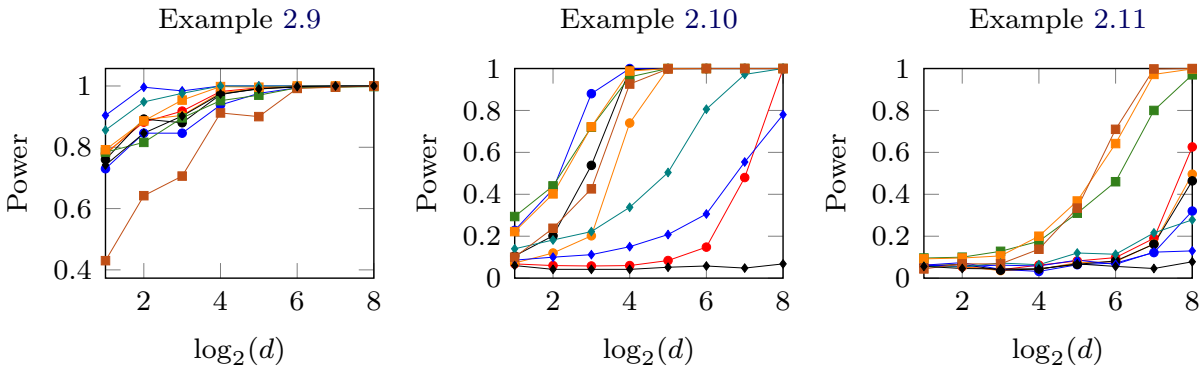


Fig. 2.7 *Powers of BD-ℓ_2 (●), BD-ℓ_1 (●), BD-*exp *(●), BD-*log *(●), FR (■), BF (◆), NN (■), MMD (◆), SHP (■), BG (◆) tests in Examples 2.9-2.11.*

$\lim_{d\to\infty} e_{h,\phi}(F,G) = 0$ for all three choices of ψ (i.e., ψ_1, ψ_2 and ψ_3). Here, graph-based tests performed much better than average distance-based tests. However, unlike BG, BF, and MMD tests, the powers of the ball divergence tests had a sharp rise in higher dimensions.

Finally, we consider three examples involving sparse alternatives, where F and G differ only in $\lfloor d^\beta \rfloor$ many coordinates for $\beta \in (0,1)$. In these examples, we have $e_{h,\psi}(F,G) \asymp d^{\beta-1}$ for all choices of ψ considered here. For our numerical study, we use $\beta = 0.7$ and $n = m = 5 + \lfloor \sqrt{d} \rfloor$.

Example 2.12. *Two distributions $F = \mathcal{N}_d(\boldsymbol{\mu}_d, \mathbf{I}_d)$ and $G = \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ differ only in their locations. The first $\lceil d^\beta \rceil$ coordinates of $\boldsymbol{\mu}_d$ are 2, and the rest are zero.*

Example 2.13. *Two distributions $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and $G = \mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma}_d)$ differ only in their scales. Here $\boldsymbol{\Sigma}_d$ is a diagonal matrix with first $\lceil d^\beta \rceil$ entries equal to 5, and the rest equal to 1.*

Example 2.14. *The distribution G differs from $F = \mathcal{N}_d(\mathbf{0}_d, 2\mathbf{I}_d)$ only in the first $\lceil d^\beta \rceil$ coordinates. These coordinate variables are independent, and they follow t distribution with 4 degrees of freedom.*

Figure 2.8 shows the powers of different tests in these three examples. In the location and scale problems in Examples 2.12 and 2.13, our findings were similar to those observed in Examples

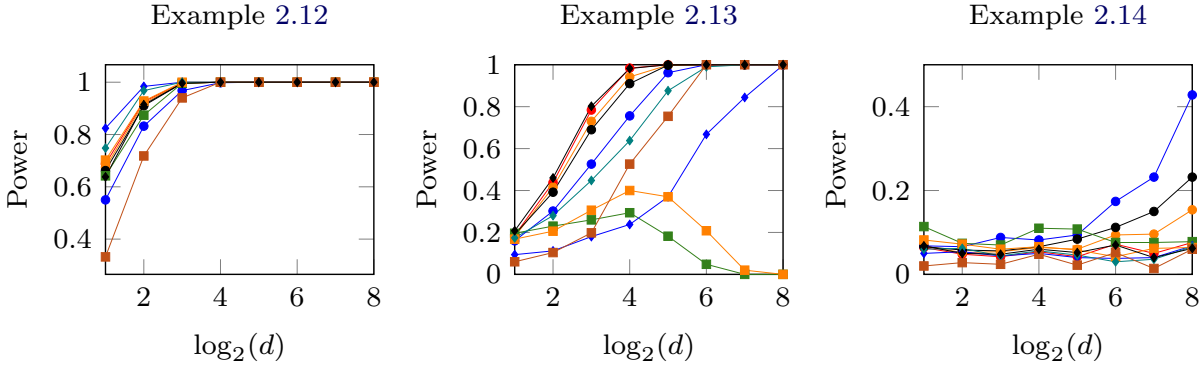


Fig. 2.8 *Powers of BD-ℓ_2 (●), BD-ℓ_1 (●), BD-*exp *(●), BD-*log *(●), FR (■), BF (◆), NN (■), MMD (◆), SHP (■), BG (◆) tests in Examples 2.12-2.14.*

2.1 and 2.2, respectively. In Example 2.12, BF and MMD tests had an edge, but the performance of the proposed tests was competitive with the rest. The SHP test had relatively low power. In Example 2.13, the BG test and the ball divergence tests outperformed their competitors in higher dimensions. In this scale problem, FR and NN tests had poor performance. Example 2.14 can be viewed as a sparse version of Example 2.8, where the two distributions differ only in their shapes. In this example, all tests based on the ℓ_2 distance failed to have satisfactory performance, but the ball divergence tests based on other distance functions performed better. The powers of these tests showed an upward trend with increasing dimensions.

Analysis of datasets generated from strongly spiked eigenvalue model

Next, we consider two examples involving data sets generated using strongly spiked eigenvalue (SSE) models (see Aoshima & Yata, 2018). In particular, we consider a scale problem (Example 2.15) and a location problem (Example 2.16) and investigate the performance of different methods when the sample sizes remain fixed (50 from each distribution), and the dimension increases.

Example 2.15. *Two probability distributions* $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}_d^\circ(1.1))$ *and* $G = \mathcal{N}_d(0, \mathbf{\Sigma}_d^\circ(1.5))$ *differ only in the scale of the first coordinate variable. Here* $\mathbf{\Sigma}_d^\circ(\gamma)$ *denotes the* $d \times d$ *diagonal matrix with the first diagonal entry* d^γ *for some* $\gamma > 0$ *and the rest equal to unity.*

Example 2.16. *Here* $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}_d^\circ(1.5))$ *and* $G = \mathcal{N}_d(0.5\mathbf{1}_d, \mathbf{\Sigma}_d^\circ(1.5))$ *differ only in their locations. Here* $\mathbf{\Sigma}_d^\circ(\gamma)$ *has the same meaning as in Example 2.15, and* $\mathbf{1}_d = (1, 1, \ldots, 1)^\top$.

Since $\lim \lambda_{max}^2(\mathbf{\Sigma}_d^\circ(\gamma))/\text{trace}(\mathbf{\Sigma}_d^{\circ^2}(\gamma)) = \lim d^{2\gamma}/(d^{2\gamma} + d - 1) = 1$ for any $\gamma > 1$, both examples belong to SSE models. For each of these examples, we considered 10 different choices of d (2^i for $i = 1, \ldots, 10$), and in each case, we repeated the experiment 500 times to estimate the power of the tests by the proportion of times they rejected the null hypothesis $H_0 : F = G$. These results are summarized in Figure 2.9.

In Example 2.15, $(X_{11} - X_{21}) \sim \mathcal{N}_1(0, 2d^{1.1})$ and $(X_{1i} - X_{2i}) \sim \mathcal{N}_1(0, 2)$ for $i = 2, \ldots, d$. Hence, as $d \to \infty$, $\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|^2/2d^{1.1} \xrightarrow{D} \chi_1^2$ (converges in distribution to a chi-square variable with one degree of freedom). Similarly, as $d \to \infty$, we have $\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|^2/2d^{1.5} \xrightarrow{D} \chi_1^2$, $\|\boldsymbol{X}_1 - \boldsymbol{Y}_1\|^2/d^{1.5} \xrightarrow{D} \chi_1^2$, and hence $\mathbb{P}\{\|\boldsymbol{X}_1 - \boldsymbol{X}_2\| \leq \|\boldsymbol{X}_1 - \boldsymbol{Y}_1\|\} \to 1$. Thus, by Lemma A2.4, we have the HDLSS consistency of the BD-ℓ_2 test. One can use similar arguments to show the consistency of BD-ℓ_1 test as well. This was the reason behind the excellent performance of these tests. In this example, the BG test performed best, followed by BD-ℓ_2 and BD-ℓ_1 tests. Except for BD-exp and BD-log, the powers of all other tests also converged to 1 as the d increased. In cases of these two tests, the pairwise distances $\sum_{i=1}^d \psi\big((X_{1i} - X_{2i})^2\big)/d$, $\sum_{i=1}^d \psi\big((Y_{1i} - Y_{2i})^2/2\big)/d$ and $\sum_{i=1}^d \psi\big((X_{1i} - Y_{1i})^2/2\big)/d$ converge in probability to the same positive constant as d goes to infinity. So, unlike BD-ℓ_2 and BD-ℓ_1, they were unable to extract substantial discriminatory information from the first coordinate.

In Example 2.16, powers of BD-ℓ_2, BF, BG, and MMD tests dropped down as d increased, but those for the graph-based tests increased steadily. Note that while the performance of the graph-based tests depends on the ordering of the pairwise distances, that of the above-mentioned four tests depends on their magnitudes. In this example, though the inter-sample distances had a tendency to take higher values than the intra-sample distances, as d diverges to infinity, $\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|^2/2d^{1.5}$, $\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|^2/2d^{1.5}$ and $\|\boldsymbol{X}_1 - \boldsymbol{Y}_1\|^2/2d^{1.5}$ all converge in distribution to a chi-square random variable with one degree of freedom, and that is why they failed to discriminate among the two populations in higher dimensions. However, BD-exp and BD-log tests outperformed all graph-based tests in this example. Note that here $\sum_{i=1}^{d} \psi\big((X_{1i} - X_{2i})^2\big)/d$ and $\sum_{i=1}^{d} \psi\big((Y_{1i} - Y_{2i})^2/2\big)/d$ converge in probability to the same limit, but $\sum_{i=1}^{d} \psi\big((X_{1i} - Y_{1i})^2/2\big)/d$ converges in probability to a limit higher than that. This explains the reason behind the excellent performance of these tests. Because of the same reason, the BD-ℓ_1 test also had increasing power. However, in this example, the first coordinate difference was the dominating term in the ℓ_1 distance, and along that coordinate, we had very little difference between the two populations. This somewhat affected the performance of the BD-ℓ_1 test. In this case, coordinate-wise standardization of the variables may improve the performance of this test. Similar coordinate-wise standardization may also lead to better performance by the BD-ℓ_2 test.

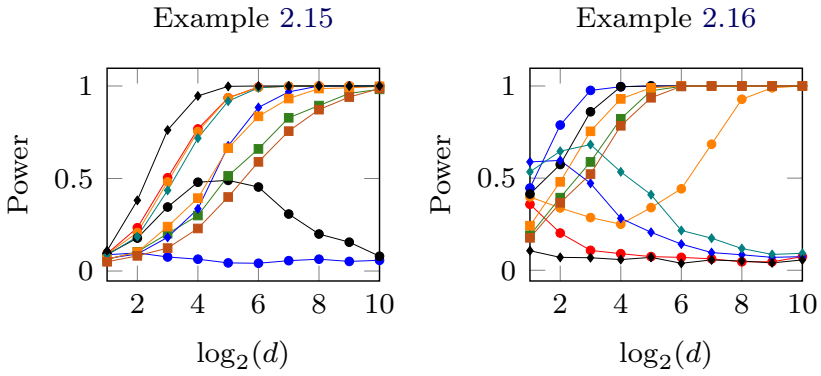


Fig. 2.9 *Powers of BD-ℓ_2 (●), BD-ℓ_1 (●), BD-*exp *(●), BD-*log *(●), FR (■), BF (◆), NN (■), MMD (◆), SHP (■), BG (◆) tests in Examples 2.15 and 2.16.*

2.3.2 Analysis of benchmark data sets

For further evaluation of the performance of different tests, we analyzed two real data sets, namely Colon data and Lightning-2 data. The Colon data set contains expression levels of 2000 genes in 40 'tumor' and 22 'normal' colon tissue samples that were analyzed with an Affymetrix oligonucleotide array. This data set is available in the R package 'rda', and its description can be found in Alon et al. (1999). Lightning 2 data set contains 637-dimensional observations from two populations with respective sample sizes 48 and 73. It is available at the UCR Time Series Classification Archive, and its description can also be found in Sarkar, Biswas & Ghosh (2020). These two data sets have

been extensively studied in the classification literature, and it is well known that in each of these data sets, there is a reasonable separation between the two distributions. So, we can assume the alternative hypothesis $H_1 : F \neq G$ to be true, and different tests can be compared based on their powers. However, when we used the full data set for testing, all tests rejected H_0 for both of these data sets. Using that single experiment based on the full data set, it was not possible to compare among different test procedures. So, we generated random sub-samples from the entire data set, keeping the sample proportions from the two distributions approximately the same as they are in the original data. Different tests were implemented using these sub-samples, and this procedure was repeated 500 times to estimate their powers. The results for different sub-sample sizes are reported in Figure 2.10.

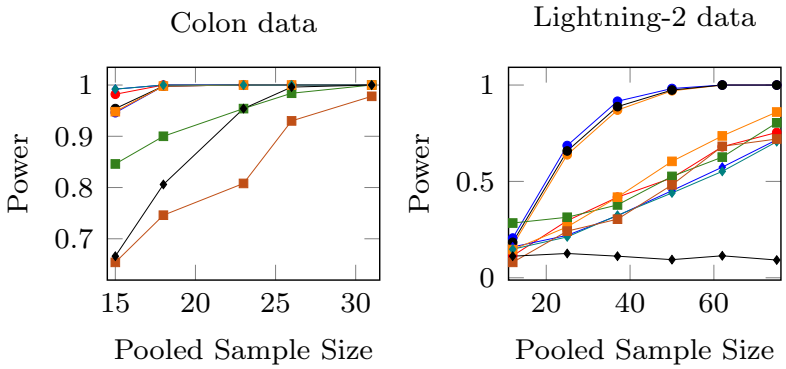


Fig. 2.10 *Powers of BD-ℓ_2 (●), BD-ℓ_1 (●), BD-*exp *(●), BD-*log *(●), FR (■), BF(◆), NN(■), MMD(◆), SHP(■), BG(◆) tests in benchmark data sets.*

In the case of Colon data, BF, MMD, and BD-ℓ_2 tests had very high powers even when the pooled sample size was 15. These three tests had comparable performance, and they performed better than others. BD-ℓ_1, BD-exp, and BD-log tests also had competitive performances. Like BF, MMD, and NN tests, they also had unit power for samples of size 18 or higher. FR, BG, and SHP tests had relatively low powers in this data set.

Figure 2.10 clearly shows the superiority of the ball divergence tests BD-ℓ_1, BD-exp and BD-log in the case of Lightning-2 data. These three tests had much higher powers compared to the rest for samples with a combined sample size larger than 20. Among the other methods, the NN test had the best overall performance. BF, MMD, FR, SHP, and BD-ℓ_2 tests also had similar powers. The BG test performed poorly in this data set.

2.4 Proofs and Mathematical details

Proof of Lemma 2.1 . Consider a random permutations π of $\{1, 2, \ldots, N = n+m\}$ and let $T^{\rho}_{n,m,\pi}$ be the permuted test statistic (permutation analogs of $T^{\rho}_{n,m}$). Let $c_{1-\alpha}$ be the upper α-th quantile of the distribution of $T^{\rho}_{n,m,\pi}$ given the pooled data $\mathcal{U} := \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m\}$. Note

that $c_{1-\alpha}$ is permutation invariant and under $H_0 : F = G$, $(T^{\rho}_{n,m}, c_{1-\alpha})$ and $(T^{\rho}_{n,m,\pi}, c_{1-\alpha})$ are identically distributed, irrespective of the values of n, m and d. Hence, we have

$$\mathbb{P}[p_{n,m} < \alpha] = \mathbb{P}[T^{\rho}_{n,m} > c_{1-\alpha}] = \mathbb{P}\left[T^{\rho}_{n,m,\pi} > c_{1-\alpha}\right] = \mathbb{E}\left[\mathbb{P}\Big\{T^{\rho}_{n,m,\pi} > c_{1-\alpha} \mid \mathcal{U}\Big\}\right] \leq \mathbb{E}\left[\alpha\right] \leq \alpha.$$

The second last inequality follows from the definition of $c_{1-\alpha}$. Hence, the levels of the permutation tests are controlled at α. ■

Lemma A2.1. *If* $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} F$ *and* $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \overset{iid}{\sim} G$ *are independent random vectors, then* $\mathbb{E}\{T^{\rho}_{n,m}\} = \frac{1}{6}\left(\frac{1}{n-2} + \frac{1}{m-2}\right) + \frac{1}{m}(p_0 - p_1) + \frac{1}{n}(p_2 - p_3) + \Theta^2_{\rho}(F, G)$, *where* $\Theta^2_{\rho}(F, G)$ *is the ball divergence measure defined in* (2.2)*, and* p_0, p_1, p_2, p_3 *are given by*

$$p_0 = \mathbb{P}\big\{\rho(\boldsymbol{Y}_1, \boldsymbol{X}_1) \leq \rho(\boldsymbol{X}_2, \boldsymbol{X}_1)\big\}, \qquad p_1 = \mathbb{P}\big\{\rho(\boldsymbol{Y}_1, \boldsymbol{X}_1) \leq \rho(\boldsymbol{X}_2, \boldsymbol{X}_1); \rho(\boldsymbol{Y}_2, \boldsymbol{X}_1) \leq \rho(\boldsymbol{X}_2, \boldsymbol{X}_1)\big\},$$
$$p_2 = \mathbb{P}\big\{\rho(\boldsymbol{X}_1, \boldsymbol{Y}_1) \leq \rho(\boldsymbol{Y}_2, \boldsymbol{Y}_1)\big\} \text{ and } \qquad p_3 = \mathbb{P}\big\{\rho(\boldsymbol{X}_1, \boldsymbol{Y}_1) \leq \rho(\boldsymbol{Y}_2, \boldsymbol{Y}_1); \rho(\boldsymbol{X}_2, \boldsymbol{Y}_1) \leq \rho(\boldsymbol{Y}_2, \boldsymbol{Y}_1)\big\}.$$

Proof. Note that $T^{\rho}_{n,m}$ can be written as $T^{\rho}_{n,m} = V_1 + V_2$, where

$$V_1 = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \left\{\frac{1}{n-2} \sum_{k=1, k\neq i,j}^{n} \delta(\mathbf{X}_k, \mathbf{X}_j, \mathbf{X}_i) - \frac{1}{m}\sum_{k=1}^{m} \delta(\mathbf{Y}_k, \mathbf{X}_j, \mathbf{X}_i)\right\}^2,$$
$$V_2 = \frac{1}{m(m-1)} \sum_{1 \leq i \neq j \leq m} \left\{\frac{1}{n} \sum_{k=1}^{n} \delta(\mathbf{X}_k, \mathbf{Y}_j, \mathbf{Y}_i) - \frac{1}{m-2}\sum_{k=1, k\neq i,j}^{m} \delta(\mathbf{Y}_k, \mathbf{Y}_j, \mathbf{Y}_i)\right\}^2,$$

and $\delta(\mathbf{s}, \mathbf{u}, \mathbf{v}) = \mathrm{I}[\rho(\mathbf{s}, \mathbf{v}) \leq \rho(\mathbf{u}, \mathbf{v})]$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\{V_1\} &= \frac{1}{(n-2)^2}\mathbb{E}\Big\{\big(\sum_{k=3}^{n} \delta(\mathbf{X}_k, \mathbf{X}_2, \mathbf{X}_1)\big)^2\Big\} + \frac{1}{m^2}\mathbb{E}\Big\{\big(\sum_{k=1}^{m} \delta(\mathbf{Y}_k, \mathbf{X}_2, \mathbf{X}_1)\big)^2\Big\} \\
&\quad - \frac{2}{m(n-2)}\mathbb{E}\Big\{\big(\sum_{k=3}^{n} \delta(\mathbf{X}_k, \mathbf{X}_2, \mathbf{X}_1)\big)\big(\sum_{k=1}^{m} \delta(\mathbf{Y}_k, \mathbf{X}_2, \mathbf{X}_1)\big)\Big\} \\
&= \frac{1}{(n-2)^2}\mathbb{E}\Big\{\sum_{k=3}^{n} \delta(\mathbf{X}_k, \mathbf{X}_2, \mathbf{X}_1) + \sum_{k,l=3, k\neq l}^{n} \delta(\mathbf{X}_k, \mathbf{X}_2, \mathbf{X}_1)\delta(\mathbf{X}_l, \mathbf{X}_2, \mathbf{X}_1)\Big\} \\
&\quad + \frac{1}{m^2}\mathbb{E}\Big\{\sum_{k=3}^{m} \delta(\mathbf{Y}_k, \mathbf{X}_2, \mathbf{X}_1) + \sum_{k,l=3, k\neq l}^{m} \delta(\mathbf{Y}_k, \mathbf{X}_2, \mathbf{X}_1)\delta(\mathbf{Y}_l, \mathbf{X}_2, \mathbf{X}_1)\Big\} \\
&\quad - \frac{2}{m(n-2)}\mathbb{E}\Big\{\sum_{k=3}^{n}\sum_{l=1}^{m} \delta(\mathbf{X}_k, \mathbf{X}_2, \mathbf{X}_1)\delta(\mathbf{Y}_l, \mathbf{X}_2, \mathbf{X}_1)\Big\}. \\
&= \frac{1}{n-2}\,\mathbb{P}\Big\{\rho(\mathbf{X}_3, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1)\Big\} + \frac{n-3}{n-2}\,\mathbb{P}\Big\{\rho(\mathbf{X}_3, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1); \rho(\mathbf{X}_4, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1)\Big\} \\
&\quad + \frac{1}{m}\,\mathbb{P}\Big\{\rho(\mathbf{Y}_1, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1)\Big\} + \frac{m-1}{m}\,\mathbb{P}\Big\{\rho(\mathbf{Y}_1, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1); \rho(\mathbf{Y}_2, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1)\Big\} \\
&\quad - 2\,\mathbb{P}\Big\{\rho(\mathbf{X}_3, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1); \rho(\mathbf{Y}_1, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1)\Big\} \\
&= \frac{1}{(n-2)}\Big\{\frac{1}{2} + \frac{n-3}{3}\Big\} + \frac{1}{m}\Big\{p_0 + (m-1)p_1\Big\} - 2p_4 = \frac{1}{3} + \frac{1}{6(n-2)} + \frac{1}{m}(p_0 - p_1) + p_1 - 2p_4,
\end{aligned}$$

where p_0 and p_1 are as before and $p_4 = \mathbb{P}\{\rho(\mathbf{X}_3, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1); \rho(\mathbf{Y}_1, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1)\}$.

Similarly, one can show that $\mathbb{E}\{V_2\} = \frac{1}{3} + \frac{1}{6(m-2)} + \frac{1}{n}(p_2 - p_3) + p_3 - 2p_5$, where p_2 and p_3 are as

before and $p_5 = \mathbb{P}\{\rho(\mathbf{Y}_3, \mathbf{Y}_1) \leq \rho(\mathbf{Y}_2, \mathbf{Y}_1); \rho(\mathbf{X}_1, \mathbf{Y}_1) \leq \rho(\mathbf{Y}_2, \mathbf{Y}_1)\}$. Hence, we have

$$\mathbb{E}\{T^{\rho}_{n,m}\} = \frac{1}{6(n-2)} + \frac{1}{6(m-2)} + \frac{1}{m}(p_0 - p_1) + \frac{1}{n}(p_2 - p_3) + \frac{2}{3} + p_1 - 2p_4 + p_3 - 2p_5.$$

Now observe that

$$\begin{aligned}
\Theta^2_{\rho}(F,G) = & \int \{F(B(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))) - G(B(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u})))\}^2 \left[dF(\boldsymbol{u})dF(\boldsymbol{v}) + dG(\boldsymbol{u})dG(\boldsymbol{v})\right] \\
= & \int F^2(B(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))dF(\boldsymbol{u})dF(\boldsymbol{v}) + \int G^2(B(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))dF(\boldsymbol{u})dF(\boldsymbol{v}) \\
& - 2\int F(B(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))G(B(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))dF(\boldsymbol{u})dF(\boldsymbol{v}) \\
& + \int F^2(B(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))dG(\boldsymbol{u})dG(\boldsymbol{v}) + \int G^2(B(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))dG(\boldsymbol{u})dG(\boldsymbol{v}) \\
& - 2\int F(B(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))G(B(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))dG(\boldsymbol{u})dG(\boldsymbol{v}) \\
= & \frac{1}{3} - 2p_4 + p_1 + p_3 - 2p_5 + \frac{1}{3} = p_1 + p_3 - 2p_4 - 2p_5 + \frac{2}{3}.
\end{aligned}$$

Hence, we have $\mathbb{E}\{T^{\rho}_{n,m}\} = \frac{1}{6}\left(\frac{1}{n-2} + \frac{1}{m-2}\right) + \frac{1}{m}(p_0 - p_1) + \frac{1}{n}(p_5 - p_3) + \Theta^2_{\rho}(F,G)$. ■

Lemma A2.2. *If* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} F$ *and* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m \overset{iid}{\sim} G$ *are independent random vectors, then* $Var(T^{\rho}_{n,m}) \leq C_0\Theta^2_{\rho}(F,G)\left(\frac{1}{n} + \frac{1}{m}\right) + C_1\left(\frac{1}{n} + \frac{1}{m}\right)^2$, *where the constants* C_0 *and* C_1 *are independent of the dimension* d.

Proof. Note that $\Theta^2_{\rho}(F,G)$ can be written as $\Theta^2_{\rho}(F,G) = A_1 + A_2$, where

$$\begin{aligned}
A_1 = & \int\int \{F(\mathbb{B}(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u})) - G(\mathbb{B}(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))\}^2 dF(\boldsymbol{u})dF(\boldsymbol{v}) \\
= & \mathbb{E}\Big(F(B(\mathbf{X}_1, \rho(\mathbf{X}_2, \mathbf{X}_1))) - G(B(\mathbf{X}_1, \rho(\mathbf{X}_2, \mathbf{X}_1)))\Big)^2, \text{ and} \\
A_2 = & \int\int \{F(\mathbb{B}(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u})) - G(\mathbb{B}(\boldsymbol{u}, \rho(\boldsymbol{v}, \boldsymbol{u}))\}^2 dG(\boldsymbol{u})dG(\boldsymbol{v}) \\
= & \mathbb{E}\Big(F(B(\mathbf{Y}_1, \rho(\mathbf{Y}_2, \mathbf{Y}_1))) - G(B(\mathbf{Y}_1, \rho(\mathbf{Y}_2, \mathbf{Y}_1)))\Big)^2.
\end{aligned}$$

It can be verified that V_1 (as defined in the proof of Lemma A2.1) can be expressed as

$$\begin{aligned}
V_1 = & \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \Big\{\frac{1}{n-2} \sum_{k=1, k\neq i,j}^{n} \delta(\mathbf{X}_k, \mathbf{X}_j, \mathbf{X}_i) - \frac{1}{m}\sum_{k=1}^{m} \delta(\mathbf{Y}_k, \mathbf{X}_j, \mathbf{X}_i)\Big\}^2 \\
= & \frac{1}{n(n-1)(n-2)^2m^2} \sum_{i\neq j=1}^{n} \sum_{u,u' \neq i,j}^{n} \sum_{v,v'=1}^{m} \Big\{\delta(\mathbf{X}_u, \mathbf{X}_j, \mathbf{X}_i)\delta(\mathbf{X}_{u'}, \mathbf{X}_j, \mathbf{X}_i) \\
& \qquad + \delta(\mathbf{Y}_v, \mathbf{X}_j, \mathbf{X}_i)\delta(\mathbf{Y}_{v'}, \mathbf{X}_j, \mathbf{X}_i) - \delta(\mathbf{X}_{u'}, \mathbf{X}_j, \mathbf{X}_i)\delta(\mathbf{Y}_v, \mathbf{X}_j, \mathbf{X}_i) \\
& \qquad\qquad - \delta(\mathbf{X}_u, \mathbf{X}_j, \mathbf{X}_i)\delta(\mathbf{Y}_{v'}, \mathbf{X}_j, \mathbf{X}_i)\Big\} \\
= & \frac{1}{n(n-1)(n-2)^2m^2} \sum_{i\neq j=1}^{n} \sum_{u,u' \neq i,j}^{n} \sum_{v,v'=1}^{m} \psi_{A_1}(\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_u, \mathbf{X}_{u'}; \mathbf{Y}_v, \mathbf{Y}_{v'}), \text{ say.}
\end{aligned}$$

Clearly, V_1 can be written as a linear combination of U-statistics of different degrees. Let $\hat{U}_{A_1}^{(4,2)}$ be the U-statistic with the kernel function $\psi_{A_1}(\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_u, \mathbf{X}_{u'}; \mathbf{Y}_v, \mathbf{Y}_{v'})$, which has the same degree as V_1. So, it determines the order of $\mathrm{Var}(V_1)$. More specifically, we get

$$V_1 = \frac{1}{(n-2)^2 m^2}\left\{4\binom{n-2}{2}\binom{m}{2}\hat{U}_{A_1}^{(4,2)}\right\} + O_P\Big(\frac{1}{n}+\frac{1}{m}\Big) \text{ and}$$

$$\mathrm{Var}(V_1) = \frac{\sigma_{1,0}^2(A_1)}{n} + \frac{\sigma_{0,1}^2(A_1)}{m} + C_2\Big(\frac{1}{n}+\frac{1}{m}\Big)^2.$$

Note that here $\sigma_{1,0}^2(A_1) = \mathrm{Var}(\psi^s_{A_1,1,0}(\mathbf{X}))$ and $\sigma_{0,1}^2(A_1) = \mathrm{Var}(\psi^s_{A_1,0,1}(\mathbf{X}))$, where we have $\psi^s_{A_1,1,0}(\mathbf{x}) = \mathbb{E}[\psi_{A_1}(\mathbf{x}, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4; \mathbf{Y}_1, \mathbf{Y}_2)]$ and $\psi^s_{A_1,0,1}(\mathbf{y}) = \mathbb{E}[\psi_{A_1}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4; \mathbf{y}, \mathbf{Y}_2)]$. Since ψ_{A_1} is uniformly bounded, the constant C_2 does not depend on d. Now, we have

$$\begin{aligned}
\psi^s_{A_1,1,0}(\mathbf{x}) &= \mathbb{E}\Big\{\frac{1}{4!2!}\sum_{\pi\in S_4}\sum_{\gamma\in S_2}\psi_{A_1}(\mathbf{X}_{\pi(1)}, \mathbf{X}_{\pi(2)}, \mathbf{X}_{\pi(3)}, \mathbf{X}_{\pi(4)}; \mathbf{Y}_{\gamma(1)}, \mathbf{Y}_{\gamma(2)}) \mid \mathbf{X}_1 = \mathbf{x}\Big\} \\
&= \frac{1}{4}\mathbb{E}\Big\{\psi_{A_1}(\mathbf{x}, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4; \mathbf{Y}_1, \mathbf{Y}_2) + \psi_{A_1}(\mathbf{X}_1, \mathbf{x}, \mathbf{X}_3, \mathbf{X}_4; \mathbf{Y}_1, \mathbf{Y}_2) \\
&\qquad + \psi_{A_1}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{x}, \mathbf{X}_4; \mathbf{Y}_1, \mathbf{Y}_2) + \psi_{A_1}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{x}; \mathbf{Y}_1, \mathbf{Y}_2)\Big\}
\end{aligned}$$

So, after simplification, one gets

$$\begin{aligned}
\psi^s_{A_1,1,0}(\mathbf{x}) &= \frac{1}{4}\Big\{\mathbb{E}\big(F(B(\mathbf{x}, \rho(\mathbf{X}_2, \mathbf{x}))) - G(B(\mathbf{x}, \rho(\mathbf{X}_2, \mathbf{x})))\big)^2\Big\} \\
&\quad + \frac{1}{4}\Big\{\mathbb{E}\big(F(B(\mathbf{X}_1, \rho(\mathbf{x}, \mathbf{X}_1))) - G(B(\mathbf{X}_1, \rho(\mathbf{x}, \mathbf{X}_1)))\big)^2\Big\} \\
&\quad + \frac{1}{2}\mathbb{E}\Big(\delta(\mathbf{x}, \mathbf{X}_2, \mathbf{X}_1) - G(B(\mathbf{X}_1, \rho(\mathbf{X}_2, \mathbf{X}_1)))\Big) \\
&\qquad\quad \times \Big(F(B(\mathbf{X}_1, \rho(\mathbf{X}_2, \mathbf{X}_1))) - G(B(\mathbf{X}_1, \rho(\mathbf{X}_2, \mathbf{X}_1)))\Big) \\
&= g_1(\mathbf{x}) + g_2(\mathbf{x}) + g_3(\mathbf{x}), \quad \text{say.}
\end{aligned}$$

Therefore, we have $\sigma_{1,0}^2 = \mathbb{E}\big\{\psi^s_{A_1,1,0}(\mathbf{X}) - \mathbb{E}(\psi^s_{A_1,1,0}(\mathbf{X}))\big\}^2 = \mathbb{E}\big\{\psi^s_{A_1,1,0}(\mathbf{X}) - A_1\big\}^2 = \mathbb{E}\big\{g_1(\mathbf{X}) + g_2(\mathbf{X}) + g_3(\mathbf{X}) - A_1\big\}^2$. So, using the inequality, $\mathbb{E}(\sum_{i=1}^p Z_1)^2 \le p\ \mathbb{E}(\sum_{i=1}^p Z_i^2)$ and the fact that $0 \le g_1(\mathbf{x}), g_2(\mathbf{x}) \le 1/4$ for all $\mathbf{x}$, we get

$$\sigma_{1,0}^2 \le 4\Big\{\mathbb{E}g_1^2(\mathbf{X}) + \mathbb{E}g_2^2(\mathbf{X}) + \mathbb{E}g_3^2(\mathbf{X}) + A_1^2\Big\} \le 4\Big\{\frac{1}{4}\mathbb{E}g_1(\mathbf{X}) + \frac{1}{4}\mathbb{E}g_2(\mathbf{X}) + \mathbb{E}g_3^2(\mathbf{X}) + A_1\Big\}.$$

Now note that $\mathbb{E}g_1(\mathbf{X}) = \mathbb{E}g_2(\mathbf{X}) = \frac{1}{4}A_1$. Also, using Cauchy-Schwartz inequality on $g_3(\mathbf{X})$, we get

$$\mathbb{E}g_3^2(\mathbf{X}) \le \frac{1}{4}\mathbb{E}\Big(\delta(\mathbf{X}, \mathbf{X}_2, \mathbf{X}_1) - G(B(\mathbf{X}_1, \rho(\mathbf{X}_2, \mathbf{X}_1)))\Big)^2 A_1 \le \frac{1}{4}A_1.$$

Hence, we have $\sigma_{1,0}^2(A_1) \le \frac{11}{2}A_1$. Similarly, we have $\sigma_{0,1}^2(A_1) \le \frac{9}{2}A_1$. Combining these, we get

$$\mathrm{Var}(V_1) \le \frac{11}{2}A_1\Big(\frac{1}{n}+\frac{1}{m}\Big) + C_2\Big(\frac{1}{n}+\frac{1}{m}\Big)^2.$$

Using the same set of arguments, for V_2 (as defined in the proof of Lemma A2.1), we also have

$$\mathrm{Var}(V_2) \le \frac{11}{2}A_2\Big(\frac{1}{n}+\frac{1}{m}\Big) + C_2\Big(\frac{1}{n}+\frac{1}{m}\Big)^2.$$

Now, the proof follows from the fact that $\mathrm{Var}(T_{n,m}^{\rho}) = \mathrm{Var}(V_1 + V_2) \le 2(\mathrm{Var}(V_1) + \mathrm{Var}(V_2))$. ■

Lemma A2.3. *Consider a random permutation π of $\{1,2,\ldots,n+m\}$. If $T^{\rho}_{n,m,\pi}$ denotes the permuted test statistic (the permutation analog of $T^{\rho}_{n,m}$), given the pooled sample $\mathcal{U} = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{n+m}\}$, the conditional expectation $T^{\rho}_{n,m,\pi}$ is given by*

$$\mathbb{E}\{T^{\rho}_{n,m,\pi} \mid \mathcal{U}\} = \frac{1}{6}\left(\frac{1}{n} + \frac{1}{m} + \frac{1}{n-2} + \frac{1}{m-2}\right).$$

Proof. For any random permutation π of $\{1,2,\ldots,n+m\}$, we have

$$\begin{aligned}
T^{\rho}_{n,m,\pi} =& \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \left\{ \frac{1}{n-2} \sum_{k=1,k\ne i,j}^{n} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(j)}, \mathbf{U}_{\pi(i)}) - \frac{1}{m} \sum_{k=n+1}^{n+m} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(j)}, \mathbf{U}_{\pi(i)}) \right\}^2 \\
&+ \frac{1}{m(m-1)} \sum_{n+1\le i\ne j\le n+m} \left\{ \frac{1}{n} \sum_{k=1}^{n} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(j)}, \mathbf{U}_{\pi(i)}) \right. \\
&\qquad \left. - \frac{1}{m-2} \sum_{k=n+1,k\ne i,j}^{n+m} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(j)}, \mathbf{U}_{\pi(i)}) \right\}^2,
\end{aligned}$$

So, the conditional expectation of $T^{\rho}_{n,m,\pi}$ for any given $\mathcal{U}$ is given by

$$\begin{aligned}
\mathbb{E}\{T^{\rho}_{n,m,\pi} \mid \mathcal{U}\} =& \mathbb{E}\left\{ \left\{ \frac{1}{n-2} \sum_{k=3}^{n} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) - \frac{1}{m} \sum_{k=n+1}^{n+m} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \right\}^2 \Big| \ \mathcal{U} \right\} \\
&+ \mathbb{E}\left\{ \left\{ \frac{1}{n} \sum_{k=1}^{n} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) - \frac{1}{m-2} \sum_{k=n+1,k\ne i,j}^{n+m} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \right\}^2 \Big| \ \mathcal{U} \right\}.
\end{aligned}$$

Now note that

$$\begin{aligned}
&\mathbb{E}\left\{ \left\{ \frac{1}{n-2} \sum_{k=3}^{n} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) - \frac{1}{m} \sum_{k=n+1}^{n+m} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \right\}^2 \Big| \ \mathcal{U} \right\} \\
&= \frac{1}{(n-2)^2} \mathbb{E}\Big\{ \sum_{k=3}^{n} \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \mid \mathcal{U} \Big\} \\
&\quad + \frac{1}{(n-2)^2} \sum_{k,l=3,k\ne l}^{n} \mathbb{E}\Big\{ \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \delta(\mathbf{U}_{\pi(l)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \mid \mathcal{U} \Big\} \\
&\quad + \frac{1}{m^2} \sum_{k=n+1}^{n+m} \mathbb{E}\Big\{ \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \mid \mathcal{U} \Big\} \\
&\quad + \frac{1}{m^2} \sum_{k,l=n+1,k\ne l}^{n+m} \mathbb{E}\Big\{ \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \delta(\mathbf{U}_{\pi(l)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \mid \mathcal{U} \Big\} \\
&\quad - \frac{2}{m(n-2)} \sum_{k=3}^{n} \sum_{l=n+1}^{n+m} \mathbb{E}\Big\{ \delta(\mathbf{U}_{\pi(k)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \delta(\mathbf{U}_{\pi(l)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(1)}) \mid \mathcal{U} \Big\} \\
&= \frac{(n-2)q_1}{(n-2)^2} + \frac{(n-2)(n-3)q_2}{(n-2)^2} + \frac{mq_1}{m^2} + \frac{m(m-1)q_2}{m^2} - \frac{2m(n-2)q_2}{m(n-2)} = (q_1 - q_2)\Big(\frac{1}{n-2} + \frac{1}{m}\Big),
\end{aligned}$$

where $q_1 = \mathbb{E}\{\delta(\mathbf{U}_{\pi(1)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(3)}) | \mathcal{U}\}$ and $q_2 = \mathbb{E}\{\delta(\mathbf{U}_{\pi(1)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(3)}) \delta(\mathbf{U}_{\pi(4)}, \mathbf{U}_{\pi(2)}, \mathbf{U}_{\pi(3)}) | \mathcal{U}\}$.

Similarly, we can also show that

$$\mathbb{E}\Big\{\Big\{\frac{1}{n}\sum_{k=1}^{n}\delta(\mathbf{U}_{\pi(k)},\mathbf{U}_{\pi(2)},\mathbf{U}_{\pi(1)})-\frac{1}{m-2}\sum_{k=n+1,k\neq i,j}^{n+m}\delta(\mathbf{U}_{\pi(k)},\mathbf{U}_{\pi(2)},\mathbf{U}_{\pi(1)})\Big\}^2\Big|\ \mathcal{U}\Big\}$$
$$=(q_1-q_2)\Big(\frac{1}{m-2}+\frac{1}{n}\Big).$$

But given the pooled sample $\mathcal{U}$, the random variables $\{\mathbf{U}_{\pi(i)}\}_{i=1}^{n+m}$ are exchangeable. So, we must have $q_1 = 1/2$ and $q_2 = 1/3$. Hence, we have

$$\mathbb{E}\{T^{\rho}_{n,m,\pi}\mid\mathcal{U}\}=\frac{1}{6}\left(\frac{1}{n}+\frac{1}{m}+\frac{1}{n-2}+\frac{1}{m-2}\right). \qquad \blacksquare$$

Proof of Lemma 2.2. Here we are interested in the quantiles of the conditional distribution of $T^{\rho}_{n,m,\pi}$ given the pooled sample $\mathcal{U}$. Since $T^{\rho}_{n,m,\pi}$ is non-negative, using Markov's inequality on the conditional random variable, we get

$$\mathbb{P}\Big\{T^{\rho}_{n,m,\pi}\geq\frac{1}{\alpha}\mathbb{E}\{T^{\rho}_{n,m,\pi}\mid\mathcal{U}\}\ \Big|\ \mathcal{U}\Big\}\leq\alpha.$$

Therefore, from the definition of the quantile $c_{1-\alpha}$, we have $c_{1-\alpha}\leq\frac{1}{\alpha}\mathbb{E}\{T^{\rho}_{n,m,\pi}\mid\mathcal{U}\}$, which holds with probability one. From Lemma A.2.3, we also have

$$\mathbb{E}\{T^{\rho}_{n,m,\pi}\mid\mathcal{U}\}=\frac{1}{6}\left(\frac{1}{n}+\frac{1}{m}+\frac{1}{n-2}+\frac{1}{m-2}\right)\leq\frac{2}{3(\min\{n,m\}-2)}.$$

This completes the proof. $\blacksquare$

Proof of Theorem 2.1. In view of Lemma A.2.1 and Lemma A.2.2, as $\min\{n,m\}$ grows to infinity, $T^{\rho}_{n,m}$ converges in probability to $\Theta^2_{\rho}(F,G)$. So, if $\Theta^2_{\rho}(F,G)>0$, under $H_1: F\neq G$, $T^{\rho}_{n,m}$ converges in probability to a positive number. On the other hand, Lemma 2.1 shows that the cut-off value of the permutation test $c_{1-\alpha}$ goes to zero almost surely. Therefore, the power of the permutation test converges to one as $\min\{n,m\}$ grows to infinity. $\blacksquare$

Proof of Lemma 2.3. To prove this lemma, we shall use the idea of Corollary 6.1 of Kim (2021). First, let us define

$$M(t)=\frac{1}{N!}\Big\{\sum_{\pi\in\mathcal{S}_N}\mathrm{I}[T^{\rho}_{n,m,\pi}\leq t]\Big\}\qquad\text{and}\qquad M_B(t)=\frac{1}{B}\Big\{\sum_{i=1}^{B}\mathrm{I}[T^{\rho}_{n,m,\pi_i}\leq t]\Big\}.$$

where M and M_B are distribution functions conditioned on the observed pooled data $\mathcal{U}$. Now,

$$\begin{aligned}
|p_{n,m}-p_{n,m,B}|&=\left|\frac{1}{N!}\Big\{\sum_{\pi\in\mathcal{S}_N}\mathrm{I}[T^{\rho}_{n,m,\pi}\geq T^{\rho}_{n,m}]\Big\}-\frac{1}{B+1}\Big\{\sum_{i=1}^{B}\mathrm{I}[T^{\rho}_{n,m,\pi_i}\geq T^{\rho}_{n,m}]+1\Big\}\right|\\
&=\left|\frac{1}{N!}\Big\{\sum_{\pi\in\mathcal{S}_N}\mathrm{I}[T^{\rho}_{n,m,\pi}< T^{\rho}_{n,m}]\Big\}-\frac{1}{B+1}\Big\{\sum_{i=1}^{B}\mathrm{I}[T^{\rho}_{n,m,\pi_i}< T^{\rho}_{n,m}]\Big\}\right|\\
&=\Big|M(T^{\rho}_{n,m}-)-\frac{B}{B+1}M_B(T^{\rho}_{n,m}-)\Big|\\
&\leq\Big|M(T^{\rho}_{n,m}-)-M_B(T^{\rho}_{n,m}-)\Big|+\Big|\frac{M_B(T^{\rho}_{n,m}-)}{B+1}\Big|\leq\sup_{t\in\mathbb{R}}|M(t)-M_B(t)|+\frac{1}{B+1}.
\end{aligned}$$

Conditioned on the pooled data $\mathcal{U}$, the Dvoretzky-Keifer-Wolfwitz inequality (see, e.g., Massart, 1990) gives us $\mathbb{P}\{\sup_{t\in\mathbb{R}}|M(t)-M_B(t)|>\epsilon\}\le 2e^{-2B\epsilon^2}$. Hence, conditioned on $\mathcal{U}$, as B grows to infinity, the randomized p-value $p_{n,m,B}$ converges almost surely to $p_{n,m}$. ■

Proof of Lemma 2.4. For $\mathbf{W}=\mathbf{X}_1-\mathbf{X}_2, \mathbf{Y}_1-\mathbf{Y}_2$ or $\mathbf{X}_1-\mathbf{Y}_1$, under (A2.1), we have $\frac{1}{d}\Big|\|W\|^2-\mathbb{E}(\|W\|^2)\Big|\xrightarrow{P}0$ as $d\to\infty$. Again under (A2.2), as $d\to\infty$, $\frac{1}{d}E(\|W\|^2)$ converges to $2\sigma_F^2$, $2\sigma_F^2$ and $\sigma_F^2+\sigma_G^2+\nu^2$ in these three cases, respectively. The result follows from these two facts. ■

Lemma A2.4. *Suppose that $\mathbf{X}_1,\mathbf{X}_2\sim F$, $\mathbf{Y}_1,\mathbf{Y}_2\sim G$ and they are independent. For a distance function ρ, assume that $\rho(\mathbf{X}_1,\mathbf{X}_2)\xrightarrow{P}\theta_1$, $\rho(\mathbf{Y}_1,\mathbf{Y}_2)\xrightarrow{P}\theta_2$ and $\rho(\mathbf{X}_1,\mathbf{Y}_2)\xrightarrow{P}\theta_3$ as $d\to\infty$. If $\theta_3>\min\{\theta_1,\theta_2\}$, then $P(T^{\rho}_{n,m}>1/3)\to 1$ as d diverges to infinity.*

Proof. Note that $T^{\rho}_{n,m}$ involves the terms $\delta(\mathbf{U}_k,\mathbf{U}_j,\mathbf{U}_i)$'s for different choices of $\mathbf{U}_i,\mathbf{U}_j,\mathbf{U}_k$ from the pooled sample. So, the behaviour of $T^{\rho}_{n,m}$ can be studied using the convergence of the $\delta(\mathbf{U}_k,\mathbf{U}_j,\mathbf{U}_i)$'s.

First, consider the case $\min\{\theta_1,\theta_2\}<\theta_3<\max\{\theta_1,\theta_2\}$. Let us assume that $\theta_1>\theta_3>\theta_2$. In such a situation, we have $\lim_{d\to\infty}\mathbb{P}[\rho(\mathbf{X}_2,\mathbf{Y}_1)\le\rho(\mathbf{Y}_2,\mathbf{Y}_1)]=0$ and $\lim_{d\to\infty}\mathbb{P}[\rho(\mathbf{Y}_1,\mathbf{X}_1)\le\rho(\mathbf{X}_2,\mathbf{X}_1)]=1$. Consider V_1 and V_2 as defined in the proof of Lemma A2.1. Now we have

$$\begin{aligned}
V_1 &= \frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}\left\{\frac{1}{n-2}\sum_{k=1,k\ne i,j}^{n}\delta(\mathbf{X}_k,\mathbf{X}_j,\mathbf{X}_i)-\frac{1}{m}\sum_{k=1}^{m}\delta(\mathbf{Y}_k,\mathbf{X}_j,\mathbf{X}_i)\right\}^2\\
&\xrightarrow{P}\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}\left\{\frac{1}{n-2}\sum_{k=1,k\ne i,j}^{n}\delta(\mathbf{X}_k,\mathbf{X}_j,\mathbf{X}_i)-1\right\}^2\\
&=\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}\left\{\frac{-1}{n-2}\sum_{k=1,k\ne i,j}^{n}\mathrm{I}[\rho(\mathbf{X}_k,\mathbf{X}_i)>\rho(\mathbf{X}_j,\mathbf{X}_i)]\right\}^2\\
&=\frac{1}{n(n-1)(n-2)^2}\Bigg\{\sum_{1\le i\ne j\ne k\le n}\mathrm{I}[\rho(\mathbf{X}_k,\mathbf{X}_i)>\rho(\mathbf{X}_j,\mathbf{X}_i)]\\
&\quad+\sum_{1\le i\ne j\le n}\sum_{k=1,k\ne i,j}^{n}\sum_{l=1,l\ne i,j,k}^{n}\mathrm{I}[\rho(\mathbf{X}_k,\mathbf{X}_i)>\rho(\mathbf{X}_j,\mathbf{X}_i)]\mathrm{I}[\rho(\mathbf{X}_l,\mathbf{X}_i)>\rho(\mathbf{X}_j,\mathbf{X}_i)]\Bigg\}\\
&=\frac{1}{n(n-1)(n-2)^2}\left\{\binom{n}{1}\binom{n-1}{2}+2\binom{n}{1}\binom{n-1}{3}\right\}=\frac{1}{3}+\frac{1}{6(n-2)}.
\end{aligned}$$

Similarly, as d diverges to infinity, we have

$$V_2\xrightarrow{P}\frac{1}{m(m-1)}\sum_{n+1\le i\ne j\le n+m}\left\{\frac{1}{m-2}\sum_{k=1,k\ne i,j}^{m}\delta(\mathbf{Y}_k,\mathbf{Y}_j,\mathbf{Y}_i)\right\}^2=\frac{1}{3}+\frac{1}{6(m-2)}.$$

Thus, $T^{\rho}_{n,m}\xrightarrow{P}\frac{2}{3}+\frac{1}{6}\left(\frac{1}{n-2}+\frac{1}{m-2}\right)$. The same result holds for $\theta_1<\theta_3<\theta_2$ as well.

Now consider the case, $\theta_3=\max\{\theta_1,\theta_2\}$. Assume that $\theta_2<\theta_1=\theta_3$. In this case, the convergence of $\mathbb{P}[\rho(\mathbf{Y}_1,\mathbf{X}_1)\le\rho(\mathbf{X}_2,\mathbf{X}_1)]$ is not clear, but $\mathbb{P}[\rho(\mathbf{X}_1,\mathbf{Y}_1)\le\rho(\mathbf{Y}_2,\mathbf{Y}_1)]$ converges to

zero as d diverges to infinity. Hence, V_2 converges to $1/3 + 1/\{6(m-2)\}$ in probability, and V_1 converges in probability to a non-negative random variable. Therefore, $P(T^{\rho}_{n,m} > 1/3)$ converges to one. Similar arguments can be given for $\theta_1 < \theta_2 = \theta_3$ as well.

Finally, consider the case $\theta_3 > \max\{\theta_1, \theta_2\}$. In this case, we have $\lim_{d\to\infty} \mathbb{P}[\rho(\mathbf{X}_2, \mathbf{Y}_1) \leq \rho(\mathbf{Y}_2, \mathbf{Y}_1)] = 0$ and $\lim_{d\to\infty} \mathbb{P}[\rho(\mathbf{Y}_1, \mathbf{X}_1) \leq \rho(\mathbf{X}_2, \mathbf{X}_1)] = 0$. Hence as d diverges to infinity, V_1 and V_2 converge in probability to $1/3 + 1/\{6(n-2)\}$ and $1/3 + 1/\{6(m-2)\}$, respectively. Thus, $T^{\rho}_{n,m}$ converges in probability to $2/3 + (1/6)\,(1/(n-2) + 1/(m-2))$.
These three cases together imply that if $\theta_3 > \min\{\theta_1, \theta_2\}$, $P(T^{\rho}_{n,m} > 1/3) \to 1$ as $d \to \infty$. ■

Proof of Lemma 2.5. Here, we have $d^{-1/2}\|\mathbf{X}_1 - \mathbf{X}_2\| \xrightarrow{P} \sigma_F\sqrt{2}$, $d^{-1/2}\|\mathbf{Y}_1 - \mathbf{Y}_2\| \xrightarrow{P} \sigma_G\sqrt{2}$ and $d^{-1/2}\|\mathbf{X}_1 - \mathbf{Y}_1\| \xrightarrow{P} \sqrt{\sigma_F^2 + \sigma_G^2 + \nu^2}$ as $d \to \infty$ (see Lemma 2.4). Let these limiting values be denoted by θ_1, θ_2, and θ_3, respectively. If $\nu^2 + (\sigma_F - \sigma_G)^2 > 0$, one can check that $\theta_3 > \sqrt{\theta_1\theta_2} \geq \min\{\theta_1, \theta_2\}$. Hence, the proof follows from Lemma A2.4. ■

Proof of Theorem 2.2. It follows from Lemma A2.4 that under the condition $\nu^2 + (\sigma_F - \sigma_G)^2 > 0$, $\mathbb{P}(T^{\ell_2}_{n,m} > 1/3)$ converges to 1 as d tends to infinity. We have also seen that the cut-off of the permutation test $c_{1-\alpha}$ has an upper bound $2/\{3\alpha(\min\{n,m\} - 2)\}$, which does not depend on the dimension d. Therefore, if $\min\{n, m\} \geq 2 + 2/\alpha$, the test based on $T^{\ell_2}_{n,m}$ rejects H_0 with probability tending to 1 as d grows to infinity. ■

Proof of Lemma 2.6 . This lemma is taken from Sarkar *&* Ghosh (2018a). The proof can be found on page 5 (see Lemma 1) of that article. ■

Proof of Theorem 2.3 . We use a sub-sequence argument to prove this theorem. Let $\{d_k\}$ be an arbitrary sub-sequence of the sequence of natural numbers. Under (A2.3) and $\liminf_{d\to\infty} e_{h,\psi}(F, G) > 0$, there exists a further subsequence $\{d'_k\}$ such that $\lim_{d'_k\to\infty} e_{h,\psi}(F, G) > 0$, and the corresponding limits of the three terms in $e_{h,\psi}(F, G)$ exist. Let θ_1, θ_2 and θ_3 be the limiting values of $d^{-1}\sum_{q=1}^{d}\{\psi(|X_1^{(q)} - X_2^{(q)}|^2)$, $d^{-1}\sum_{q=1}^{d}\{\psi(|Y_1^{(q)} - Y_2^{(q)}|^2)$ and $d^{-1}\sum_{q=1}^{d}\{\psi(|X_1^{(q)} - Y_1^{(q)}|^2)$, respectively, along the sub-sequence $\{d'_k\}$. Since $\lim_{d'_k\to\infty} e_{h,\psi}(F, G) > 0$, we have $2\theta_3 > \theta_1 + \theta_2$. Hence, using Lemma A2.4, we get $\mathbb{P}(T^{h,\psi}_{n,m} > 1/3) \to 1$ as $d'_k \to \infty$. Since $\{d'_k\}$ is the sub-sequence of an arbitrary sequence $\{d_k\}$, we can conclude that $\mathbb{P}(T^{h,\psi}_{n,m} > 1/3) \to 1$ as $d \to \infty$. Now, using arguments similar to those in the proof of Theorem 2.2, one can establish the consistency of the level α test when $\min\{n, m\} \geq 2 + 2/\alpha$. ■

Lemma A2.5. *If* $\mathbf{X}_1, \mathbf{X}_2 \overset{iid}{\sim} F$ *and* $\mathbf{Y}_1, \mathbf{Y}_2 \overset{iid}{\sim} G$ *are independent random vectors, then*

$$\Theta^2_{\ell_2}(F, G) \geq \left[\mathbb{P}\{\|\mathbf{X}_1 - \mathbf{Y}_1\| \leq \|\mathbf{Y}_2 - \mathbf{Y}_1\|\} - 1/2\right]^2 + \left[\mathbb{P}\{\|\mathbf{Y}_1 - \mathbf{X}_1\| \leq \|\mathbf{X}_2 - \mathbf{X}_1\|\} - 1/2\right]^2.$$

Proof. We know that for any random variable Z, $(\mathbb{E}\{Z\})^2 \le \mathbb{E}\{Z^2\}$. Using this fact twice, we get

$$\begin{aligned}\Theta^2_{\ell_2}(F,G) &= \int \{F(B(\boldsymbol{u},\ell_2(\boldsymbol{v},\boldsymbol{u}))) - G(B(\boldsymbol{u},\ell_2(\boldsymbol{v},\boldsymbol{u})))\}^2 \, [dF(\boldsymbol{u})dF(\boldsymbol{v}) + dG(\boldsymbol{u})dG(\boldsymbol{v})] \\ &\ge \left\{\int F(B(\boldsymbol{u},\ell_2(\boldsymbol{v},\boldsymbol{u}))) - G(B(\boldsymbol{u},\ell_2(\boldsymbol{v},\boldsymbol{u}))) \, dF(\boldsymbol{u})dF(\boldsymbol{v})\right\}^2 \\ &\qquad + \left\{\int F(B(\boldsymbol{u},\ell_2(\boldsymbol{v},\boldsymbol{u}))) - G(B(\boldsymbol{u},\ell_2(\boldsymbol{v},\boldsymbol{u}))) \, dG(\boldsymbol{u})dG(\boldsymbol{v})\right\}^2 \\ &= \{\mathbb{P}\{\|\mathbf{Y}_1 - \mathbf{X}_1\| \le \|\mathbf{X}_2 - \mathbf{X}_1\|\} - 1/2\}^2 + \{\mathbb{P}\{\|\mathbf{X}_1 - \mathbf{Y}_1\| \le \|\mathbf{Y}_2 - \mathbf{Y}_1\|\} - 1/2\}^2\end{aligned}$$

The last equality follows from the fact that if $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \overset{iid}{\sim} F_0$, where F_0 is a continuous distribution, we have $\int F_0\big(B(u,\ell_2(v,u))\big)\, dF_0(u)dF_0(v) = \mathbb{P}\{\|\mathbf{X}_3 - \mathbf{X}_1\| \le \|\mathbf{X}_2 - \mathbf{X}_1\|\} = \frac{1}{2}$. ■

Lemma A2.6. *Consider two d-dimensional random variables $\mathbf{X} = (\xi_1, 0, \ldots, 0)^\top$ and $\mathbf{Y} = (\xi_2, 0, \ldots, 0)^\top$, where $\xi_1 \sim \mathcal{N}_1(\mu_1, 1)$ and $\xi_2 \sim \mathcal{N}_1(\mu_2, 1)$ are independent. Let P_1 and P_2 denote the distributions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. If $\mu_1 = cn^{-1/2}$ and $\mu_2 = -cm^{-1/2}$ for some $c > 0$, then there exists a constant $C > 0$ independent of the dimension d such that $\Theta^2(P_1, P_2) \ge C\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\big)^2$. This lower bound is tight up to a constant factor.*

Proof. Let $\xi_{11}, \xi_{12} \overset{iid}{\sim} \mathcal{N}_1(\mu_1, 1)$ and $\xi_{21}, \xi_{22} \overset{iid}{\sim} \mathcal{N}_1(\mu_2, 1)$ be independent random variables. In view of Lemma A2.5, for $\mathbf{X}_1 = (\xi_{11}, 0, \ldots, 0)^\top, \mathbf{X}_2 = (\xi_{12}, 0, \ldots, 0)^\top \sim P_1$ and $\mathbf{Y}_1 = (\xi_{21}, 0, \ldots, 0)^\top$, $\mathbf{Y}_2 = (\xi_{22}, 0, \ldots, 0)^\top \sim P_2$, it is enough to prove that

$$[\mathbb{P}\{\|\mathbf{X}_1 - \mathbf{Y}_1\| \le \|\mathbf{Y}_2 - \mathbf{Y}_1\|\} - 1/2]^2 + [\mathbb{P}\{\|\mathbf{Y}_1 - \mathbf{X}_1\| \le \|\mathbf{X}_2 - \mathbf{X}_1\|\} - 1/2]^2 \ge C\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)^2,$$

Now, we derive the lower bounds for these two terms separately. Note that

$$\begin{aligned}\Big|\mathbb{P}\{\|\mathbf{X}_1 - \mathbf{Y}_1\| \le \|\mathbf{Y}_2 - \mathbf{Y}_1\|\} - 1/2\Big| &= \Big|\mathbb{P}\{|\xi_{11} - \xi_{21}| \le |\xi_{22} - \xi_{22}|\} - 1/2\Big| \\ &= \Big|\mathbb{P}\{|\xi_{11} - \xi_{21}|^2 - |\xi_{22} - \xi_{21}|^2 \le 0\} - 1/2\Big| \\ &= \Big|\mathbb{P}\{(\xi_{11} + \xi_{22} - 2\xi_{21})(\xi_{11} - \xi_{22}) \le 0\} - 1/2\Big|.\end{aligned}$$

So, taking $T_0 = \xi_{11} - \xi_{22}$ and $S_0 = \xi_{11} + \xi_{22} - 2\xi_{21}$, we get

$$\Big|\mathbb{P}\{\|\mathbf{X}_1 - \mathbf{Y}_1\| \le \|\mathbf{Y}_2 - \mathbf{Y}_1\|\} - 1/2\Big| = \Big|\mathbb{P}\{S_0T_0 \le 0\} - 1/2\Big| = \Big|\mathbb{E}\{\mathbb{P}\{S_0T_0 \le 0 \mid S_0\} - 1/2\}\Big|.$$

Here, T_0 and S_0 jointly follow a bivariate normal distribution with $\mathbb{E}(T_0) = c(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}})$, $\mathbb{E}(S_0) = c(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}})$, $\mathrm{Var}(T_0) = 2$, $\mathrm{Var}(S_0) = 6$ and $\mathrm{Cov}(T_0, S_0) = 0$. Therefore,

$$\begin{aligned}\Big|\mathbb{E}\big\{\mathbb{P}\{S_0T_0 \le 0 \mid S_0\} - 1/2\big\}\Big| &= \left|\mathbb{E}\left\{\Phi\left(-c\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)\frac{S_0}{|S_0|\sqrt{2}}\right) - 1/2\right\}\right| \\ &\ge \frac{c}{\sqrt{2}}\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)\phi(\sqrt{2}c),\end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density function and the distribution function of the standard normal variate, respectively. Here, the last inequality is obtained by using the mean value theorem

and the fact that $\phi(t)$ is decreasing in $|t|$. Therefore, we get

$$\Big|\mathbb{P}\{\|\mathbf{X}_1 - \mathbf{Y}_1\| \le \|\mathbf{Y}_2 - \mathbf{Y}_1\|\} - 1/2\Big| \ge \frac{c}{\sqrt{2}}\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)\phi(\sqrt{2}c).$$

Similarly, we can derive the same lower bound for $\Big|\mathbb{P}\{\|\mathbf{Y}_1 - \mathbf{X}_1\| \le \|\mathbf{X}_2 - \mathbf{X}_1\|\} - 1/2\Big|$ as well. So, we can find a constant $C > 0$ independent of the dimension d such that

$$\Theta^2_{\ell_2}(P_1, P_2) \ge C\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)^2.$$

To show that this lower bound is tight, notice that

$$\Theta^2(P_1, P_2) \le \sup_A |P_1(A) - P_2(A)| \le KL(P_1, P_2) = \frac{c^2}{2}(\mu_1 - \mu_2)^2 = \frac{c^2}{2}\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)^2,$$

where $KL(\cdot, \cdot)$ denotes the Kullback-Leibler divergence between two probability measures. Here, the first inequality follows trivially, and the second one is known as Pinsker's inequality (see Lemma 2.5 in Tsybakov, 2009). Hence the lower bound is tight up to a constant term. ■

Proof of Theorem 2.4. The minimax lower bound can be obtained based on the standard application of Neyman-Pearson lemma (see Baraud, 2002; Kim, Balakrishnan *&* Wasserman, 2020). Let the distributions of the sample under the null and alternative hypotheses be denoted as Q_0 and Q_1, respectively. Then, following our notations, we have

$$R_{n,m,d}(\epsilon) \ge 1 - \alpha - \sup_A |Q_0(A) - Q_1(A)| \ge 1 - \alpha - \sqrt{\frac{1}{2}KL(Q_0, Q_1)},$$

where the second inequality is obtained from Pinsker's inequality (see Tsybakov, 2009). Now suppose that P_1 and P_2 are the distributions corresponding to $\mathbf{X} = (\xi_1, 0, \ldots, 0)^\top$ and $\mathbf{Y} = (\xi_2, 0, \ldots, 0)^\top$, where ξ_1 and ξ_2 are independent random variables following normal distributions with the unit variance and means

$$\mu_1 = \frac{\sqrt{2}(1 - \alpha - \zeta)}{\sqrt{n}} \qquad \text{and} \qquad \mu_2 = -\frac{\sqrt{2}(1 - \alpha - \zeta)}{\sqrt{m}},$$

respectively. Let P_0 be the distribution of $(\xi, 0, \ldots, 0)^\top$ where ξ is a standard normal random variable. Define $k(\alpha, \zeta) := (1 - \alpha - \zeta)^2\Big(\phi\big(\sqrt{2}(1 - \alpha - \zeta)\big)\Big)^2$. Then by Lemma A2.5, $(P_1, P_2) \in \mathcal{F}(c\lambda(n, m))$ for all $0 < c < k(\alpha, \zeta)$. Now taking $Q_0 = P_0^{(n+m)}$ and $Q_1 = P_1^n P_2^m$, we have

$$KL(Q_0, Q_1) = \frac{n}{2}\mu_1^2 + \frac{m}{2}\mu_2^2 = 2(1 - \alpha - \zeta)^2.$$

Therefore, $R_{n,m,d}(c\lambda(n, m)) \ge \zeta$ for all $0 < c < k(\alpha, \zeta)$. Since ζ and $k(\alpha, \zeta)$ do not depend on n, m and d, this trivially satisfies the condition $\liminf_{n,m,d\to\infty} R_{n,m,d}(c\lambda(n, m)) \ge \zeta$ for all $0 < c < k(\alpha, \zeta)$. ■

Proof of Theorem 2.5 . Here we want to show that for every positive α and ζ, there exists a constant $K(\alpha, \zeta)$ such that

$$\limsup_{n,m,d\to\infty} \sup_{(F,G)\in\mathcal{F}(c\lambda(n,m))} \mathbb{P}^{n,m}_{F,G}\{T_{n,m} \le c_{1-\alpha}\} \le \zeta$$

for all $c > K(\alpha, \zeta)$. Let us first choose a constant K_1 such that

$$K_1\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)^2 \geq \frac{1}{\alpha}\mathbb{E}\{T^{\pi}_{n,m} \mid \mathcal{U}\} = \frac{1}{6\alpha}\left(\frac{1}{n} + \frac{1}{m} + \frac{1}{n-2} + \frac{1}{m-2}\right).$$

Now, take any $(F, G) \in \mathcal{F}(c\lambda(n, m))$ such that $c > K_1$. Using the fact that $c_{1-\alpha} \leq \frac{1}{\alpha}\mathbb{E}\{T^{\pi}_{n,m} \mid \mathcal{U}\}$ (see the proof of Lemma 2.2), we get

$$\begin{aligned}\mathbb{P}^{n,m}_{F,G}\{T_{n,m} \leq & c_{1-\alpha}\} \leq \mathbb{P}^{n,m}_{F,G}\{T_{n,m} \leq \frac{1}{\alpha}\mathbb{E}\{T^{\pi}_{n,m} \mid \mathcal{U}\}\} \\ &= \mathbb{P}^{n,m}_{F,G}\{-T_{n,m} + \mathbb{E}_{F,G}\{T_{n,m}\} \geq \mathbb{E}_{F,G}\{T_{n,m}\} - \frac{1}{\alpha}\mathbb{E}\{T^{\pi}_{n,m} \mid \mathcal{U}\}\}.\end{aligned}$$

Since $\mathbb{E}_{F,G}\{T_{n,m}\} \geq \Theta^2_{\rho}(F, G) \geq c\lambda(n, m) \geq K_1\lambda(n, m) \geq \frac{1}{\alpha}\mathbb{E}\{T^{\rho}_{n,m,\pi} \mid \mathcal{U}\}$, using the Chebyshev's inequality, one gets

$$\begin{aligned}&\mathbb{P}^{n,m}_{F,G}\{T_{n,m} \leq c_{1-\alpha}\} \leq \mathbb{P}^{n,m}_{F,G}\{-T_{n,m} + \mathbb{E}_{F,G}\{T_{n,m}\} \geq \mathbb{E}_{F,G}\{T_{n,m}\} - \frac{1}{\alpha}\mathbb{E}\{T^{\pi}_{n,m} \mid \mathcal{U}\}\} \\ &\leq \frac{Var_{F,G}(T_{n,m})}{\big(\mathbb{E}_{F,G}\{T_{n,m}\} - \frac{1}{\alpha}\mathbb{E}\{T^{\pi}_{n,m} \mid \mathcal{U}\}\big)^2} \leq \frac{C_1\Theta^2_{\rho}(F, G)\Big(\frac{1}{n} + \frac{1}{m}\Big) + C_2\Big(\frac{1}{n} + \frac{1}{m}\Big)^2}{\Big(a_{n,m} + \Theta^2_{\rho}(F, G) - \frac{1}{6\alpha}\Big(\frac{1}{n} + \frac{1}{m} + \frac{1}{n-2} + \frac{1}{m-2}\Big)\Big)^2} \\ &\leq \frac{C_1\Theta^2_{\rho}(F, G)\Big(\frac{1}{n} + \frac{1}{m}\Big) + C_2\Big(\frac{1}{n} + \frac{1}{m}\Big)^2}{\Big(\Theta^2_{\rho}(F, G) - \frac{1}{6\alpha}\Big(\frac{1}{n} + \frac{1}{m} + \frac{1}{n-2} + \frac{1}{m-2}\Big)\Big)^2},\end{aligned}$$

where $a_{n,m} = \frac{1}{6}\left(\frac{1}{n-2} + \frac{1}{m-2}\right) + \frac{1}{m}(p_0 - p_1) + \frac{1}{n}(p_2 - p_3)$. This implies that

$$\limsup_{n,m,d\to\infty} \sup_{(F,G)\in\mathcal{F}(c\lambda(n,m))} \mathbb{P}^{n,m}_{F,G}\{T_{n,m} \leq c_{1-\alpha}\} \leq (C_1 c + C_2)/\Big(c - \frac{1}{3\alpha}\Big)^2.$$

One can notice that this upper bound is a decreasing function in c, and as c grows to infinity, it goes to zero. Hence, for any $0 < \zeta < 1 - \alpha$, there exists a constant $K_2 > 0$ such that this upper bound is smaller than ζ. Now let $K(\alpha, \zeta) = \max\{K_1, K_2\}$. Then, for $c > K(\alpha, \zeta)$, the maximum type II error rate is asymptotically upper bounded by ζ. This establishes the theorem. ■

Proof of Theorem 2.6 . If the two distributions F and G are such that $\lim_{d\to\infty} \Theta^2_{\ell_2}(F, G)/\lambda(n, m) = \infty$, then from Theorem 2.5, we have $\lim_{d\to\infty} \mathbb{P}^{n,m}_{F,G}\{T_{n,m} \leq c_{1-\alpha}\} = 0$. Hence, the power of the test converges to 1. ■

Proof of Theorem 2.7 . Using similar arguments as in the proofs of Theorems 2.4 and 2.5, one can show that if h and ψ are strictly increasing functions, then for testing $H_0 : \Theta^2_{\varphi_{h,\psi}}(F, G) = 0$ against $H_1 : \Theta^2_{\varphi_{h,\psi}}(F, G) > \epsilon$, the minimax rate of separation is $\lambda(n, m) = (1/\sqrt{n} + 1/\sqrt{m})^2$, and the permutation test based on $T^{h,\psi}_{n,m}$ is minimax rate optimal. Hence, one gets a similar conclusion as in Theorem 2.6. ■

Proof of Proposition 2.1 . Here, we use Lemma A2.5 to establish the condition of Theorem 2.6 for different β. Assume that $\mathbf{X}_1, \mathbf{X}_2 \sim F = \otimes^d_{i=1}\mathcal{N}_1(1/d^{\beta}, 1)$, $\mathbf{Y}_1, \mathbf{Y}_2 \sim G = \otimes^d_{i=1}\mathcal{N}_1(-1/d^{\beta}, 1)$,

and they are independent. Then

$$\begin{aligned}\Theta^2_{\ell_2}(F,G) &\geq \Big[\mathbb{P}\big(\|\mathbf{X}_1-\mathbf{Y}_1\|\leq\|\mathbf{Y}_2-\mathbf{Y}_1\|\big)-\frac{1}{2}\Big]^2+\Big[\mathbb{P}\big(\|\mathbf{Y}_1-\mathbf{X}_1\|\leq\|\mathbf{X}_2-\mathbf{X}_1\|\big)-\frac{1}{2}\Big]^2\\ &=\Big[\mathbb{P}\big(\|\mathbf{X}_1-\mathbf{Y}_1\|^2-\|\mathbf{Y}_2-\mathbf{Y}_1\|^2\leq 0\big)-\frac{1}{2}\Big]^2\\ &\qquad+\Big[\mathbb{P}\big(\|\mathbf{Y}_1-\mathbf{X}_1\|^2-\|\mathbf{X}_2-\mathbf{X}_1\|^2\leq 0\big)-\frac{1}{2}\Big]^2\\ &=\left[\mathbb{P}\Big(\sum_{i=1}^d T_i^*S_i^*\leq 0\Big)-\frac{1}{2}\right]^2+\left[\mathbb{P}\Big(\sum_{i=1}^d T_i'S_i'\leq 0\Big)-\frac{1}{2}\right]^2,\end{aligned}$$

where $T_i^*=X_1^{(i)}-Y_2^{(i)}$, $S_i^*=X_1^{(i)}+Y_2^{(i)}-2Y_1^{(i)}$, $T_i'=Y_1^{(i)}-X_2^{(i)}$ and $S_i'=Y_1^{(i)}+X_2^{(i)}-2X_1^{(i)}$ $(i=1,2,\ldots,d)$. Clearly, T_i^*,S_i^* are independent, and so are T_i',S_i'. Here $S_1^*,S_2^*,\ldots,S_d^*\overset{iid}{\sim}N(\frac{2}{d^\beta},6)\big)$, and $S_i^{'}$ has the same distribution as $-S_i^*$ for all $i=1,2,\ldots,d$. Now,

$$\begin{aligned}&\Big[\mathbb{P}\big(\|\mathbf{X}_1-\mathbf{Y}_1\|\leq\|\mathbf{Y}_2-\mathbf{Y}_1\|\big)-1/2\Big]^2+\Big[\mathbb{P}\big(\|\mathbf{Y}_1-\mathbf{X}_1\|\leq\|\mathbf{X}_2-\mathbf{X}_1\|\big)-1/2\Big]^2\\ &=\left[\mathbb{E}\left\{\Phi\left(-\frac{2}{d^\beta}\frac{\sum_{i=1}^d S_i^*}{\sqrt{2\sum_{i=1}^d S_i^{*2}}}\right)\right\}-\frac{1}{2}\right]^2+\left[\mathbb{E}\left\{\Phi\left(\frac{2}{d^\beta}\frac{\sum_{i=1}^d S_i'}{\sqrt{2\sum_{i=1}^d S_i'^2}}\right)\right\}-\frac{1}{2}\right]^2.\\ &=2\left[\mathbb{E}\left\{\Phi\left(-\frac{2}{d^\beta}\frac{\sum_{i=1}^d S_i^*}{\sqrt{2\sum_{i=1}^d S_i^{*2}}}\right)\right\}-\frac{1}{2}\right]^2\end{aligned}$$

Hence, studying the behaviour of $Z(\beta)=\big(2\sum_{i=1}^d S_i^*\big)/\big(d^\beta\sqrt{2\sum_{i=1}^d S_i^{*2}}\big)$ for different values of β, one can get the conditions for the consistency of our test.

Note that $\sum_{i=1}^d S_i^*/d^{\beta+1/2}\sim N(2/d^{2\beta-1/2},6/d^{2\beta})$. Hence for $\beta<1/4$, $\frac{1}{d^{\beta+1/2}}\sum_{i=1}^d S_i^*\xrightarrow{P}\infty$ and $\sum_{i=1}^d S_i^{*2}/d\xrightarrow{P}6$. So, for $\beta<1/4$, $Z(\beta)\xrightarrow{P}\infty$. For $\beta=1/4$, $\sum_{i=1}^d S_i^*/d^{\beta+1/2}\xrightarrow{P}2$. So, $Z(\beta)\xrightarrow{P}2/\sqrt{3}$. Therefore, for $\beta\leq 1/4$, we have

$$\liminf_{d\to\infty}\left[\mathbb{E}\left\{\Phi\left(-\frac{2}{d^\beta}\frac{\sum_{i=1}^d S_i^*}{\sqrt{2\sum_{i=1}^d S_i^{*2}}}\right)\right\}-\frac{1}{2}\right]^2>0,$$

which in turn implies that $\liminf_{d\to\infty}\Theta^2_{\ell_2}(F,G)>0$. This proves Proposition 2.1(a).

For $1/4<\beta<1/2$, notice that $d^{2\beta-1/2}\sum_{i=1}^d S_i^*/d^{\beta+1/2}\sim N(2,6d^{2\beta-1})$. So, as d tends to infinity, $d^{2\beta-1/2}\sum_{i=1}^d S_i^*/d^{\beta+1/2}\xrightarrow{P}2$. Now, if we take $n\asymp m\asymp d^\gamma$, to match this convergence rate so that $\Theta^2_{\ell_2}(F,G)/\lambda(n,m)$ diverges to infinity, we require the following

$$\lim_{d\to\infty}d^\gamma\left[\mathbb{E}\left\{\Phi\left(-\frac{2}{d^{2\beta-1/2}}\frac{d^{\beta-1}\sum_{i=1}^d S_i^*}{\sqrt{2\sum_{i=1}^d S_i^{*2}/d}}\right)\right\}-\frac{1}{2}\right]^2=\infty.$$

This is possible when $\gamma>4\beta-1$. Also, note that for $\beta=1/2$, $\sum_{i=1}^d S_i^*/d^{1/2}$ forms a tight sequence. In this case, we need

$$\lim_{d\to\infty} d^{\gamma}\left[\mathbb{E}\left\{\Phi\Big(-\frac{2}{d^{\beta}}\frac{d^{-1/2}\sum_{i=1}^{d}S_i^*}{\sqrt{2\sum_{i=1}^{d}S_i^{*2}/d}}\Big)\right\}-\frac{1}{2}\right]^2=\infty,$$

which is satisfied when $\gamma > 1 = 4\beta - 1$. This proves Proposition 2.1(b).

Now for $\beta > 1/2$, we have $\sum_{i=1}^{d} S_i^*/d^{1/2} \sim N(2/d^{\beta-1/2}, 6)$, and hence it is a tight sequence of random variables. In this scenario, we require

$$\lim_{d\to\infty} d^{\gamma}\left[\mathbb{E}\left\{\Phi\Big(-\frac{2}{d^{\beta}}\frac{d^{-1/2}\sum_{i=1}^{d}S_i^*}{\sqrt{2\sum_{i=1}^{d}S_i^{*2}/d}}\Big)\right\}-\frac{1}{2}\right]^2=\infty,$$

which is satisfied if $\gamma > 2\beta$. Also notice that when $\beta > 1/2$ and $\gamma < 2\beta - 1$, the Kullback-Leibler Divergence ($KL(Q_1, Q_0) \asymp d^{\gamma-2\beta+1}$) converges to zero with increasing dimensions. Hence, in this scenario, the asymptotic type II error rate of any test remains bounded below by $1-\alpha$, i.e., no tests have asymptotic power more than the nominal level α. This completes the proof of Proposition 2.1(c). ■

Chapter 3

Test of Spherical Symmetry for High-Dimensional Data

An inherent property of nature is that it tends to exhibit some form of symmetry within itself. In the nineteenth century, such symmetric patterns were approximated by the normal distribution (see, e.g., Lehmann, 2012 for a history of statistical methods). However, with time, more general notions of symmetry were introduced. One of the most popular notions is spherical symmetry or elliptic symmetry (i.e., spherical symmetry after standardization) (see, e.g. Chmielewski, 1981; Fang, Kotz & Ng, 1990; Fourdrinier, Strawderman & Wells, 2018). This is an important class of distributions, and testing the sphericity of a distribution is an important statistical problem. Here, we want to test the null hypothesis of spherical symmetry of the underlying distribution P based on a sample $\mathcal{D} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ of n independent realizations of the random vector $\mathbf{X} \sim \mathrm{P}$.

We know that a random vector $\mathbf{X}$ follows a spherically symmetric distribution if and only if $\mathbf{X}$ and $\mathbf{HX}$ have the same distribution (i.e., $\mathbf{X} \overset{D}{=} \mathbf{HX}$) for any orthogonal matrix $\mathbf{H}$. However, we can also characterize spherical symmetry using the following lemma.

Lemma 3.1. *A d-dimensional random vector $\mathbf{X}$ is spherically symmetric if and only if $\mathbf{X} \overset{D}{=} \|\mathbf{X}\|\mathbf{U}$, where $\mathbf{U}$ is independent of $\|\mathbf{X}\|$, and it is uniformly distributed over $\mathcal{S}^{d-1}$.*

For proof of Lemma 3.1, see page 31 in Fang, Kotz & Ng (1990). Using this characterization, we construct a new measure of spherical asymmetry $\zeta(\mathrm{P})$ for any probability distribution P. Let φ_1 be the characteristic function of $\mathbf{X} \sim \mathrm{P}$ and φ_2 be that of its spherically symmetric variant $\mathbf{X}' = \|\mathbf{X}\|\mathbf{U}$, where $\mathbf{U} \sim \mathrm{Unif}(\mathcal{S}^{d-1})$ and $\|\mathbf{X}\|$ are independent. We define

$$\zeta(\mathrm{P}) = \int \left|\varphi_1(\mathbf{t}) - \varphi_2(\mathbf{t})\right|^2 \mathrm{d}W(\mathbf{t}), \tag{3.1}$$

where W is a non-negative measure equivalent to the Lebesgue measure, and the integral is taken in the principal value sense. The following proposition proves the characterization property of $\zeta(\mathrm{P})$.

Proposition 3.1. *For any distribution* P, $\zeta(\mathrm{P})$ *is non-negative, and it takes the value* 0 *if and only if* P *is spherically symmetric.*

For suitable choices of W, $\zeta(\mathrm{P})$ has nice closed-form expressions. For instance, if W is taken as the probability measure corresponding to the d-dimensional Gaussian distribution $\mathcal{N}_d(\mathbf{0}_d, \frac{1}{d}\mathbf{I}_d)$, one can derive an alternative form for $\zeta(\mathrm{P})$ using the following lemma.

Lemma 3.2. *For any two non-zero vectors* $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^d$,

$$\int \exp\{i\langle \mathbf{t}, \mathbf{X}_1 - \mathbf{X}_2\rangle\} \frac{d^{d/2}}{(2\pi)^{d/2}} e^{-d\|\mathbf{t}\|^2/2} \mathrm{d}\mathbf{t} = \exp\Big\{ -\frac{1}{2d}\|\mathbf{X}_1 - \mathbf{X}_2\|^2 \Big\}.$$

Using Lemma 3.2, we get an alternative representation of $\zeta(\mathrm{P})$ given in the following theorem.

Theorem 3.1. *If* $\mathbf{X}_1, \mathbf{X}_2$ *are independent copies of* $\mathbf{X} \sim \mathrm{P}$ *and* W *is the probability measure corresponding to* $\mathcal{N}_d(\mathbf{0}_d, \frac{1}{d}\mathbf{I}_d)$, *then* $\zeta(\mathrm{P})$ *can be expressed as*

$$\zeta(\mathrm{P}) = \mathbb{E}\Big\{ \exp\Big\{ -\frac{\|\mathbf{X}_1 - \mathbf{X}_2\|^2}{2d} \Big\} \Big\} + \mathbb{E}\Big\{ \exp\Big\{ -\frac{\|\mathbf{X}_1' - \mathbf{X}_2'\|^2}{2d} \Big\} \Big\} - 2\mathbb{E}\Big\{ \exp\Big\{ -\frac{\|\mathbf{X}_1 - \mathbf{X}_2'\|^2}{2d} \Big\} \Big\},$$

where $\mathbf{X}_i' = \|\mathbf{X}_i\|\mathbf{U}_i$ *for* $i = 1, 2$, *and* $\mathbf{U}_1, \mathbf{U}_2 \overset{iid}{\sim} \mathrm{Unif}(\mathcal{S}^{d-1})$ *are independent of* $\mathbf{X}_1$ *and* $\mathbf{X}_2$.

There are several other choices of W for which we have a closed-form expression for $\zeta(\mathrm{P})$. But, unless mentioned otherwise, throughout this chapter, we shall use the measure W considered in Theorem 3.1. For this choice of W, $\zeta(\mathrm{P})$ has some interesting features.

Remark 3.1. *Theorem 3.1 shows that* $\zeta(\mathrm{P})$ *can be expressed as a function of the pairwise Euclidean distances between* $\mathbf{X}_1, \mathbf{X}_2$ *and their spherically symmetric variants* $\mathbf{X}_1', \mathbf{X}_2'$. *So,* $\zeta(\mathrm{P})$ *is invariant under rotation. In particular, we have* $\zeta(\mathcal{N}_d(\mathbf{0}, \mathbf{\Sigma})) = \zeta(\mathcal{N}_d(\mathbf{0}, \mathbf{H}\mathbf{\Sigma}\mathbf{H}^T))$ *for any orthogonal matrix* $\mathbf{H}$. *For the exact expression of* $\zeta(\mathcal{N}_d(\mathbf{0}, \mathbf{\Sigma}))$, *interested readers are referred to Section 3.5.1.*

Remark 3.2. $\zeta(\mathrm{P})$ *can also be viewed as the energy distance (see, e.g., Székely & Rizzo, 2004) or the maximum mean discrepancy (MMD) (see, e.g., Gretton et al., 2007) between the distributions of* $\mathbf{X}$ *and* $\|\mathbf{X}\|\mathbf{U}$. *MMD is used to quantify the distributional difference using embedding into a Reproducing Kernel Hilbert Space (RKHS). Here* $K(\mathbf{x}, \mathbf{y}) = \exp\{-\|\mathbf{x} - \mathbf{y}\|^2/(2d)\}$ *is the kernel associated with that RKHS, and we use this notation throughout the rest of the chapter.*

3.1 Estimation of $\zeta(\mathrm{P})$

Let $\mathcal{D} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ be a random sample of size n from a d-dimensional distribution P. Note that the term $\mathbb{E}[K(\mathbf{X}_1, \mathbf{X}_2)]$ in $\zeta(\mathrm{P})$ can be easily estimated by its empirical analog, but $\zeta(\mathrm{P})$ involves two other terms $\mathbb{E}[K(\mathbf{X}_1', \mathbf{X}_2')]$ and $\mathbb{E}[K(\mathbf{X}_1, \mathbf{X}_2')]$, which are not estimable from $\mathcal{D}$ alone. Therefore, we adopt the following data augmentation approach.

- Generate $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$, a random sample of size n from $\mathrm{Unif}(\mathcal{S}^{d-1})$. Define $\mathbf{X}_i' = \|\mathbf{X}_i\|\mathbf{U}_i$ for $i = 1, 2, \ldots, n$ and the augmented data set $\mathcal{D}_A = \{(\mathbf{X}_i, \mathbf{X}_i') : i = 1, \ldots, n\}$.
- Using observations from $\mathcal{D}_A$ we propose an estimator of $\zeta(\mathrm{P})$ as

$$\hat{\zeta}_n = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \Big\{ K(\mathbf{X}_i, \mathbf{X}_j) + K(\mathbf{X}_i', \mathbf{X}_j') - 2K(\mathbf{X}_i, \mathbf{X}_j') \Big\}.$$

Clearly $\hat{\zeta}_n$ can be viewed as a U-statistic with the symmetrized kernel function

$$g\big((\mathbf{x}_1, \mathbf{x}_1'), (\mathbf{x}_2, \mathbf{x}_2')\big) = K(\mathbf{x}_1, \mathbf{x}_2) + K(\mathbf{x}_1', \mathbf{x}_2') - K(\mathbf{x}_1, \mathbf{x}_2') - K(\mathbf{x}_2, \mathbf{x}_1'). \tag{3.2}$$

Note that $\hat{\zeta}_n$ is an unbiased estimator of $\zeta(\mathrm{P})$. Since g is bounded, using the bounded difference inequality, we establish the following bound for the deviation of $\hat{\zeta}_n$ from its population analog $\zeta(\mathrm{P})$.

Theorem 3.2. *If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent copies of a d-dimensional random vector $\mathbf{X} \sim \mathrm{P}$, then $\mathbb{P}\Big\{\big|\hat{\zeta}_n - \zeta(\mathrm{P})\big| > \epsilon\Big\} \leq 2\exp\{-\frac{n\epsilon^2}{32}\}$, and this inequality holds irrespective of the dimension d.*

Theorem 3.2 shows that $\hat{\zeta}_n$ is a strongly consistent estimator of $\zeta(\mathrm{P})$, and the exponential bound is free from d. The large sample distribution of $\hat{\zeta}_n$ can be derived using the theory of U-statistics (see Lee, 1990). The asymptotic null distribution of $\hat{\zeta}_n$ is given by Theorem 3.3.

Theorem 3.3. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent copies of the random vector $\mathbf{X}$, which follows a spherically symmetric distribution P. Define $\mathbf{X}_i' = \|\mathbf{X}_i\|\mathbf{U}_i$ for $i = 1, 2, \ldots, n$, where $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$ are i.i.d. $\mathrm{Unif}(\mathcal{S}^{d-1})$ and independent of $\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_n$. Let λ_k $(k = 1, 2, \ldots)$ be the eigenvalue corresponding to the eigenfunction ψ_k of the integral equation*

$$\mathbb{E}\{g\big((\mathbf{x}, \mathbf{x}'), (\mathbf{X}, \mathbf{X}')\big)\psi_k\big(\mathbf{X}, \mathbf{X}'\big)\} = \lambda_k \psi_k(\mathbf{x}, \mathbf{x}'),$$

where g is as in equation (3.2). Then as n goes to infinity, $n\hat{\zeta}_n \xrightarrow{D} \sum_{k=1}^{\infty} \lambda_k (Z_i^2 - 1)$, where $\{Z_k : k \geq 1\}$ is a sequence of i.i.d. standard normal variables.

In Theorem 3.3, the limiting null distribution of $\hat{\zeta}_n$ (after appropriate adjustment for location and scale) turns out to be a weighted sum of independent chi-squares due to the first-order degeneracy of g under H_0 (i.e., spherical symmetry). However, under H_1, g is a non-degenerate kernel. Therefore, the limiting distribution of $\hat{\zeta}_n$ (after an appropriate adjustment for location and scale) turns out to be Gaussian, which is asserted by the following theorem.

Theorem 3.4. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent copies of a random vector $\mathbf{X}$, which follows a non-spherical distribution P. Define $\mathbf{X}_i' = \|\mathbf{X}_i\|\mathbf{U}_i$ $(i = 1, 2, \ldots, n)$, where $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$ are i.i.d. $\mathrm{Unif}(\mathcal{S}^{d-1})$ and independent of $\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_n$. Then as $n \to \infty$, $\sqrt{n}\big(\hat{\zeta}_n - \zeta(\mathrm{P})\big) \xrightarrow{D} N(0, 4\sigma^2)$, where g is as in Equation (3.2) and $\sigma^2 = Var\Big\{\mathbb{E}\big\{g\big((\mathbf{X}_1, \mathbf{X}_1'), (\mathbf{X}_2, \mathbf{X}_2')\big) \mid (\mathbf{X}_1, \mathbf{X}_1')\big\}\Big\}$.*

3.2 Test of Spherical Symmetry

Theorem 3.2 shows that $\hat{\zeta}_n$ is a strongly consistent estimator of $\zeta(\mathrm{P})$. From Proposition 3.1, it is also clear that under H_0, $\hat{\zeta}_n$ converges almost surely to zero, but under H_1, it converges to a positive constant. Therefore, the power of a test that rejects H_0 for higher values of $\hat{\zeta}_n$ converges to one as the sample size increases. However, it is difficult to find the critical value based on the asymptotic null distribution of $n\hat{\zeta}_n$ since it involves an ℓ_2 sequence $\{\lambda_k\}$ (i.e., $\sum_{k=1}^{\infty} \lambda_k^2$ is finite), which depends on the underlying distribution P (see Theorem 3.3). Though our test has a similarity with the test based on Maximum Mean Discrepancy (MMD) (see Gretton et al., 2012), it can not be calibrated correctly using the permutation method. This is due to the existing dependence between the observations and their spherically symmetric variants even under H_0. However, under H_0, $\mathbf{X}_i$

and $\mathbf{X}_i'$ are exchangeable for every $i = 1, 2, \ldots, n$. Therefore, a random swap between $\mathbf{X}_i$ and $\mathbf{X}_i'$ does not change the null distribution of $\hat{\zeta}_n$. Utilizing this fact, we construct a novel resampling algorithm to compute the cut-off. The algorithm is given below.

Resampling algorithm

A. Given the augmented data $\mathcal{D}_A$ compute the test statistic $\hat{\zeta}_n$.

B. Let $\boldsymbol{S} = (S_1, \ldots, S_n)$ be an element in $\{0,\ 1\}^n$. Define $\mathbf{Y}_i = S_i\mathbf{X}_i + (1 - S_i)\mathbf{X}_i'$ and $\mathbf{Y}_i' = (1 - S_i)\mathbf{X}_i + S_i\mathbf{X}_i'$ for $i = 1, 2, \ldots, n,$. Use $(\mathbf{Y}_1, \mathbf{Y}_1'), \ldots, (\mathbf{Y}_n, \mathbf{Y}_n')$ to compute $\hat{\zeta}_n(\boldsymbol{S})$, the resampling analogue of $\hat{\zeta}_n$.

C. Repeat step B for all possible $\boldsymbol{S}$ to get the critical value for a level α $(0 < \alpha < 1)$ test given by

$$c_{1-\alpha} = \inf\{t \in \mathbb{R} : \frac{1}{2^n} \sum_{\boldsymbol{S} \in \{0,\ 1\}^n} \mathrm{I}[\hat{\zeta}_n(\boldsymbol{S}) \leq t] \geq 1 - \alpha\}.$$

We reject H_0 if $\hat{\zeta}_n$ is larger than $c_{1-\alpha}$. The p-value of this conditional test is given by

$$p_n = \frac{1}{2^n} \sum_{\boldsymbol{S} \in \{0,\ 1\}^n} \mathrm{I}[\hat{\zeta}_n(\boldsymbol{S}) \geq \hat{\zeta}_n]. \tag{3.3}$$

So, alternatively, we can reject H_0 if $p_n < \alpha$. The following lemma shows that this resampling algorithm gives a valid level α test.

Lemma 3.3. *Let $\hat{\zeta}_n$ be the estimator of $\zeta(\mathrm{P})$ as defined in Theorem 3.1. If p_n denotes the conditional p-value as defined in (3.3), then under $H_0 : \zeta(\mathrm{P}) = 0$ we have $\mathbb{P}\{p_n < \alpha\} \leq \alpha$ irrespective of the values of n and d.*

Interestingly, we can control the threshold $c_{1-\alpha}$ by a deterministic sequence that does not depend on d and converges to 0 as n increases. This is shown in the following theorem.

Lemma 3.4. *If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent copies of a d-dimensional random vector $\mathbf{X} \sim \mathrm{P}$, then for any α $(0 < \alpha < 1)$, the inequality $c_{1-\alpha} \leq 2(\alpha(n-1))^{-1}$ holds with probability one.*

So, irrespective of the value of d, $c_{1-\alpha}$ is of order $O_P(n^{-1})$, and it converges to zero almost surely as n diverges to infinity. For any fixed d, we can also show that for any spherical distribution P, $n\hat{\zeta}_n(\boldsymbol{S})$ converges in distribution to a weighted sum of independent chi-squares as n diverges to infinity. Since, under H_1, $\hat{\zeta}_n$ converges to a positive number, the conditional p-value p_n converges to zero as n increases. Hence, the power of the resulting conditional test converges to one. This is formally stated in the following theorem.

Theorem 3.5. *For any fixed alternative, the power of the conditional test based on p_n converges to one as n diverges to infinity.*

Though the above resampling algorithm leads to a consistent level α test, it has a computational complexity of the order $O(n^2 2^n)$. So, it is not computationally feasible to implement

even if the sample size is moderately large. Therefore, in practice, we generate $\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_B$ uniformly from $\{0,\ 1\}^n$ and compute the randomized p-value

$$p_{n,B} = \frac{1}{B+1}\Big(\sum_{i=1}^{B} I\{\hat{\zeta}_n(\boldsymbol{S}_i) \geq \hat{\zeta}_n\} + 1\Big).$$

We reject H_0 if $p_{n,B} < \alpha$. The following theorem shows that $p_{n,B}$ closely approximates p_n for large B and thereby justifies the use $p_{n,B}$ for the practical implementation of the test.

Theorem 3.6. *Given the augmented data $\mathcal{D}_A$, $|p_{n,B} - p_n| \xrightarrow{a.s} 0$ as B diverges to infinity.*

3.3 Asymptotic properties of the test

In this section, we study some large sample properties of our test. First, we investigate the robustness of our test against contamination alternatives. Next, we establish its minimax rate optimality against a suitable class of nonparametric alternatives and prove its consistency even when the dimension of data increases with the sample size. Finally, we show that our test is efficient in the Pitman sense under contiguous contamination alternatives.

3.3.1 Robustness

Consider a distribution F, which is not spherically symmetric. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be i.i.d. random vectors from a contaminated distribution $F_\delta = (1-\delta)F + \delta G$ $(0 < \delta < 1)$. The following lemma shows the effect of this contamination on $\zeta(\cdot)$ by providing a relation among $\zeta(F), \zeta(G)$ and $\zeta(F_\delta)$.

Lemma 3.5. *For any $\delta \in (0,1)$, we have $\zeta(F_\delta) = (1-\delta)^2\zeta(F) + \delta^2\zeta(G) + 2\delta(1-\delta)\zeta'(G,F)$, where $\zeta'(G,F) = \mathbb{E}\{K(\mathbf{X}_1, \mathbf{Y}_1)\} + \mathbb{E}\{K(\mathbf{X}_1', \mathbf{Y}_1')\} - \mathbb{E}\{K(\mathbf{X}_1, \mathbf{Y}_1')\} - \mathbb{E}\{K(\mathbf{X}_1', \mathbf{Y}_1)\}$. Here $\mathbf{X}_1 \sim G$ and $\mathbf{Y}_1 \sim F$ are independent, $\mathbf{X}_1' = \|\mathbf{X}_1\|\mathbf{U}_1$, $\mathbf{Y}_1' = \|\mathbf{Y}_1\|\mathbf{U}_2$, where $\mathbf{U}_1, \mathbf{U}_2$ are i.i.d. $\mathrm{Unif}(\mathcal{S}^{d-1})$ and $K(\mathbf{x}, \mathbf{y}) = \exp\{-\frac{1}{2d}\|\mathbf{x}-\mathbf{y}\|^2\}$.*

Note that if G is spherically symmetric, $\zeta(G) = 0$ and $\zeta'(G,F) = 0$. Therefore, for any fixed δ, we have $\zeta(F_\delta) = (1-\delta)^2\zeta(F)$. This shows that $\zeta(\cdot)$ has a bounded Gateaux derivative. The following theorem also shows that for any $\delta \in (0,1)$ and G spherically symmetric, the power of our test for the contaminated alternative F_δ converges to one as the sample size increases.

Theorem 3.7. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be n independent realizations of $\mathbf{X} \sim F_\delta = (1-\delta)F + \delta G$, where $0 < \delta < 1$, and $\zeta(F) > 0$. Then the minimum power of the proposed test over the class of the spherical distributions G, i.e., $\inf_{G:\zeta(G)=0} \mathbb{P}\{\hat{\zeta}_n > c_{1-\alpha}\}$, converges to one as n diverges to infinity.*

This theorem shows that our test is asymptotically robust against outliers for any fixed proportion of contamination. Since $\zeta(F_{1-\delta}) = \delta^2\zeta(F)$, it shows the convergence of the power of our test for $F_{1-\delta}$ as well. So, if a sample from a spherically symmetric distribution has a small proportion of contamination by observations from a non-spherical distribution, our test can successfully detect the presence of those contaminations when the sample size is large. The result in Theorem 3.7

holds even for a sequence $\{\delta_n\}$ that remains bounded away from one. However, if that is not the case, the asymptotic power of the test will depend on the convergence rate of $1-\delta_n$ and may yield non-trivial limits for certain choices of $\{\delta_n\}$. This is explored in the following subsection.

3.3.2 Minimax rate optimality and high-dimensional behaviour

Let us consider a testing problem involving a pair of hypotheses $H_0 : \zeta(\mathrm{P}) = 0$ and $H_1' : \zeta(\mathrm{P}) > \epsilon(n)$ where $\epsilon(n)$ is a positive number that depends on the sample size n. Let $\mathcal{F}(\epsilon(n)) := \{\mathrm{P} \mid \zeta(\mathrm{P}) > \epsilon(n)\}$ be the class of alternatives under H_1', and $\mathbb{T}_n(\alpha)$ be the class of all level α test. The minimax type II error rate for this problem is defined as

$$R_n\big(\epsilon(n)\big) = \inf_{\phi \in \mathbb{T}_n(\alpha)} \sup_{F \in \mathcal{F}(\epsilon(n))} \mathbb{P}_F^{(n)}\{\phi = 0\},$$

where $\mathbb{P}_F^{(n)}$ denotes the probability corresponding to the joint distribution of $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$, and the $\mathbf{X}_i$s are independent copies of $\mathbf{X} \sim F$. Here, we want to find an optimum choice of $\epsilon(n)$ (call it $\epsilon_0(n)$) such that it satisfies conditions (a) and (b) mentioned in Section 2.2.1. Theorem 3.8 below shows that $\epsilon_0(n)$ cannot be of smaller order than $O(n^{-1})$. In other words, for any $0 < \beta < 1-\alpha$ and any $\phi \in \mathbb{T}_n(\alpha)$, we can always find a distribution F with $\zeta(F)$ of the order $O(n^{-1})$ or smaller such that the type II error rate of the test ϕ is more than β, i.e., $\mathbb{P}_F^{(n)}\{\phi = 0\} \geq \beta$.

Theorem 3.8. *For $0 < \beta < 1-\alpha$, there exists a constant $c_0(\alpha, \beta)$ such that the minimax type II error rate $R_n(cn^{-1})$ is lower bounded by β for all n and all $0 < c < c_0(\alpha, \beta)$.*

Remark 3.3. *Consider the distribution $F_{\delta_n} = (1-\delta_n)F + \delta_n G$, where $\zeta(G) = 0$ and $\zeta(F) > 0$ (note that $\zeta(F_{\delta_n}) = (1-\delta_n)^2\zeta(F)$). If δ_n is such that $n(1-\delta_n)^2 \to 0$ as $n \to \infty$ (i.e. $\zeta(F_{\delta_n})$ is of smaller asymptotic order than $O(n^{-1})$), then the power of any level α test for the alternative F_{δ_n} will fall below the nominal level α.*

In the next theorem, we establish that in the case of $\epsilon_0(n) = n^{-1}$, our test based on $\hat{\zeta}_n$ satisfies the condition (b) stated above. Therefore, these two theorems (Theorem 3.8 and 3.9) together show that the minimax rate of separation is $\epsilon_0(n) = n^{-1}$, and our proposed test has the minimax rate optimality for the class of alternatives $\mathcal{F}(\epsilon(n))$.

Theorem 3.9. *For any $\beta \in (0, 1-\alpha)$, there exists a constant $C_0(\alpha, \beta)$ (independent of d) such that asymptotically the maximum type II error of the test based on $\hat{\zeta}_n$ over $\mathcal{F}(cn^{-1})$ is uniformly bounded above by β for all $c > C_0(\alpha, \beta)$, i.e., $\limsup_{n\to\infty} \sup_{F \in \mathcal{F}(c\lambda(n))} \mathbb{P}_F^{(n)}(\hat{\zeta}_n \leq c_{1-\alpha}) \leq \beta$ for all $c > C_0(\alpha, \beta)$.*

Remark 3.4. *Consider the same example as in Remark 3.3. Since $\zeta(F_{\delta_n}) = (1-\delta_n)^2\zeta(F)$, from Theorem 3.9 it is clear that if $n(1-\delta_n)^2 \to \infty$ as $n \to \infty$ (i.e. $\zeta(F_{\delta_n})$ is of higher asymptotic order than $O(n^{-1})$), the power of our proposed test for the alternative F_{δ_n} converges to one.*

Note that the constant $C(\alpha, \beta)$ in Theorem 3.9 does not depend on the dimension d. However, $\zeta(F)$ may vary with the dimension. We know that under certain regularity conditions (see, e.g., Hall, Marron & Neeman, 2005; Ahn et al., 2007; Jung & Marron, 2009), as the dimension increases,

pairwise distances among the observations (after appropriate scaling) converge to a constant, and all observations tend to lie on the surface of a sphere of increasing radius. So, in such situations, $\zeta(F)$ converges to 0 as d increases. Therefore, one may be curious to know how this test performs if the dimension and the sample size increase simultaneously. Theorem 3.9 answers this question. It shows that as long as $n\zeta(F)$ diverges to infinity, irrespective of whether d is fixed or it increases with the sample size, the power of our test converges to one. This high-dimensional consistency of our test is asserted by the following theorem.

Theorem 3.10. *Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent copies of $\mathbf{X} \sim F^{(d)}$, a d-dimensional distribution. If $d = d(n)$ grows with the sample size n such that $n\zeta(F^{(d)})$ diverges to infinity as n increases, then the power of the test based on $\hat{\zeta}_n$ converges to one as n and d both diverge to infinity.*

So, even if $\zeta(F^{(d)})$ converges to 0 as d increases with n, the power of our test converges to one as long as $\zeta(F^{(d)})$ converges at a slower rate than $O(n^{-1})$. However, if the distance convergence does not hold and we have $\liminf_{d\to\infty} \zeta(F^{(d)}) > 0$, the power of our test converges to one even if the sample size increases at a very slow rate. An example of such a distribution is given in Section 3.4 (see Example 3.7(a)). For such examples, one can expect the test to have good performance even in the HDLSS setup, where n is fixed (but suitably large), and d diverges to infinity. However, in the case of distance concentration in the HDLSS set-up, where we have $\liminf_{d\to\infty} \zeta(F^{(d)}) = 0$, we need to increase the sample size suitably to get good performance. This is further explored in our simulation studies.

3.3.3 Pitmann Efficiency

Now, consider the alternative $F_{1-\beta_n n^{-1/2}} = (1 - \beta_n n^{-1/2})G + (\beta_n n^{-1/2})F$, but assume that $\{\beta_n\}$ is a sequence of positive numbers converging to some $\beta \in (0, \infty)$. Let f and g be the densities corresponding to F and G, respectively. To study the asymptotic behaviour of our test for such an alternative, we first study the asymptotic behaviour of $n\hat{\zeta}_n$ and its resample analog $n\hat{\zeta}_n(\boldsymbol{S})$. The following result shows that under suitable assumption on F and G, the sequence of alternative asymmetric distributions $F_{1-\beta_n n^{-1/2}}$ is contiguous and locally asymptotically normal.

Proposition 3.2. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent copies of $\mathbf{X} \sim G$. If $\int \big(f(\mathbf{u})/g(\mathbf{u}) - 1\big)^2 g(\mathbf{u})\mathrm{d}\mathbf{u}$ is finite, as n grows to infinity, we have*

$$\left| \log\left\{ \prod_{i=1}^{n} \Big(1 + \frac{\beta_n}{\sqrt{n}}\Big\{\frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)} - 1\Big\}\Big)\right\} - \frac{\beta_n}{\sqrt{n}} \sum_{i=1}^{n} \left\{\frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)} - 1\right\} + \frac{\beta_n^2}{2}\mathbb{E}\left\{\frac{f(\mathbf{X}_1)}{g(\mathbf{X}_1)} - 1\right\}^2 \right| \xrightarrow{P} 0.$$

Using Proposition 3.2 and Le Cam's third lemma, we establish the local asymptotic behaviour of $n\hat{\zeta}_n$ in the following theorem.

Theorem 3.11. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent copies of $\mathbf{X} \sim F_{1-\beta_n n^{-1/2}}$ and $\mathbf{X}_i' = \|\mathbf{X}_i\|\mathbf{U}_i$ $(i = 1, \ldots, n)$, where $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$ are i.i.d.* Unif$(\mathcal{S}^{d-1})$. *Also let λ_k be the eigenvalues with*

corresponding eigenfunction ψ_k $(k = 1, 2, \ldots)$ *of the integral equation* $\mathbb{E}\{g\big((\mathbf{x}_1, \mathbf{x}_1'), (\mathbf{X}_1, \mathbf{X}_1')\big)\ \psi_k\big((\mathbf{X}_1, \mathbf{X}_1')\big)\} = \lambda_k \psi_k\big((\mathbf{x}_1, \mathbf{x}_1')\big)$, *where* g *is as in equation (3.2). Then, as* n *tends to infinity,*

$$n\hat{\zeta}_n \xrightarrow{D} \sum_{k=1}^{\infty} \lambda_k \left(\big(Z_k + \beta\ \mathbb{E}_F\{\psi_i(\mathbf{X}_1, \mathbf{X}_1')\}\big)^2 - 1 \right),$$

where Z_i *is a sequence of i.i.d. standard normal random variables.*

Theorem 3.11 shows that for $\beta > 0$, the local limit distribution of $n\hat{\zeta}_n$ is stochastically larger than its limiting null distribution given in Theorem 3.3. The following theorem establishes that the local limiting distribution of the permuted statistic $n\hat{\zeta}_n(\boldsymbol{S})$ under the sequence of alternatives $F_{1-\beta_n n^{-1/2}}$ is identical to the asymptotic null distribution of $n\hat{\zeta}_n$.

Theorem 3.12. *Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *be independent copies of* $\mathbf{X} \sim \mathrm{P}_n$ *and* $\hat{\zeta}_n(\boldsymbol{S})$ *be the resampling analog of our test statistic obtained using the resampling algorithm as in Section 3.2. Then under any fixed alternative (i.e.,* $\mathrm{P}_n = F$ *for some asymmetric distribution* F*) or a contiguous alternative (i.e.,* $\mathrm{P}_n = F_{1-\beta_n n^{-1/2}}$*), as* n *grows to infinity,* $n\hat{\zeta}_n(\boldsymbol{S}) \xrightarrow{D} \sum_{k=1}^{\infty} \lambda_k (Z_k^2 - 1)$*, where* $\{Z_k\}$ *is a sequence of i.i.d standard normal random variables and* $\{\lambda_k\}$ *is an* ℓ_2 *sequence of real numbers.*

Theorems 3.11 and 3.12 together show that under $F_{1-\beta_n n^{-1/2}}$, the power of our test converges to a non-trivial limit, and as β starts increasing from zero, the power of our test gradually increases from α to one. This establishes that our test is efficient in the Pitman sense. However, the exact expression of the limit is not analytically tractable.

We now present a small simulation study using n independent observations from $F_{1-\beta_n n^{-1/2}}$ in $\mathbb{R}^{10}$, where G is $\mathcal{N}_{10}(\mathbf{0}_{10}, \mathbf{I}_{10})$, F is $\mathcal{N}_{10}(0, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = 0.5\ \mathbf{I}_{10} + 0.5\ \mathbf{J}_{10}$ and $\mathbf{J}_d$ is the $d \times d$ matrix with all entries equal to one. We considered three different sequences (a) $\beta_n = 5n^{-0.1}$, (b) $\beta_n = 5$ and (c) $\beta_n = 5n^{0.1}$ and evaluated the performance of our test against these alternatives. The p-value of the test was approximated using the randomized p-value with $B = 500$, and the power of the test was evaluated based on 1000 repetitions of each experiment. In Table 3.1, we see that for case (a), the power of our test showed a decreasing trend with increasing sample size. For case (b), the power exhibited convergence towards 0.37, which can be considered as the Pitman efficiency of our test when $\beta_n = 5$. For case (c), we see that the power of our test converged to one with increasing sample size. This behaviour of our test supports our theoretical findings in this section.

Table 3.1 *Powers of the proposed test for the alternative* $F_{1-\beta_n n^{-1/2}} = (1 - \beta_n n^{-1/2})\mathcal{N}_{10}(\mathbf{0}_{10}, \mathbf{I}_{10}) + \beta_n n^{-1/2}\mathcal{N}_{10}(\mathbf{0}_{10}, 0.5\mathbf{I}_{10} + 0.5\mathbf{J}_{10})$ *when* $\beta_n = 5n^{\gamma}$.

Sample Size	$\gamma = -0.1$	$\gamma = 0$	$\gamma = 0.1$
50	0.145	0.361	0.789
100	0.147	0.375	0.932
250	0.108	0.373	0.991
500	0.104	0.376	0.999

3.4 Numerical studies

In this section, we investigate the empirical performance of our test. First, we study its finite sample level properties and then compare its empirical power with the tests based on optimal transport (Huang & Sen, 2023), density functions (Diks & Tong, 1999) and projection pursuit technique (Fang, Zhu & Bentler, 1993). Henceforth, we refer to these tests as the OT test, the DT test, and the PP test, respectively. Throughout this chapter, all tests are considered to have a 5% nominal level. The OT test is distribution-free, and the PP test is asymptotically distribution-free. Following the suggestion of the authors, for these two tests we used the cut-offs based on the asymptotic distributions of their test statistics. Our test and the DT test were calibrated using the resampling method, where the cut-off was computed based on 500 iterations. Each experiment was repeated 1000 times to estimate the power of a test by the proportion of times it rejected H_0.

3.4.1 Analysis of simulated data sets

First, we investigate the level property of our test by generating random samples from some spherically symmetric distributions.

Example 3.1. *We consider observations from standard multivariate (a) Gaussian, (b) Cauchy, and (c) $\boldsymbol{t}_4$ (t distribution with 4 degrees of freedom) distributions.*

In each case, we computed the powers for different sample sizes ($n = 20, 40, 60$) and dimensions ($d = 2^i, i = 1, 2, \ldots, 10$). They are reported in Figure 5.1. This figure clearly shows that for all three distributions, our test rejected H_0 in nearly 5% cases. The other three competing tests also exhibited similar behaviour, but to avoid repetition, we do not report them here.

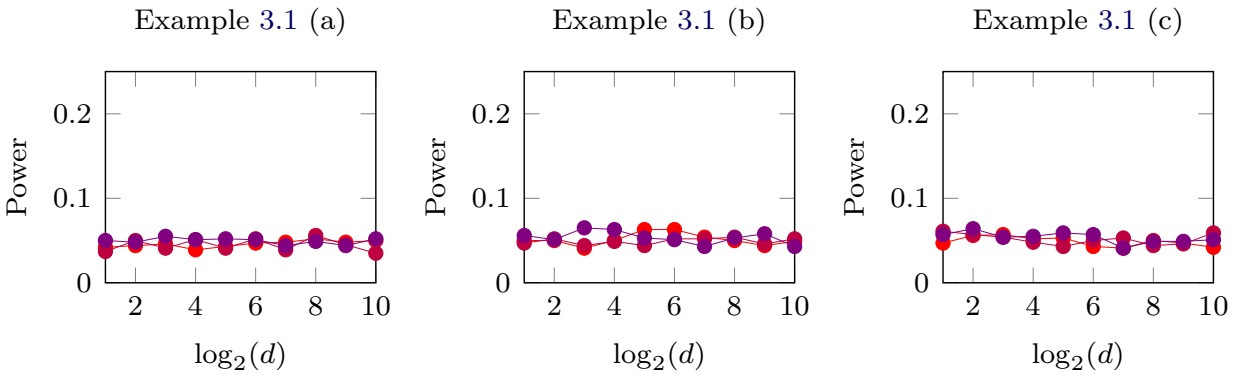


Fig. 3.1 *Observed levels of the proposed test for observations generated from the standard (a) Gaussian, (b) Cauchy, and (c) $\boldsymbol{t}_4$ distributions with sample size $n = 20$(●), $n = 40$ (●) and $n = 60$ (●).*

Next, we consider some non-spherical distributions to compare the powers of different tests.

Example 3.2. *We consider (a) Gaussian, (b) Cauchy, and (c) $\boldsymbol{t}_4$ distributions with the center at the origin and the scatter matrix of the form $(1 - \rho)\, \mathbf{I}_d + \rho\, \mathbf{J}_d$, where $\rho \in (0, 1)$.*

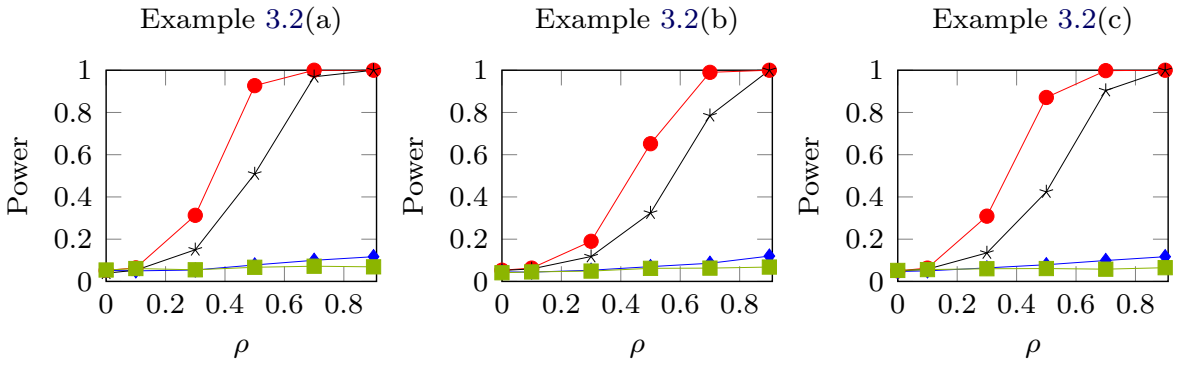


Fig. 3.2 *Powers of the proposed test (●), OT test (♦), DT test (⋆) and PP test (■) in Examples 3.2(a)-(c).*

For Example 3.2, we carried out different tests based on 100 observations when $d = 5$. As ρ increases from 0 to 1 (i.e., the distribution deviates more from sphericity), a test is expected to have increasing power. But, for OT and PP tests, these increments were negligible (see Figure 3.2). In these examples, our test had the best performance followed by the DT test.

Next, we consider four examples involving symmetric but non-elliptic distributions. In Examples 3.3 and 3.4, we deal with ℓ_p-symmetric distributions (see, e.g., Gupta & Song, 1997; Dutta, Ghosh & Chaudhuri, 2011) with $p = \infty$ and $p = 1$, respectively.

Example 3.3. *Let* $\mathbf{X} = R\mathbf{U}$*, where* $R \sim \text{Unif}(9, 10)$ *and* $\mathbf{U}$ *are independent. We take* $\mathbf{U} = \mathbf{Y}/\|\mathbf{Y}\|_\infty$ *where* $\mathbf{Y} = (Y_1, \ldots, Y_5)$*, for the* Y_i*'s* $(i = 1, 2, \ldots, 5)$ *being i.i.d.* $\text{Unif}(-1, 1)$.

Example 3.4. *Let* $\mathbf{X}$ *be as in Example 3.3, but take* $\mathbf{U} = \mathbf{Y}/\|\mathbf{Y}\|_1$*, where* $Y_1, \ldots, Y_5$ *are independent Laplace*$(0, 1)$ *random variables.*

We carried out our experiment for different sample sizes, and the results are reported in Figure 3.3. In these examples, OT and PP tests had powers close to the nominal level of 0.05. In Example 3.3, the proposed test and the DT test had comparable performance, but in Example 3.4, our test significantly outperformed the DT test.

In Example 3.5, we consider an angular symmetric distribution, and in Example 3.6, we deal with a mixture of normal distributions.

Example 3.5. *Let* $\mathbf{X} = R\mathbf{U}$*, where* $\mathbf{U} \sim \text{Unif}(\mathcal{S}^4)$*, but here* R *and* $\mathbf{U}$ *are dependent. For any given* $\mathbf{U} = \mathbf{u}(= (u_1, u_2, \ldots, u_5)^\top)$*, the conditional distribution of* R *is uniform on* $(0, \theta_\mathbf{u})$*, where* $\theta_\mathbf{u} = 10\ \text{I}[u_1u_2 > 0] + 50\ \text{I}[u_1u_2 \le 0; u_3u_4u_5 > 0] + 100\ \text{I}[u_1u_2 \le 0; u_3u_4u_5 \le 0]$.

Example 3.6. *We consider an equal mixture of four normal distributions with the same dispersion matrix* $\mathbf{I}_5$*, but mean vectors* $\mathbf{1}_5$*,*$-\mathbf{1}_5$*,* $\boldsymbol{\beta} = (1, -1, 1, -1, 1)^\top$ *and* $-\boldsymbol{\beta}$*, respectively.*

We carried out our experiment for different sample sizes, and the results are reported in Figure 3.3. In these examples also, our test outperformed its competitors. The DT test had the second-best performance, but its power was much lower compared to our test.

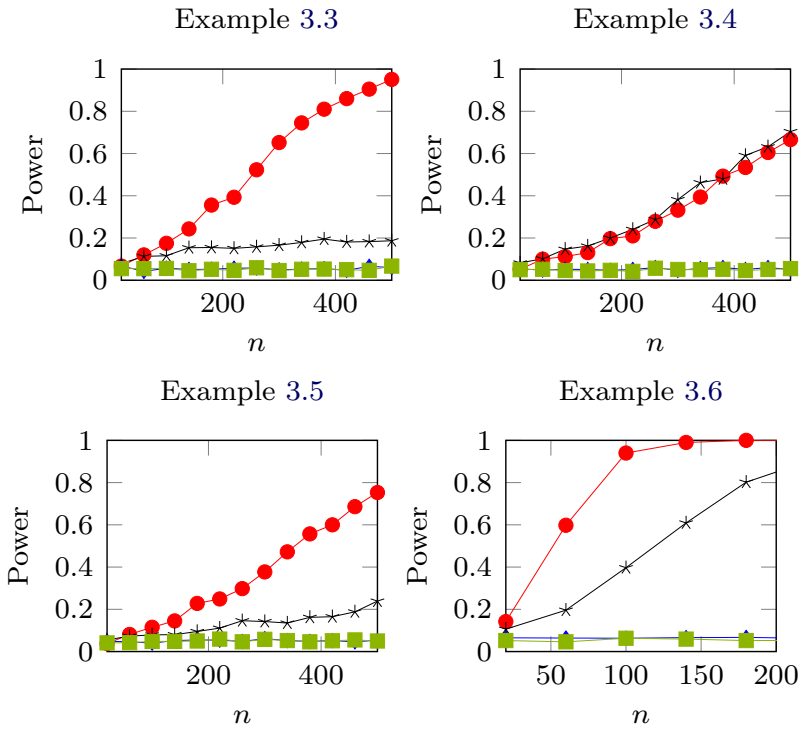


Fig. 3.3 *Powers of the proposed test (●), OT test (♦), DT test (⋆) and PP test (■) in Examples 3.3-3.6.*

Finally, we consider some high-dimensional examples involving normal spiked covariance model (see Johnstone, 2001). In these examples, the DT test based on the usual bandwidth did not work well. So, after consulting with the author, we used the bandwidth $(0.25)^2\hat{\sigma}_0^2$ (where $\hat{\sigma}_0^2 = \frac{1}{n(n-1)}\sum_{1\le i<j\le n}\|\mathbf{X}_i - \mathbf{X}_j\|^2$) which makes an adjustment for the scale in higher dimensions.

Example 3.7. *We consider a normal spiked covariance model with mean zero and a diagonal covariance matrix with entries (a)* $(d, 1, 1, \ldots, 1)$ *and (b)* $(d^{0.5}, 1, 1, \ldots, 1)$*, respectively. In these examples, we consider* $n = 20$ *and different values of* d*, while in (c) we use the same model as (b), but increase the sample size* $n = 20 + [d^{1.5}]$ *with the dimension* d*.*

Figure 3.4 shows that in Example 3.7 (a), our test and the DT test performed well, while the other tests had a non-satisfactory performance. Among these two tests, our test had an edge in higher dimensions. But, in Example 3.7 (b), all tests, including ours, performed poorly. Note that in Example 3.7(a) and 3.7(b), the measure of sphericity (see, e.g., John, 1972; Jung & Marron, 2009) converges to 0 and 1, respectively, as d increases. So, in higher dimension, the data cloud in Example 3.7 (b) turns out to be similar to that from a spherical distribution, whereas in Example 3.7 (a), it has significant deviations from sphericity. This explains the reason behind the diametrically opposite behaviour of our test in these two examples. However, Corollary 3.10 suggests that even in Example 3.7 (b), our test can perform well if we increase the sample size along with the dimension at a suitable rate. We observed this in Example 3.7 (c), where the power curve of our test had a

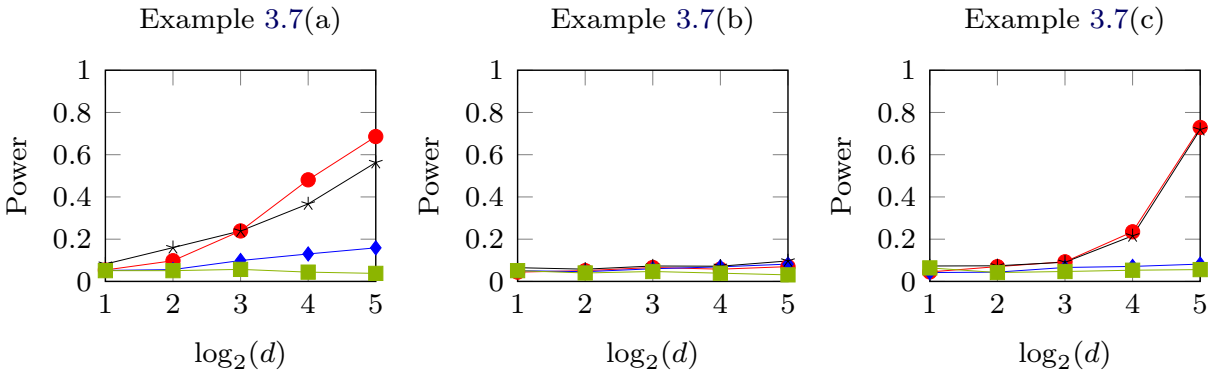


Fig. 3.4 *Powers of the proposed test (●), OT test (♦), DT test (⋆) and PP test (■) in Examples 3.7(a)-(c).*

sharp increasing trend. The DT test also showed a similar pattern. But the other competing tests had poor performance even in this set-up.

3.4.2 Analysis of a benchmark dataset

For further evaluation of different tests, we analyze the "MAGIC Gamma Telescope" data set available at the UCI machine learning repository. This data set was generated by a Monte Carlo program called CORSIKA described in Heck et al. (1998). It was used to simulate the registration of high-energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using imaging techniques. The observations are classified based on the patterns in the images the particles generate, called the shower images. Based on these shower images, the particles are classified as "primary gamma" and "hadronic shower". In our analysis, we first divided the entire data set into two parts based on the class labels "primary gamma" and "hadronic shower" and called them "MAGIC-1" and "MAGIC-2", respectively. Here the observations are 10-dimensional (see Heck et al., 1998for details). While MAGIC-1 contains 12332 observations, there are 6688 observations in MAGIC-2. When we used the full data sets (after centering by subtracting the corresponding spatial medians) for testing, all four tests rejected the null hypothesis of spherical symmetry in both

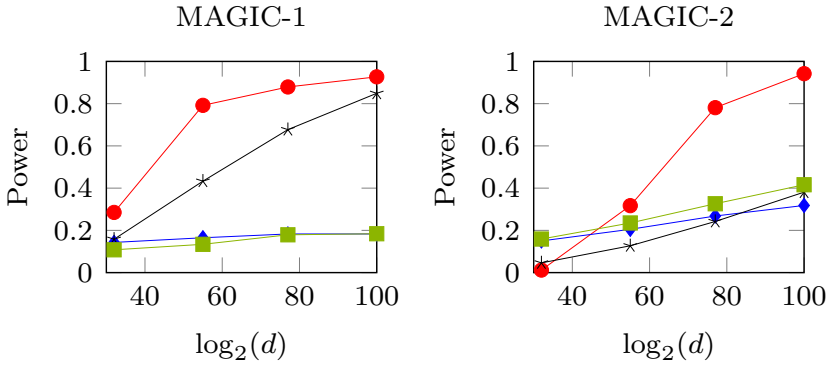


Fig. 3.5 *Powers of the proposed test (●), OT test (♦), DT test (⋆) and PP test (■) in the Magic Gamma Telescope data set.*

cases. This gives an indication that the underlying distributions are non-spherical, and different can be compared based on their powers. However, it was difficult to compare among different tests using a single experiment based on the whole data set. Therefore, to compare the performance of different tests, we generated random sub-samples from these two data sets, and in each case, we repeated the procedure 1000 times. The power of a test was computed by the proportion of times it rejected the null hypothesis. The results for different sub-sample sizes are reported in Figure 3.5.

In the MAGIC-1 data set, our test significantly outperformed all other tests for all sample sizes considered here. For the MAGIC-2 data set, the OT and PP tests had better performance for a lower sample size. However, the power of our test increased sharply with the sample size, while those of other tests increased at a slower rate. Figure 3.5 clearly shows that our test outperformed all its competitors for sample sizes larger than 55.

3.5 Proofs and mathematical details

Proof of Proposition 3.1. If φ_1 and φ_2 are characteristic functions of $\mathbf{X} \sim \mathrm{P}$ and $\mathbf{X}'$, we have

$$\zeta(\mathrm{P}) = \int \left|\varphi_1(\mathbf{t}) - \varphi_2(\mathbf{t})\right|^2 \mathrm{d}W(\mathbf{t}).$$

Since $W(\cdot)$ is non-negative, the fact $\zeta(\mathrm{P}) \geq 0$ follows from the non-negativity of the integrand. If P is spherically symmetric, then $\varphi_1 = \varphi_2$ and hence $\zeta(\mathrm{P}) = 0$. Conversely, $\zeta(\mathrm{P}) = 0$ implies $\varphi_1(\mathbf{t}) = \varphi_2(\mathbf{t})$ over the set $\{\mathbf{t} \mid W(\mathbf{t}) > 0\}$. Since $W(\cdot)$ is equivalent to the Lebesgue measure, this implies $\varphi_1(\mathbf{t}) = \varphi_2(\mathbf{t})$ almost everywhere with respect to the Lebesgue measure. Again, both φ_1 and φ_2 continuous functions. So, $\zeta(\mathrm{P}) = 0$ implies $\varphi_1(\mathbf{t}) = \varphi_2(\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^d$ and hence $\mathbf{X}$ and $\|\mathbf{X}\|\mathbf{U}$ are identically distributed, i.e., P is spherically symmetric. ■

Proof of Lemma 3.2. Note that

$$\int \exp\left\{i\langle \mathbf{T}, \mathbf{X}_1 - \mathbf{X}_2\rangle\right\} \frac{d^{d/2}}{(2\pi)^{d/2}} e^{-d\|\mathbf{T}\|^2/2} \mathrm{d}\mathbf{T} = \mathbb{E}\Big\{ \exp\left\{i\langle \mathbf{T}, \mathbf{X}_1 - \mathbf{X}_2\rangle\right\}\Big\}, \tag{3.4}$$

where $\mathbf{T} \sim \mathcal{N}_d(\mathbf{0}_d, d^{-1}\mathbf{I}_d)$. The right side of equation (3.4) is the characteristic function of $\mathbf{T}$ evaluated at $\mathbf{X}_1 - \mathbf{X}_2$. Hence, we have the desired result. ■

Proof of Theorem 3.1. First note that $|\varphi_1(\mathbf{t}) - \varphi_2(\mathbf{t})|^2 = \big(\varphi_1(\mathbf{t}) - \varphi_2(\mathbf{t})\big)\big(\varphi_1(-\mathbf{t}) - \varphi_2(-\mathbf{t})\big)$. So, expanding the characteristic functions in terms of expectations, we get

$$\begin{aligned} |\varphi_1(\mathbf{t}) - \varphi_2(\mathbf{t})|^2 =& \mathbb{E}\Big\{ \exp\left\{i\langle \mathbf{t}, \mathbf{X}_1 - \mathbf{X}_2\rangle\right\}\Big\} + \mathbb{E}\Big\{ \exp\left\{i\langle \mathbf{t}, \mathbf{X}_1' - \mathbf{X}_2'\rangle\right\}\Big\} \\ & - \mathbb{E}\Big\{ \exp\left\{i\langle \mathbf{t}, \mathbf{X}_1 - \mathbf{X}_2'\rangle\right\}\Big\} - \mathbb{E}\Big\{ \exp\left\{i\langle \mathbf{t}, \mathbf{X}_1' - \mathbf{X}_2\rangle\right\}\Big\}. \end{aligned} \tag{3.5}$$

From Lemma 3.2, for any $\mathbf{V}_1$ and $\mathbf{V}_2$, we have

$$\int \mathbb{E}\Big\{ \exp\left\{i\langle \mathbf{t}, \mathbf{V}_1 - \mathbf{V}_2\rangle\right\}\Big\} \frac{d^{d/2}}{(2\pi)^{d/2}} e^{-d\|\mathbf{t}\|^2/2} \mathrm{d}\mathbf{t} = \mathbb{E}\Big\{ \exp\Big\{ -\frac{1}{2d}\|\mathbf{V}_1 - \mathbf{V}_2\|^2\Big\}\Big\}.$$

Applying this to all four terms in (3.5) (note that the last two terms are equal), we get the result. ■

Proof of Theorem 3.2. As introduced in Section 3.1, we can write our estimator as

$$\hat{\zeta}_n = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} g\big((\mathbf{X}_i, \mathbf{X}_i'), (\mathbf{X}_j, \mathbf{X}_j')\big),$$

$$\text{where } g\big((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{y}_1, \mathbf{y}_2)\big) = \exp\Big\{ - \frac{\|\mathbf{x}_1 - \mathbf{y}_1\|^2}{2d} \Big\} + \exp\Big\{ - \frac{\|\mathbf{x}_2 - \mathbf{y}_2\|^2}{2d} \Big\} - \exp\Big\{ - \frac{\|\mathbf{x}_1 - \mathbf{y}_2\|^2}{2d} \Big\} - \exp\Big\{ - \frac{\|\mathbf{x}_2 - \mathbf{y}_1\|^2}{2d} \Big\}.$$

Now let $\hat{\zeta}_n^{(i)}$ denote our estimator when the i^{th} observation $(\mathbf{X}_i, \mathbf{X}_i')$ is replaced by an independent copy $(\mathbf{Y}_i, \mathbf{Y}_i')$ of the same. Note that

$$|\hat{\zeta}_n - \hat{\zeta}_n^{(i)}| \le \frac{2}{n(n-1)} \Big\{ \sum_{j=1}^{i-1} \Big| g\big((\mathbf{X}_j, \mathbf{X}_j'), (\mathbf{X}_i, \mathbf{X}_i')\big) - g\big((\mathbf{X}_j, \mathbf{X}_j'), (\mathbf{Y}_i, \mathbf{Y}_i')\big) \Big| + \sum_{j=i+1}^{n} \Big| g\big((\mathbf{X}_i, \mathbf{X}_i'), (\mathbf{X}_j, \mathbf{X}_j')\big) - g\big((\mathbf{Y}_i, \mathbf{Y}_i'), (\mathbf{X}_j, \mathbf{X}_j')\big) \Big| \Big\}$$

Since $|g(\cdot,\cdot)| \le 2$, this implies $|\hat{\zeta}_n - \hat{\zeta}_n^{(i)}| \le \frac{8(n-1)}{n(n-1)} \le \frac{8}{n}$. So, applying bounded difference inequality (see page 37 in Wainwright, 2019), we get

$$\mathbb{P}\{\big|\hat{\zeta}_n - \zeta(\mathrm{P})\big| > \epsilon\} \le \exp\Big\{ - \frac{2\epsilon^2}{\sum_{i=1}^n \frac{64}{n^2}} \Big\} = \exp\Big\{ - \frac{n\epsilon^2}{32} \Big\}. \qquad \blacksquare$$

Proof of Theorem 3.3. Note that our estimator $\hat{\zeta}_n$ is a U-statistic with the kernel

$$g\big((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{y}_1, \mathbf{y}_2)\big) = \exp\Big\{ - \frac{\|\mathbf{x}_1 - \mathbf{y}_1\|^2}{2d} \Big\} + \exp\Big\{ - \frac{\|\mathbf{x}_2 - \mathbf{y}_2\|^2}{2d} \Big\} - \exp\Big\{ - \frac{\|\mathbf{x}_1 - \mathbf{y}_2\|^2}{2d} \Big\} - \exp\Big\{ - \frac{\|\mathbf{x}_2 - \mathbf{y}_1\|^2}{2d} \Big\}$$

of degree 2. The first order Hoeffding projection of $g(\cdot,\cdot)$ is

$$g_1\big((\mathbf{x}_1, \mathbf{x}_2)\big) = \mathbb{E}\Big\{K(\mathbf{x}_1, \mathbf{X}_1)\Big\} + \mathbb{E}\Big\{K(\mathbf{x}_2, \mathbf{X}_1')\Big\} - \mathbb{E}\Big\{K(\mathbf{x}_1, \mathbf{X}_1')\Big\} - \mathbb{E}\Big\{K(\mathbf{x}_2, \mathbf{X}_1)\Big\}, \tag{3.6}$$

where $K(\mathbf{x}, \mathbf{y}) = \exp\{-\|\mathbf{x} - \mathbf{y}\|^2/(2d)\}$, $\mathbf{X}_1 \sim \mathrm{P}$ and $\mathbf{X}_1' = \|\mathbf{X}_1\|\mathbf{U}_1$ for $\mathbf{U}_1 \sim \mathrm{Unif}(\mathcal{S}^{d-1})$ independent of $\mathbf{X}_1$. Under H_0, $\mathbf{X}_1$ and $\mathbf{X}_1'$ are identically distributed and hence $g_1\big((\mathbf{x}_1, \mathbf{x}_2)\big) = 0$. Therefore, using Theorem 1 from Lee (1990) page 79, we get that $n\hat{\zeta}_n$ converges in distribution to $\sum_{i=1}^{\infty} \lambda_i(Z_i^2 - 1)$, where the Z_i's are independent standard normal random variables and λ_i's are the eigenvalues of the integral equation

$$\int g\big((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{y}_1, \mathbf{y}_2)\big) f\big((\mathbf{y}_1, \mathbf{y}_2)\big) \mathrm{d}F\big((\mathbf{y}_1, \mathbf{y}_2)\big) = \lambda f\big((\mathbf{x}_1, \mathbf{x}_2)\big),$$

for F being the joint distribution of $(\mathbf{X}_1, \mathbf{X}_1')$. $\blacksquare$

Proof of Theorem 3.4. Under H_1, the function $g_1\big((\mathbf{x}_1, \mathbf{x}_2)\big)$ defined in equation (3.6) is non-degenerate. Therefore, using Theorem 1 from Lee (1990) page 76, we get the asymptotic normality of $\sqrt{n}(\hat{\zeta}_n - \zeta(\mathrm{P}))$ with the mean zero and the variance $4\sigma_1^2$, where $\sigma_1^2 = \mathrm{Var}\big(g_1\big((\mathbf{X}_1, \mathbf{X}_1')\big)\big)$. $\blacksquare$

Proof of Lemma 3.3. Define $\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_n$ as in Theorem 3.3. Under H_0, $(\mathbf{X}_i, \mathbf{X}'_i)$ and $(\mathbf{X}'_i, \mathbf{X}_i)$ are identically distributed for each $i = 1, 2, \ldots, n$. Therefore, the joint distribution of $(\hat{\zeta}_n, c_{1-\alpha})$ is identical to the joint distribution of $(\hat{\zeta}_n(\boldsymbol{S}), c_{1-\alpha})$, where $\boldsymbol{S}$ is an element of $\{0,1\}^n$. Let $\mathcal{D}_A = \{(\mathbf{X}_i, \mathbf{X}'_i) \mid i = 1, 2, \ldots, n\}$ denote the augmented data. Then

$$\mathbb{P}\{\hat{\zeta}_n > c_{1-\alpha}\} = \mathbb{P}\{\hat{\zeta}_n(\boldsymbol{S}) > c_{1-\alpha}\} = \mathbb{E}\big\{\mathbb{P}\{\hat{\zeta}_n(\boldsymbol{S}) > c_{1-\alpha} \mid \mathcal{D}_A\}\big\} \leq \alpha.$$

The last inequality follows from the definition of $c_{1-\alpha}$ (given in page 42). ■

Proof of Lemma 3.4. Here $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent copies of $\mathbf{X} \sim \mathrm{P}$ and $\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_n$ are as in Theorem 3.3. Define $\mathcal{D}_A$ as in the proof of Lemma 3.3. Let $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$ be uniformly distributed on the set $\{0,1\}^n$. For $i = 1, 2, \ldots, n$, define $\mathbf{Y}_i = S_i\mathbf{X}_i + (1 - S_i)\mathbf{X}'_i$ and $\mathbf{Y}'_i = (1 - S_i)\mathbf{X}_i + S_i\mathbf{X}'_i$. Since $g\big((\mathbf{Y}_i, \mathbf{Y}'_i), (\mathbf{Y}_i, \mathbf{Y}'_i)\big) \geq 0$ for all $i = 1, 2, \ldots, n$, we have

$$n(n-1)\hat{\zeta}_n(\boldsymbol{S}) = \sum_{1 \leq i \neq j \leq n} g\big((\mathbf{Y}_i, \mathbf{Y}'_i), (\mathbf{Y}_j, \mathbf{Y}'_j)\big) \leq \sum_{1 \leq i,j \leq n} g\big((\mathbf{Y}_i, \mathbf{Y}'_i), (\mathbf{Y}_j, \mathbf{Y}'_j)\big) = n^2\zeta(F_n),$$

where F_n denotes the empirical probability distribution of $(\mathbf{Y}_1, \mathbf{Y}'_1), (\mathbf{Y}_2, \mathbf{Y}'_2), \ldots, (\mathbf{Y}_n, \mathbf{Y}'_n)$. Since $\zeta(F_n)$ is a non-negative random variable, using Markov inequality, for any $\epsilon > 0$, we have

$$\mathbb{P}\big\{\hat{\zeta}_n(\boldsymbol{S}) > \epsilon \mid \mathcal{D}_A\big\} \leq \mathbb{P}\big\{n^2\zeta(F_n) > n(n-1)\epsilon \mid \mathcal{D}_A\big\} \leq \frac{n^2}{n(n-1)\epsilon}\mathbb{E}\big\{\zeta(F_n) \mid \mathcal{D}_A\big\}.$$

Now, taking $\epsilon = \frac{n}{(n-1)\alpha}\mathbb{E}\big\{\zeta(F_n) \mid \mathcal{D}_A\big\}$, we get $\mathbb{P}\big\{\hat{\zeta}_n(\boldsymbol{S}) > \epsilon \mid \mathcal{D}_A\big\} \leq \alpha$. So, from the definition of $c_{1-\alpha}$ (see the resampling algorithm in Section 3.2, page 42), we have $c_{1-\alpha} \leq \frac{n}{(n-1)\alpha}\mathbb{E}\big\{\zeta(F_n) \mid \mathcal{D}_A\big\}$. Also, note that

$$\mathbb{E}\{\zeta(F_n) \mid \mathcal{D}_A\} = \frac{n(n-1)}{n^2}\mathbb{E}\Big\{g\big((\mathbf{Y}_1, \mathbf{Y}'_1), (\mathbf{Y}_2, \mathbf{Y}'_2)\big) \mid \mathcal{D}_A\Big\} + \frac{1}{n}\mathbb{E}\Big\{g\big((\mathbf{Y}_1, \mathbf{Y}'_1), (\mathbf{Y}_1, \mathbf{Y}'_1)\big) \mid \mathcal{D}_A\Big\}.$$

$$\begin{aligned}\text{Now } \mathbb{E}\Big\{ \exp\{-\frac{\|\mathbf{Y}_1 - \mathbf{Y}_2\|^2}{2d}\} \mid \mathcal{D}_A\Big\} \\ = \frac{1}{2^n}\sum_{\boldsymbol{S} \in \{0,1\}^n} \exp\{-\frac{\|S_1\mathbf{X}_1 + (1-S_1)\mathbf{X}'_1 - S_2\mathbf{X}_2 + (1-S_2)\mathbf{X}'_2\|^2}{2d}\} = \nabla_n \text{ (say).}\end{aligned}$$

Similarly, one can also show that

$$\begin{aligned}\mathbb{E}\Big\{ \exp\{-\frac{\|\mathbf{Y}'_1 - \mathbf{Y}'_2\|^2}{2d}\} \mid \mathcal{D}_A\Big\} &= \mathbb{E}\Big\{ \exp\{-\frac{\|\mathbf{Y}_1 - \mathbf{Y}'_2\|^2}{2d}\} \mid \mathcal{D}_A\Big\} \\ &= \mathbb{E}\Big\{ \exp\{-\frac{\|\mathbf{Y}'_1 - \mathbf{Y}_2\|^2}{2d}\} \mid \mathcal{D}_A\Big\} = \nabla_n.\end{aligned}$$

Hence, $\mathbb{E}\Big\{g\big((\mathbf{Y}_1, \mathbf{Y}'_1), (\mathbf{Y}_2, \mathbf{Y}'_2)\big) \mid \mathcal{D}_A\Big\} = \nabla_n + \nabla_n - \nabla_n - \nabla_n = 0$. Similarly, one can also show that

$$\mathbb{E}\Big\{g\big((\mathbf{Y}_1, \mathbf{Y}'_1), (\mathbf{Y}_1, \mathbf{Y}'_1)\big) \mid \mathcal{D}_A\Big\} = 2\Big(1 - \nabla_n\Big) \leq 2$$

Combining these, we get $c_{1-\alpha} \leq 2(\alpha(n-1))^{-1}$. This completes the proof. ■

Proof of Theorem 3.5. Note that $\hat{\zeta}_n$ is a consistent estimator of $\zeta(\mathrm{P})$ (follows from Theorem 3.2), where $\zeta(\mathrm{P}) = 0$ under H_0 and positive under H_1 (follows from Proposition 3.1). So, by Lemma 3.4, under H_1, the power of the test $\mathbb{P}(\hat{\zeta}_n > c_{1-\alpha})$ converges to one as n diverges to infinity. ■

Proof of Theorem 3.6. Let us define the distribution functions $M_B(t) = \frac{1}{B}\left\{\sum_{i=1}^{B} I[\hat{\zeta}_n(\boldsymbol{S}_i) \le t]\right\}$ and $M(t) = \frac{1}{2^n}\left\{\sum_{\boldsymbol{S}\in\{0,1\}^n} I[\hat{\zeta}_n(\boldsymbol{S}) \le t]\right\}$ conditioned on the augmented data $\mathcal{D}_A$. Then

$$\begin{aligned}
|p_n - p_{n,B}| &= \left|\frac{1}{2^n}\left\{\sum_{\boldsymbol{S}\in\{0,1\}^n} I[\hat{\zeta}_n(\boldsymbol{S}) \ge \hat{\zeta}_n]\right\} - \frac{1}{B+1}\left\{\sum_{i=1}^{B} I[\hat{\zeta}_n(\boldsymbol{S}_i) \ge \hat{\zeta}_n] + 1\right\}\right| \\
&= \left|\frac{1}{2^n}\left\{\sum_{\boldsymbol{S}\in\{0,1\}^n} I[\hat{\zeta}_n(\boldsymbol{S}) < \hat{\zeta}_n]\right\} - \frac{1}{B+1}\left\{\sum_{i=1}^{B} I[\hat{\zeta}_n(\boldsymbol{S}_i) < \hat{\zeta}_n]\right\}\right| \\
&= \left|M(\hat{\zeta}_n-) - \frac{B}{B+1}M_B(\hat{\zeta}_n-)\right| \\
&\le \left|M(\hat{\zeta}_n-) - M_B(\hat{\zeta}_n-)\right| + \left|\frac{M_B(\hat{\zeta}_n-)}{B+1}\right| \le \sup_{t\in\mathbb{R}}\left|M(t) - M_B(t)\right| + \frac{1}{B+1}.
\end{aligned}$$

Conditioned on $\mathcal{D}_A$, the Dvoretzky-Keifer-Wolfwitz inequality (Massart (1990)) gives us $\mathbb{P}\{\sup_{t\in\mathbb{R}}|M(t) - M_B(t)| > \epsilon\} \le 2e^{-2B\epsilon^2}$. Hence, conditioned on $\mathcal{D}_A$, as B grows to infinity the randomized p-value $p_{n,B}$ converges almost surely to p_n. ■

Proof of Lemma 3.5. Recall that $\zeta(\mathrm{P})$ can be expressed as

$$\zeta(\mathrm{P}) = \mathbb{E}\{K(\mathbf{X}_1, \mathbf{X}_2)\} + \mathbb{E}\{K(\mathbf{X}_1', \mathbf{X}_2')\} - 2\mathbb{E}\{K(\mathbf{X}_1, \mathbf{X}_2')\},$$

where $\mathbf{X}_1, \mathbf{X}_2$ are independent copies of $\mathbf{X}\sim\mathrm{P}$, $\mathbf{X}_1' = \|\mathbf{X}_1\|\mathbf{U}_1, \mathbf{X}_2' = \|\mathbf{X}_2\|\mathbf{U}_2$ with $\mathbf{U}_1, \mathbf{U}_2 \overset{iid}{\sim} \mathrm{Unif}(\mathcal{S}^{d-1})$ independent of $\mathbf{X}_1, \mathbf{X}_2$ and $K(\mathbf{x}, \mathbf{y}) = \exp\left\{-\frac{1}{2d}\|\mathbf{x}-\mathbf{y}\|^2\right\}$. Now if $\mathbb{P} = (1-\delta)F + \delta G$ $(0 < \delta < 1)$, then

$$\begin{aligned}
\mathbb{E}\{K(\mathbf{X}_1, \mathbf{X}_2)\} &= \int K(\mathbf{x}_1, \mathbf{x}_2)\mathrm{d}\mathbb{P}(\mathbf{x}_1)\mathrm{d}\mathbb{P}(\mathbf{x}_2) \\
&= (1-\delta)^2\int K(\mathbf{x}_1, \mathbf{x}_2)\mathrm{d}F(\mathbf{x}_1)\mathrm{d}F(\mathbf{x}_2) + 2\delta(1-\delta)\int K(\mathbf{x}_1, \mathbf{x}_2)\mathrm{d}G(\mathbf{x}_1)\mathrm{d}F(\mathbf{x}_2) \\
&\quad + \delta^2\int K(\mathbf{x}_1, \mathbf{x}_2)\mathrm{d}G(\mathbf{x}_1)\mathrm{d}G(\mathbf{x}_2).
\end{aligned}$$

If μ_0 denotes the uniform distribution on $(\mathcal{S}^{d-1})$, then

$$\begin{aligned}
\mathbb{E}\{K(\mathbf{X}_1', \mathbf{X}_2')\} &= \int K(\|\mathbf{x}_1\|\mathbf{u}_1, \|\mathbf{x}_2\|\mathbf{u}_2)\mathrm{d}\mathbb{P}(\mathbf{x}_1)\mathrm{d}\mathbb{P}(\mathbf{x}_2)\mathrm{d}\mu_0(\mathbf{u}_1)\mathrm{d}\mu_0(\mathbf{u}_2) \\
&= (1-\delta)^2\int K(\|\mathbf{x}_1\|\mathbf{u}_1, \|\mathbf{x}_2\|\mathbf{u}_2)\mathrm{d}F(\mathbf{x}_1)\mathrm{d}F(\mathbf{x}_2)\mathrm{d}\mu_0(\mathbf{u}_1)\mathrm{d}\mu_0(\mathbf{u}_2) \\
&\quad + \delta^2\int K(\|\mathbf{x}_1\|\mathbf{u}_1, \|\mathbf{x}_2\|\mathbf{u}_2)\mathrm{d}G(\mathbf{x}_1)\mathrm{d}G(\mathbf{x}_2)\mathrm{d}\mu_0(\mathbf{u}_1)\mathrm{d}\mu_0(\mathbf{u}_2) \\
&\quad + 2\delta(1-\delta)\int K(\|\mathbf{x}_1\|\mathbf{u}_1, \|\mathbf{x}_2\|\mathbf{u}_2)\mathrm{d}G(\mathbf{x}_1)\mathrm{d}F(\mathbf{x}_2)\mathrm{d}\mu_0(\mathbf{u}_1)\mathrm{d}\mu_0(\mathbf{u}_2)
\end{aligned}$$

$$\begin{aligned}\text{and } \mathbb{E}\{K(\mathbf{X}_1,\mathbf{X}_2')\} &= \int K(\mathbf{x}_1,\|\mathbf{x}_2\|\mathbf{u}_2)\mathrm{d}\mathbb{P}(\mathbf{x}_1)\mathrm{d}\mathbb{P}(\mathbf{x}_2)\mathrm{d}\mu_0(\mathbf{u}_2)\\ &= (1-\delta)^2\int K(\mathbf{x}_1,\|\mathbf{x}_2\|\mathbf{u}_2)\mathrm{d}F(\mathbf{x}_1)\mathrm{d}F(\mathbf{x}_2)\mathrm{d}\mu_0(\mathbf{u}_2)\\ &+ \delta(1-\delta)\int K(\mathbf{x}_1,\|\mathbf{x}_2\|\mathbf{u}_2)\ [dG(\mathbf{x}_1)\mathrm{d}F(\mathbf{x}_2)\mathrm{d}\mu_0(\mathbf{u}_2)+dF(\mathbf{x}_1)\mathrm{d}G(\mathbf{x}_2)\mathrm{d}\mu_0(\mathbf{u}_2)]\\ &+ \delta^2\int K(\mathbf{x}_1,\|\mathbf{x}_2\|\mathbf{u}_2)\mathrm{d}G(\mathbf{x}_1)\mathrm{d}G(\mathbf{x}_2)\mathrm{d}\mu_0(\mathbf{u}_2).\end{aligned}$$

Therefore, if $\mathrm{P}=(1-\delta)F+\delta G$, then we have

$$\zeta(\mathrm{P})=(1-\delta)^2\zeta(F)+\delta^2\zeta(G)+2\delta(1-\delta)\zeta'(G,F),$$

where $\zeta'(G,F)=\mathbb{E}\{K(\mathbf{Y}_1,\mathbf{Y}_2)\}+\mathbb{E}\{K(\mathbf{Y}_1',\mathbf{Y}_2')\}-\mathbb{E}\{K(\mathbf{Y}_1,\mathbf{Y}_2')\}-\mathbb{E}\{K(\mathbf{Y}_1',\mathbf{Y}_2)\}$. Here $\mathbf{Y}_1\sim G$ and $\mathbf{Y}_2\sim F$ are independent, $\mathbf{Y}_1'=\|\mathbf{Y}_1\|\mathbf{U}_1$ and $\mathbf{Y}_2'=\|\mathbf{Y}_2\|\mathbf{U}_2$ where $\mathbf{U}_1,\mathbf{U}_2\overset{iid}{\sim}\mathrm{Unif}(\mathcal{S}^{d-1})$ independent of $\mathbf{Y}_1$ and $\mathbf{Y}_2$. This completes the proof. ■

Lemma A3.1. *If* $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ *are independent copies of* $\mathbf{X}\sim\mathrm{P}$*, then*

$$Var(\hat{\zeta}_n)\le\binom{n}{2}^{-1}\left[4(n-2)\zeta(\mathrm{P})+4\right].$$

Proof. Recall that $\hat{\zeta}_n$ is a U-statistic with the kernel

$$g\big((\mathbf{x}_1,\mathbf{x}_1'),(\mathbf{x}_2,\mathbf{x}_2')\big)=K(\mathbf{x}_1,\mathbf{x}_2)+K(\mathbf{x}_1',\mathbf{x}_2')-K(\mathbf{x}_1,\mathbf{x}_2')-K(\mathbf{x}_1',\mathbf{x}_2),$$

where $K(\mathbf{x},\mathbf{y})=\exp\{-\frac{1}{2d}\|\mathbf{x}-\mathbf{y}\|^2\}$. Then by the theory of U-statistics (see page 12 in Lee, 1990),

$$\mathrm{Var}(\hat{\zeta}_n)=\binom{n}{2}^{-1}\left[\binom{2}{1}\binom{n-2}{1}\mathrm{Var}\Big(g_1\big(\mathbf{X}_1,\mathbf{X}_1'\big)\Big)+\binom{2}{2}\binom{n-2}{0}\mathrm{Var}\Big(g\big((\mathbf{X}_1,\mathbf{X}_1'),(\mathbf{X}_2,\mathbf{X}_2')\big)\Big)\right],$$

$$\begin{aligned}\text{where } g_1\big((\mathbf{x}_1,\mathbf{x}_1')\big) &= \mathbb{E}\big\{\exp\big(-\frac{1}{2d}\|\mathbf{x}_1-\mathbf{X}_2\|^2\big)\big\}-\mathbb{E}\big\{\exp\big(-\frac{1}{2d}\|\mathbf{x}_1-\mathbf{X}_2'\|^2\big)\big\}\\ &\quad+\mathbb{E}\big\{\exp\big(-\frac{1}{2d}\|\mathbf{x}_1'-\mathbf{X}_2'\|^2\big)\big\}-\mathbb{E}\big\{\exp\big(-\frac{1}{2d}\|\mathbf{x}_1'-\mathbf{X}_2\|^2\big)\big\}.\end{aligned}$$

Since, $|g(\cdot,\cdot)|\le 2$, $\mathrm{Var}\Big(g\big((\mathbf{X}_1,\mathbf{X}_1'),(\mathbf{X}_2,\mathbf{X}_2')\big)\Big)$ is bounded by 4. Now, to find a bound for the first term, note that

$$\begin{aligned}g_1\big((\mathbf{x}_1,\mathbf{x}_1')\big) &= \mathbb{E}\{\int\exp(i\langle\mathbf{t},\mathbf{x}_1-\mathbf{X}_2\rangle)\phi_0(\mathbf{t})\mathrm{d}\mathbf{t}\}-\mathbb{E}\{\int\exp(i\langle\mathbf{t},\mathbf{x}_1-\mathbf{X}_2'\rangle)\phi_0(\mathbf{t})\mathrm{d}\mathbf{t}\}\\ &\quad-\mathbb{E}\{\int\exp(i\langle\mathbf{t},\mathbf{x}_1'-\mathbf{X}_2\rangle)\phi_0(\mathbf{t})\mathrm{d}\mathbf{t}\}+\mathbb{E}\{\int\exp(i\langle\mathbf{t},\mathbf{x}_1'-\mathbf{X}_2'\rangle)\phi_0(\mathbf{t})\mathrm{d}\mathbf{t}\}\\ &\qquad\qquad\text{(where }\phi_0(\mathbf{t})\text{ denotes the density of }\mathcal{N}_d(\mathbf{0},d^{-1}\mathbf{I}_d))\\ &= \int\big(\exp(i\langle\mathbf{t},\mathbf{x}_1\rangle)-\exp(i\langle\mathbf{t},\mathbf{x}_1'\rangle)\big)\big(\mathbb{E}\{\exp(-i\langle\mathbf{t},\mathbf{X}_2\rangle)\}-\mathbb{E}\{\exp(-i\langle\mathbf{t},\mathbf{X}_2'\rangle)\}\big)\phi_0(\mathbf{t})\mathrm{d}\mathbf{t}\\ &= \int(\exp(i\langle\mathbf{t},x_1\rangle)-\exp(i\langle\mathbf{t},\mathbf{x}_1'\rangle))(\varphi_1(-\mathbf{t})-\varphi_2(-\mathbf{t}))\phi_0(\mathbf{t})\mathrm{d}\mathbf{t}\\ &= \int(\exp(i\langle\mathbf{t},\mathbf{x}_1\rangle)-\exp(i\langle\mathbf{t},\mathbf{x}_1'\rangle))\overline{(\varphi_1(\mathbf{t})-\varphi_2(\mathbf{t}))}\phi_0(\mathbf{t})\mathrm{d}\mathbf{t}.\end{aligned}$$

Then, using Cauchy-Schwartz inequality, we have

$$g_1^2\big((\mathbf{x}_1,\mathbf{x}_1')\big) \le \int |\exp(i\langle \mathbf{t},\mathbf{x}_1\rangle) - \exp(i\langle \mathbf{t},\mathbf{x}_1'\rangle)|^2 \phi_0(\mathbf{t})\mathrm{d}\mathbf{t} \times \int |\varphi_1(\mathbf{t}) - \varphi_2(\mathbf{t})|^2 \phi_0(\mathbf{t})\mathrm{d}\mathbf{t} \le 2\zeta(\mathrm{P}).$$

This gives us

$$\mathrm{Var}(\hat{\zeta}_n) \le \binom{n}{2}^{-1} [4(n-2)\zeta(\mathrm{P}) + 4]\,. \qquad \blacksquare$$

Proof of Theorem 3.7. Here $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent copies of $\mathbf{X} \sim F_\delta = (1-\delta)F + \delta G$, where $\zeta(G) = 0$ and $\zeta(F) = \gamma_0 > 0$. So, we have $\zeta(F_\delta) = (1-\delta)^2\gamma_0$ (follows from Lemma 3.5) and

$$\mathbb{P}\{\hat{\zeta}_n > c_{1-\alpha}\} = 1 - \mathbb{P}\{\hat{\zeta}_n \le c_{1-\alpha}\} \ge 1 - \mathbb{P}\{\hat{\zeta}_n \le 2((n-1)\alpha)^{-1}\}.$$

Here, the last inequality follows from Lemma 3.4. Now, choose a large n so that $2((n-1)\alpha)^{-1} < \zeta(F_\delta)$. Then we have

$$\begin{aligned}
\mathbb{P}\{\hat{\zeta}_n \le 2((n-1)\alpha)^{-1}\} &= \mathbb{P}\{\hat{\zeta}_n - \zeta(F_\delta) \le 2((n-1)\alpha)^{-1} - \zeta(F_\delta)\} \\
&\le \frac{\mathrm{Var}(\hat{\zeta}_n)}{\big(\zeta(F_\delta) - 2((n-1)\alpha)^{-1}\big)^2} && \text{(by Chebyshev's inequality)} \\
&\le \frac{\binom{n}{2}^{-1}[4(n-2)\zeta(F_\delta)+4]}{\big(\zeta(F_\delta) - 2((n-1)\alpha)^{-1}\big)^2} && \text{(by Lemma A3.1)} \\
&= \frac{\binom{n}{2}^{-1}\big[4(n-2)(1-\delta)^2\gamma_0+4\big]}{\big((1-\delta)^2\gamma_0 - 2((n-1)\alpha)^{-1}\big)^2}.
\end{aligned}$$

Note that the upper bound in the above inequality does not depend on G, and hence

$$\inf_{G:\zeta(G)=0} \mathbb{P}\{\hat{\zeta}_n > c_{1-\alpha}\} \ge 1 - \frac{\binom{n}{2}^{-1}\big[4(n-2)(1-\delta)^2\gamma_0+4\big]}{\big((1-\delta)^2\gamma_0 - 2((n-1)\alpha)^{-1}\big)^2}.$$

So, as n goes to infinity, $\inf_{G:\zeta(G)=0} \mathbb{P}\{\hat{\zeta}_n > c_{1-\alpha}\}$ goes to one. This completes the proof. $\blacksquare$

Proof of Theorem 3.8. We prove this theorem using a simple application of the Neyman-Pearson lemma. Let Q_1 and Q_2 be the joint distribution of the sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ under the null and alternative hypotheses, respectively. Then, we can provide a lower bound of the minimax risk $R_n(\epsilon)$ as follows

$$R_n(\epsilon) \ge 1 - \alpha - d_{TV}(Q_1, Q_2) \ge 1 - \alpha - \sqrt{\frac{1}{2}KL(Q_1, Q_2)}.$$

The first inequality follows using the fact that for any unbiased test ϕ with $\mathbb{P}_{Q_1}\{\phi = 1\} \le \alpha$, and

$$\mathbb{P}_{Q_1}\{\phi = 1\} + \mathbb{P}_{Q_2}\{\phi = 0\} = 1 - (\mathbb{P}_{Q_1}\{\phi = 0\} - \mathbb{P}_{Q_2}\{\phi = 0\}) \ge 1 - d_{TV}(Q_1, Q_2),$$

where d_{TV} denotes the total variation distance between Q_1 and Q_2. The second inequality follows from Pinsker's inequality (see. Tsybakov, 2009). Suppose G is a spherically symmetric distribution with density g and F is a distribution with density f and $\zeta(F) = \gamma_0 > 0$. Also assume that $\int (f(\mathbf{u})/g(\mathbf{u}) - 1)^2 g(\mathbf{u})\mathrm{d}\mathbf{u} = \gamma_1 < \infty$. Then setting $Q_1 = \otimes_{i=1}^n \Big((1 - \frac{\delta}{\sqrt{n}})G + \frac{\delta}{\sqrt{n}}F\Big)$ for some $\delta > 0$

and $Q_2 = \otimes_{i=1}^n G$, we have

$$\begin{aligned} KL(Q_1, Q_2) &= \int \log \prod_{i=1}^n \left\{1 + \frac{\delta}{\sqrt{n}}\left(\frac{f(\mathbf{u}_i)}{g(\mathbf{u}_i)} - 1\right)\right\} \prod_{i=1}^n d\Big(\big(1 - \frac{\delta}{\sqrt{n}}\big)G + \frac{\delta}{\sqrt{n}}F\Big)(\mathbf{u}_i) \\ &= n \int \log \left\{1 + \frac{\delta}{\sqrt{n}}\left(\frac{f(\mathbf{u}_1)}{g(\mathbf{u}_1)} - 1\right)\right\} d\Big(\big(1 - \frac{\delta}{\sqrt{n}}\big)G + \frac{\delta}{\sqrt{n}}F\Big)(\mathbf{u}_1) \\ &= n\big(1 - \frac{\delta}{\sqrt{n}}\big) \int \log \left\{1 + \frac{\delta}{\sqrt{n}}\left(\frac{f(\mathbf{u}_1)}{g(\mathbf{u}_1)} - 1\right)\right\} g(\mathbf{u}_1)\mathrm{d}\mathbf{u}_1 \\ &\qquad + n\frac{\delta}{\sqrt{n}} \int \log \left\{1 + \frac{\delta}{\sqrt{n}}\left(\frac{f(\mathbf{u}_1)}{g(\mathbf{u}_1)} - 1\right)\right\} f(\mathbf{u}_1)\mathrm{d}\mathbf{u}_1. \end{aligned}$$

Using the inequality $\log(1+y) \le y$, we get

$$\begin{aligned} KL(Q_1, Q_2) &\le n\left[\frac{\delta}{\sqrt{n}}\big(1 - \frac{\delta}{\sqrt{n}}\big) \int \big(\frac{f(\mathbf{u})}{g(\mathbf{u})} - 1\big) g(\mathbf{u}) d\mathbf{u} + \frac{\delta^2}{n} \int \big(\frac{f(\mathbf{u})}{g(\mathbf{u})} - 1\big) f(\mathbf{u})\mathrm{d}\mathbf{u}\right] \\ &= \delta^2 \left[\int \frac{f^2(\mathbf{u})}{g(\mathbf{u})} - 1\right] = \delta^2 \int \left(\frac{f(\mathbf{u})}{g(\mathbf{u})} - 1\right)^2 g(\mathbf{u})\mathrm{d}\mathbf{u} = \delta^2 \gamma_1. \end{aligned}$$

Also by Lemma 3.5, we have

$$\zeta\left(\Big(\big(1 - \frac{\delta}{\sqrt{n}}\big)G + \frac{\delta}{\sqrt{n}}F\Big)\right) = \frac{\delta^2}{n}\zeta(F) = \frac{\delta^2}{n}\gamma_0.$$

Now for any $0 < \beta < 1 - \alpha$, if we choose $\delta = \sqrt{2/\gamma_1}(1 - \alpha - \beta)$, we get

$$\zeta\left(\Big(\big(1 - \frac{\delta}{\sqrt{n}}\big)G + \frac{\delta}{\sqrt{n}}F\Big)\right) = \frac{1}{n}\left(\frac{2\gamma_0(1-\alpha-\beta)^2}{\gamma_1}\right).$$

Now define, $c(\alpha, \beta) = \big(2\gamma_0(1-\alpha-\beta)^2\big)/\gamma_1$. Then we have $\big((1 - \frac{\delta}{\sqrt{n}})G + \frac{\delta}{\sqrt{n}}F\big) \in \mathcal{F}(cn^{-1}) = \{F \mid \zeta(F) > cn^{-1}\}$ for all $0 < c < c(\alpha, \beta)$. For this choice of alternative, we also have $R_n(cn^{-1}) \ge \beta$ for all $0 < c < c(\alpha, \beta)$. Since β and $c(\alpha, \beta)$ do not depend on n, this trivially satisfies the condition $\liminf_{n\to\infty} R_n(cn^{-1}) \ge \beta$ for all $0 < c < c(\alpha, \beta)$. ■

Proof of Theorem 3.9. Here we want to show that for every $0 < \alpha < 1$ and $0 < \beta < 1 - \alpha$, there exists a constant $C(\alpha, \beta) > 0$ such that

$$\limsup_{n\to\infty} \sup_{F \in \mathcal{F}(cn^{-1})} \mathbb{P}_F^{(n)}\{\hat{\zeta}_n \le c_{1-\alpha}\} \le \beta$$

for all $c > C(\alpha, \beta)$. Now take any $\mathrm{P} \in \mathcal{F}(cn^{-1})$ with $c > 4/\alpha$ (i.e. $\zeta(\mathrm{P}) > 4/n\alpha$). Using the fact $c_{1-\alpha} \le 2((n-1)\alpha)^{-1}$ and Chebyshev's inequality, we have

$$\begin{aligned} \mathbb{P}_F^{(n)}\{\hat{\zeta}_n \le c_{1-\alpha}\} \le \mathbb{P}_F^{(n)}\{\hat{\zeta}_n \le 2((n-1)\alpha)^{-1}\} &\le \mathbb{P}_F^{(n)}\{\zeta(\mathrm{P}) - \hat{\zeta}_n \ge \zeta(\mathrm{P}) - 2((n-1)\alpha)^{-1}\} \\ &\le \frac{\mathrm{Var}\big(\hat{\zeta}_n\big)}{\big(\zeta(\mathrm{P}) - 2((n-1)\alpha)^{-1}\big)^2}, \end{aligned}$$

which holds since $\zeta(\mathrm{P}) - 2((n-1)\alpha)^{-1} > 4(n\alpha)^{-1} - 2((n-1)\alpha)^{-1} = \frac{2n-4}{n(n-1)\alpha} > 0$ for all $n \ge 2$. Now,

$$\frac{\mathrm{Var}\big(\hat{\zeta}_n\big)}{\big(\zeta(\mathrm{P}) - 2((n-1)\alpha)^{-1}\big)^2} \le \frac{\binom{n}{2}^{-1}\left[4(n-2)\zeta(\mathrm{P}) + 4\right]}{\big(\zeta(\mathrm{P}) - 2((n-1)\alpha)^{-1}\big)^2} \qquad \text{(follows from Lemma A3.1)} \qquad (3.7)$$

which implies that

$$\limsup_{n\to\infty} \sup_{F\in\mathcal{F}(cn^{-1})} \mathbb{P}_F^{(n)}\{\hat{\zeta}_n \le c_{1-\alpha}\} \le \frac{4c+4}{\big(c-2/\alpha\big)^2}.$$

It is easy to see that the upper bound is a monotonically decreasing function of c for $c > 4/\alpha$, and it converges to 0 as c increases. Hence, for any $\beta < 1-\alpha$, there exists a $r(\alpha,\beta)$ such that the upper bound is smaller than β whenever $c > r(\alpha,\beta)$. Now set $C(\alpha,\beta) = \max\{r(\alpha,\beta), 4/\alpha\}$. Then for any $c > C(\alpha,\beta)$, the maximum type II error rate of our test is upper bounded by β. ■

Proof of Theorem 3.10. If the distribution $\mathcal{F}^{(d)}$ is such that $n\zeta(F^{(d)})$ diverges to infinity, then from the proof of Theorem 3.9 (see equation (3.7)) we see that $\lim \mathbb{P}\{\hat{\zeta}_n \le c_{1-\alpha}\} = 0$. Hence, under the above condition, the power of our test converges to one. ■

Proof of Proposition 3.2. It is easy to see that the likelihood ratio of $F_{1-\frac{\beta}{\sqrt{n}}} = (1-\beta_n/\sqrt{n})G + \beta_n/\sqrt{n}F$ and G is $\Big(1+\frac{\beta_n}{\sqrt{n}}\big(f(\mathbf{u})/g(\mathbf{u})-1\big)\Big)$. Hence if $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} G$, then the log-likelihood ratio is given by,

$$L_N = \log\Big\{\prod_{i=1}^n \frac{dF_{1-\frac{\beta_n}{\sqrt{n}}}}{dG}(\mathbf{X}_i)\Big\} = \sum_{i=1}^n \log\Big\{\frac{dF_{1-\frac{\beta_n}{\sqrt{n}}}}{dG}(\mathbf{X}_i)\Big\} = \sum_{i=1}^n \log\Big(1+\frac{\beta_n}{\sqrt{n}}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)\Big).$$

Using the fact that $\log(1+y) = y - \frac{y^2}{2} + \frac{1}{2}y^2h(y)$ where $h(\cdot)$ is continuous and $\lim_{y\to 0} h(y) = 0$, we get

$$\begin{aligned} L_N = &\sum_{i=1}^n \frac{\beta_n}{\sqrt{n}}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big) - \sum_{i=1}^n \frac{\beta_n^2}{2n}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)^2 \\ &+ \sum_{i=1}^n \frac{\beta_n^2}{2n}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)^2 h\Big(\frac{\beta_n}{\sqrt{n}}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)\Big). \end{aligned}$$

Under the assumption $\int (f(\mathbf{u})/g(\mathbf{u})-1)^2 g(\mathbf{u})\mathrm{d}\mathbf{u} < \infty$ and $\beta_n \to \beta$, as n grows to infinity, we have

$$\sum_{i=1}^n \frac{\beta_n^2}{n}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)^2 \overset{a.s.}{\to} \beta^2\mathbb{E}\Big(\big(f(\mathbf{X}_1)/g(\mathbf{X}_1)-1\big)^2\Big).$$

Hence, we only need to show that

$$\sum_{i=1}^n \frac{\beta_n^2}{n}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)^2 h\Big(\frac{\beta_n}{\sqrt{n}}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)\Big)$$

converges to zero in probability. Notice that

$$\begin{aligned} &\sum_{i=1}^n \frac{\beta_n^2}{n}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)^2 h\Big(\frac{\beta_n}{\sqrt{n}}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)\Big) \\ &\le \max_{1\le i\le n} |h\Big(\frac{\beta_n}{\sqrt{n}}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)\Big)| \sum_{i=1}^n \frac{\beta_n^2}{n}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)^2. \end{aligned}$$

Therefore, it suffices to show that $\max_{1\le i\le N}\Big|h\Big(\frac{\beta_n}{\sqrt{n}}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)\Big)\Big|$ converges to zero in probability, which follows if $\max_{1\le i\le N}\Big|\frac{\beta_n}{\sqrt{n}}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i)-1\big)\Big|$ converges to zero in probability

(as $\lim_{y\to 0} h(y) = 0$ and it is continuous). Note that

$$\begin{aligned}
&\mathbb{P}\Big\{ \max_{1\le i\le n} \big|\frac{1}{\sqrt{n}}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i) - 1\big)\big| > \epsilon\Big\} \\
&\le \sum_{i=1}^{n} \mathbb{P}\Big\{\big|\frac{1}{\sqrt{n}}\big(f(\mathbf{X}_i)/g(\mathbf{X}_i) - 1\big)\big| > \epsilon\Big\} \\
&= n\mathbb{P}\Big\{\big|\frac{1}{\sqrt{n}}\big(f(\mathbf{X}_1)/g(\mathbf{X}_1) - 1\big)\big| > \epsilon\Big\} \\
&= n\mathbb{E}\Big\{I\big(\big|\frac{1}{\sqrt{n}}\big(f(\mathbf{X}_1)/g(\mathbf{X}_1) - 1\big)\big| > \epsilon\big)\Big\} \\
&\le n\mathbb{E}\Big\{\frac{\big(f(\mathbf{X}_1)/g(\mathbf{X}_1) - 1\big)^2}{n\epsilon^2} I\big|\frac{1}{\sqrt{n}}\big(f(\mathbf{X}_1)/g(\mathbf{X}_1) - 1\big)\big| > \epsilon\Big\} \\
&\le \frac{1}{\epsilon^2}\mathbb{E}\Big\{\big(f(\mathbf{X}_1)/g(\mathbf{X}_1) - 1\big)^2 I\big(\big|\frac{1}{\sqrt{n}}\big(f(\mathbf{X}_1)/g(\mathbf{X}_1) - 1\big)\big| > \epsilon\big)\Big\}.
\end{aligned}$$

Since $I\big(\big|\frac{1}{\sqrt{n}}\big(f(\mathbf{X}_1)/g(\mathbf{X}_1) - 1\big)\big| > \epsilon\big)$ converges to zero in probability, the right-hand side converges to zero by the Dominated Convergence Theorem. Hence, we have

$$\left|\log\Big\{\prod_{i=1}^{n} \frac{dF_{1-\beta_n/\sqrt{n}}}{dG}(\mathbf{X}_i)\Big\} - \frac{\beta_n}{\sqrt{n}}\sum_{i=1}^{n}\Big(f(\mathbf{X}_i)/g(\mathbf{X}_i) - 1\Big) + \frac{\beta_n^2}{2}\mathbb{E}\Big\{f(\mathbf{X}_1)/g(\mathbf{X}_1) - 1\Big\}^2\right| \to 0,$$

in probability as n goes to infinity. This completes the proof. ■

Proof of Theorem 3.11. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} G$ and $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n \overset{iid}{\sim} \text{Unif}(\mathcal{S}^{d-1})$ be independent and $\zeta(G) = 0$. For $i = 1, 2, \ldots, n$, define, $\mathbf{X}_i' = \|\mathbf{X}_i\|\mathbf{U}_i$. To prove this theorem, we only need to find the limit distribution of $\sqrt{n}\big(\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i, \mathbf{X}_i') - \mathbb{E}\big\{h(\mathbf{X}_1, \mathbf{X}_1')\big\}\big)$ for some square-integrable function h under the contiguous alternative $F_{1-\beta/\sqrt{n}} = \big(1 - \frac{\beta_n}{\sqrt{n}}\big)G + \frac{\beta_n}{\sqrt{n}}F$, where $\zeta(F) > 0$ and $\beta_n \to \beta$ as $n \to \infty$. Using the bivariate central limit theorem, we can say that as n diverges to infinity, the joint distribution of

$$\sqrt{n}\big(\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i, \mathbf{X}_i') - \mathbb{E}\big\{h(\mathbf{X}_1, \mathbf{X}_1')\big\}\big) \text{ and } \frac{\beta_n}{\sqrt{n}}\sum_{i=1}^{n}\big(\frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)} - 1\big) - \frac{\beta_n^2}{2}\mathbb{E}\big\{\frac{f(\mathbf{X}_1)}{g(\mathbf{X}_1)} - 1\big\}^2$$

converges to a bivariate normal distribution with the mean and the variance given by

$$\mu = \begin{pmatrix} 0 \\ -\frac{\beta^2}{2}\mathbb{E}\big\{\frac{f(\mathbf{X}_1)}{g(\mathbf{X}_1)} - 1\big\}^2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \text{Var}\big(h(\mathbf{X}_1, \mathbf{X}_1')\big) & \tau \\ \tau & -\frac{\beta^2}{2}\mathbb{E}\big\{\frac{f(\mathbf{X}_1)}{g(\mathbf{X}_1)} - 1\big\}^2 \end{pmatrix}, \text{where}$$

$$\begin{aligned}
\tau &= \mathbb{E}\Big\{\big\{h(\mathbf{X}_1, \mathbf{X}_1') - \mathbb{E}\{h(\mathbf{X}_1, \mathbf{X}_1')\}\big\}\beta\big\{\frac{f(\mathbf{X}_1)}{g(\mathbf{X}_1)} - 1\big\}\Big\} \\
&= \beta\int \big\{h(\mathbf{x}, \|\mathbf{x}\|\mathbf{u}) - \mathbb{E}\{h(\mathbf{X}_1, \mathbf{X}_1')\}\big\}\big(f(\mathbf{x}) - g(\mathbf{x})\big)\mathrm{d}\mathbf{x}\mathrm{d}\mu_0(\mathbf{u}),
\end{aligned}$$

for μ_0 being the probability measure corresponding to the distribution $\text{Unif}(\mathcal{S}^{d-1})$. Now using Le Cam's third lemma (see. Van der Vaart, 1998), as n diverges to infinity, under $F_{1-\beta_n/\sqrt{n}}$, we have

$$\sqrt{n}\big(\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i, \mathbf{X}_i') - \mathbb{E}\big\{h(\mathbf{X}_1, \mathbf{X}_1')\big\}\big) \xrightarrow{D} \mathcal{N}_1\Big(\tau, \text{Var}\big(h(\mathbf{X}_1, \mathbf{X}_1')\big)\Big).$$

Now, using similar arguments as in Theorem 1 on page 79 from Lee (1990) and contiguity arguments, under $F_{1-\beta_n/\sqrt{n}}$, we get

$$n\hat{\zeta}_n \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i \left(\big(Z_i + \beta\, \mathbb{E}_F\{\psi_i(\mathbf{X}_1, \mathbf{X}_1')\}\big)^2 - 1 \right),$$

where Z_i's are i.i.d. $\mathcal{N}_1(0,1)$ variables and ψ_i's are defined in the statement of the theorem. ■

Proof of Theorem 3.12. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} G$ and $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n \overset{iid}{\sim} \text{Unif}(\mathcal{S}^{d-1})$ be independent and $\mathbf{X}_i' = \|\mathbf{X}_i\|\mathbf{U}_i$ for $i = 1, 2, \ldots, n$. Also let $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$ be the vector of i.i.d random variables from the Bernoulli(0.5) distribution. Then the resampled test statistic $\hat{\zeta}_n(\boldsymbol{S})$ can be written as

$$\frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} g\big((\mathbf{Y}_i, \mathbf{Y}_i'), (\mathbf{Y}_j, \mathbf{Y}_j')\big),$$

where $\mathbf{Y}_i = S_i\mathbf{X}_i + (1 - S_i)\mathbf{X}_i'$ and $\mathbf{Y}_i' = (1 - S_i)\mathbf{X}_i + S_i\mathbf{X}_i'$. Since, under any fixed alternative F, $\{(\mathbf{Y}_i, \mathbf{Y}_i')\}$ is a sequence of i.i.d. random vectors, we can apply Theorem 1 from Lee (1990) page 79 to find the limiting distribution of $\hat{\zeta}_n(\boldsymbol{S})$. Note that $\mathbf{Y}_i \overset{D}{=} \mathbf{Y}_i'$ and hence $g_1\big((\mathbf{y}, \mathbf{y}')\big) = \mathbb{E}\{g\big((\mathbf{y}, \mathbf{y}'), (\mathbf{Y}_1, \mathbf{Y}_1')\big)\} = 0$. Therefore, $\hat{\zeta}_n(\boldsymbol{S})$ is a U-statistic with a first-order degenerate kernel function, and we get

$$n\hat{\zeta}_n(\boldsymbol{S}) \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1),$$

where $\{\lambda_i\}$ is a square integrable sequence and $\{Z_i\}$ is a sequence of i.i.d. $\mathcal{N}_1(0,1)$ random variables.

Now suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} F_{1-\beta_n n^{-1/2}} = (1 - \frac{\beta}{\sqrt{n}})G + \frac{\beta}{\sqrt{n}}F$ where $\beta_n \to \beta$, $\zeta(G) = 0$ and $\zeta(F) > 0$. Then, to find the limiting distribution of $n\zeta_n(\boldsymbol{S})$, we need to find the joint limiting distribution of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_\ell\big((\mathbf{Y}_i, \mathbf{Y}_i')\big) \text{ and } \frac{\beta_n}{\sqrt{n}} \sum_{i=1}^{n} \big(\frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)} - 1\big) - \frac{\beta_n^2}{2}\mathbb{E}\big\{\frac{f(\mathbf{X}_1)}{g(\mathbf{X}_1)} - 1\big\}^2,$$

where $\{\varphi_\ell\}$ are the solutions of the integral equation

$$\int g((\mathbf{x}_1, \mathbf{x}_1'), (\mathbf{y}_1, \mathbf{y}_1'))\varphi_\ell((\mathbf{y}_1, \mathbf{y}_1'))\mathrm{d}\nu^*\big((\mathbf{y}_1, \mathbf{y}_1')\big) = \lambda\varphi_\ell\big((\mathbf{x}_1, \mathbf{x}_1')\big), \tag{3.8}$$

for ν^* being the joint distribution of $(\mathbf{Y}_1, \mathbf{Y}_1')$ when $\mathbf{X}_1 \sim G$. By the bivariate central limit theorem, we can see that the joint distribution of these two variables converges to a bivariate normal distribution with mean and variance given by

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ -\frac{\beta^2}{2}\mathbb{E}\big\{\frac{f(\mathbf{X}_1)}{g(\mathbf{X}_1)} - 1\big\}^2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \text{Var}\big(\varphi_\ell(\mathbf{Y}_1, \mathbf{Y}_1')\big) & \tau \\ \tau & -\frac{\beta^2}{2}\mathbb{E}\big\{\frac{f(\mathbf{X}_1)}{g(\mathbf{X}_1)} - 1\big\}^2 \end{pmatrix}, \text{where}$$

$$\begin{aligned} \tau &= \mathbb{E}\Big\{\big\{\varphi_\ell(\mathbf{Y}_1, \mathbf{Y}_1')\big\}\beta\big\{\frac{f(\mathbf{X}_1)}{g(\mathbf{X}_1)} - 1\big\}\Big\} \\ &= \beta\, \mathbb{E}_F\{\varphi_\ell(\mathbf{Y}_1, \mathbf{Y}_1')\} = \beta\, \big\{\frac{1}{2}\mathbb{E}\{\varphi_\ell\big((\mathbf{X}_1, \mathbf{X}_1')\big)\} + \frac{1}{2}\mathbb{E}\{\varphi_\ell\big((\mathbf{X}_1', \mathbf{X}_1)\big)\}\big\}. \end{aligned}$$

Now note that $g\big((\mathbf{x}_1, \mathbf{x}_1'), (\mathbf{y}_1, \mathbf{y}_1')\big) = -g\big((\mathbf{x}_1', \mathbf{x}_1), (\mathbf{y}_1, \mathbf{y}_1')\big)$. Then using (3.8), we get $\varphi_\ell\big((\mathbf{x}_1, \mathbf{x}_1')\big) = -\varphi_\ell\big((\mathbf{x}_1', \mathbf{x}_1)\big)$. Using this, we get $\tau = 0$. So, using Le Cam's third lemma (see. Van der Vaart, 1998), as n diverges to infinity, under $F_{1-\beta_n/\sqrt{n}}$, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_\ell\big((\mathbf{Y}_i, \mathbf{Y}_i')\big) \xrightarrow{D} \mathcal{N}_1\Big(0, \mathrm{Var}\big(\varphi_\ell\big((\mathbf{Y}_1, \mathbf{Y}_1')\big)\big)\Big) = \mathcal{N}_1(0, 1).$$

Using similar arguments as in Theorem 1 on page 79 from Lee (1990) and contiguity, we get $n\hat{\zeta}_n(\boldsymbol{S}) \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1)$, where $\{\lambda_i\}$ is a square-integrable sequence and $\{Z_i\}$ is a sequence of i.i.d. standard normal random variables. ■

3.5.1 Expression of $\zeta(\mathrm{P})$ for Gaussian distributions

In this section, we present a closed-form expression for $\zeta(\mathrm{P})$ when P is $\mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma})$. But before that, we present some preliminary lemmas.

Lemma A3.2. *If $\mathbf{X}_1 \sim \mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma}_1)$ and $\mathbf{X}_2 \sim \mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma}_2)$ are independent d-dimensional random vectors,*

$$\mathbb{E}\Big\{ \exp\big(-\frac{1}{2d}\|\mathbf{X}_1 - \mathbf{X}_2\|^2\big)\Big\} = \Big|\frac{1}{d}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) + \mathbf{I}_d\Big|^{-1/2},$$

where $\big|\mathbf{A}\big|$ denotes the determinant of a matrix $\mathbf{A}$, and $\mathbf{I}_d$ is the $d \times d$ identity matrix.

Proof of Lemma A3.2. Here $\mathbf{N}_0 = \mathbf{X}_1 - \mathbf{X}_2$ follows $\mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$. So, we have

$$\mathbb{E}\Big\{ \exp\big(-\frac{1}{2d}\|\mathbf{N}_0\|^2\big)\Big\} = \int \frac{1}{(2\pi)^{d/2}\Big|\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\Big|^{1/2}} \exp\Big(-\frac{1}{2d}\|\mathbf{u}\|^2 - \frac{1}{2}\mathbf{u}^\top(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\mathbf{u}\Big)\mathrm{d}\mathbf{u}.$$

Note that the exponent on the right side is the same as that of the density of a normal distribution with mean $\mathbf{0}_d$ and variance-covariance matrix $\big(\frac{1}{d}\mathbf{I}_d + (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\big)^{-1}$. So, we have

$$\mathbb{E}\Big\{ \exp\big(-\frac{1}{2d}\|\mathbf{N}_0\|^2\big)\Big\} = \frac{1}{\Big|\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\Big|^{1/2}\Big|\frac{1}{d}\mathbf{I}_d + (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\Big|^{1/2}} = \Big|\frac{1}{d}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) + \mathbf{I}_d\Big|^{-1/2}.$$

This completes the proof. ■

Lemma A3.3. *If $\mathbf{X}$ follows a d-variate distribution with density f, $\mathbf{U} \sim \mathrm{Unif}(\mathcal{S}^{d-1})$, and they are independent, then $\|\mathbf{X}\|\mathbf{U}$ has the density $\int f(\mathbf{H}^\top\mathbf{u})\nu_0(\mathbf{H})$, where ν_0 is the Haar measure on the set of all $d \times d$ orthogonal matrices.*

Proof of Lemma A3.3. Let μ_0 be the probability measure corresponding to distribution $\mathrm{Unif}(\mathcal{S}^{d-1})$ and g be a bounded continuous function. Then we have

$$\mathbb{E}\{g(\|\mathbf{X}\|\mathbf{U})\} = \int\int g(\|\mathbf{x}\|\mathbf{u}) f(\mathbf{x})\mathrm{d}\mu_0(\mathbf{u})\mathrm{d}\mathbf{x} = \int\int g(\mathbf{H}\mathbf{x}) f(\mathbf{x})\mathrm{d}\nu_0(\mathbf{H})\mathrm{d}\mathbf{x}.$$

Since for any fixed $\mathbf{x}$ and $\mathbf{u}$, there exists a unique orthogonal matrix $\mathbf{H}$ such that $\mathbf{Hx} = \|\mathbf{x}\|\mathbf{u}$, the second equality follows by substitution. Now if we substitute $\mathbf{Hx} = \mathbf{v}$, we get

$$\mathbb{E}\{g(\|\mathbf{X}\|\mathbf{U})\} = \int g(\mathbf{v})\Big(\int f(\mathbf{H}^\top\mathbf{v})\mathrm{d}\nu_0(\mathbf{H})\Big)\Big|\mathbf{H}^\top\Big|\mathrm{d}\mathbf{v} = \int g(\mathbf{v})\Big(\int f(\mathbf{H}^\top\mathbf{v})\mathrm{d}\nu_0(\mathbf{H})\Big)\mathrm{d}\mathbf{v}.$$

Since g is an arbitrary bounded continuous function, the result follows. ■

Proposition A3.1. *If* $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma})$, *then*

$$\begin{aligned}\zeta(\mathrm{P}) = \Big|\frac{2}{d}\mathbf{\Sigma} + \mathbf{I}_d\Big|^{-1/2} + \int\int \Big|\frac{1}{d}(\mathbf{H}_1\mathbf{\Sigma}\mathbf{H}_1^\top + \mathbf{H}_2\mathbf{\Sigma}\mathbf{H}_2^\top) + \mathbf{I}_d\Big|^{-1/2}\mathrm{d}\nu_0(\mathbf{H}_1)\mathrm{d}\nu_0(\mathbf{H}_2)\\ - 2\int \Big|\frac{1}{d}(\mathbf{\Sigma} + \mathbf{H}\mathbf{\Sigma}\mathbf{H}^\top) + \mathbf{I}_d\Big|^{-1/2}\mathrm{d}\nu_0(\mathbf{H}),\end{aligned}$$

where ν_0 *is the Haar measure on the set of all orthogonal matrices of order* $d \times d$.

Proof of Proposition A3.1. Recall that our measure can be written as (see Theorem 3.1)

$$\zeta(\mathrm{P}) = \mathbb{E}\big\{\exp\{-\frac{1}{2d}\|\mathbf{X}_1 - \mathbf{X}_2\|^2\}\big\} + \mathbb{E}\big\{\exp\{-\frac{1}{2d}\|\mathbf{X}_1' - \mathbf{X}_2'\|^2\}\big\} - 2\mathbb{E}\big\{\exp\{-\frac{1}{2d}\|\mathbf{X}_1 - \mathbf{X}_2'\|^2\}\big\}.$$

Let us look at the individual terms separately. The first term on the right side is the same as the term in Lemma 3.2 with $\mathbf{X}_1$ and $\mathbf{X}_2$ being i.i.d. $\mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma})$. Therefore,

$$\mathbb{E}\big\{\exp\{-\frac{1}{2d}\|\mathbf{X}_1 - \mathbf{X}_2\|^2\}\big\} = \Big|\mathbf{I}_d + \frac{2}{d}\mathbf{\Sigma}\Big|^{-1/2}.$$

Now note that the second term can be written as

$$\begin{aligned}&\mathbb{E}\big\{\exp\{-\frac{1}{2d}\|\mathbf{X}_1' - \mathbf{X}_2'\|^2\}\big\}\\ &= \int \exp\{-\frac{1}{2d}\|\mathbf{x}_1 - \mathbf{x}_2\|^2\}\frac{1}{(2\pi)^d\big|\mathbf{\Sigma}\big|}\exp\Big\{-\frac{1}{2}\mathbf{x}_1^\top(\mathbf{H}_1\mathbf{\Sigma}\mathbf{H}_1^\top)^{-1}\mathbf{x}_1 - \frac{1}{2}\mathbf{x}_2^\top(\mathbf{H}_2\mathbf{\Sigma}\mathbf{H}_2^\top)^{-1}\mathbf{x}_2\Big\}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2\mathrm{d}\nu_0(\mathbf{H}_1)\mathrm{d}\nu_0(\mathbf{H}_2)\\ &= \int \frac{\big|\mathbf{\Sigma}\big|^{-1}\big|\mathbf{H}_1\mathbf{\Sigma}\mathbf{H}_1^\top\big|^{1/2}\big|\mathbf{H}_2\mathbf{\Sigma}\mathbf{H}_2^\top\big|^{1/2}}{\Big|\frac{1}{d}\big(\mathbf{H}_1\mathbf{\Sigma}\mathbf{H}_1^\top + \mathbf{H}_2\mathbf{\Sigma}\mathbf{H}_2^\top\big) + \mathbf{I}_d\Big|^{1/2}}\mathrm{d}\nu_0(\mathbf{H}_1)\mathrm{d}\nu_0(\mathbf{H}_2)\\ &= \int \Big|\frac{1}{d}\big(\mathbf{H}_1\mathbf{\Sigma}\mathbf{H}_1^\top + \mathbf{H}_2\mathbf{\Sigma}\mathbf{H}_2^\top\big) + \mathbf{I}_d\Big|^{-1/2}\mathrm{d}\nu_0(\mathbf{H}_1)\mathrm{d}\nu_0(\mathbf{H}_2).\end{aligned}$$

Similarly, one can also show that

$$\mathbb{E}\big\{\exp\{-\frac{1}{2d}\|\mathbf{X}_1 - \mathbf{X}_2'\|^2\}\big\} = \int \Big|\frac{1}{d}\big(\mathbf{H}\mathbf{\Sigma}\mathbf{H}^\top + \mathbf{\Sigma}\big) + \mathbf{I}_d\Big|^{-1/2}\mathrm{d}\nu_0(\mathbf{H}).$$

This completes the proof. ■

Chapter 4

Distribution-free Tests of Spherical Symmetry

The test of spherical symmetry proposed in Chapter 3 is consistent against general alternatives and applicable to high-dimensional data. However, it uses a resampling algorithm for calibration, which makes it computationally demanding. In this chapter, we propose some graph-based distribution-free tests for spherical symmetry which takes care of this problem.

Lemma 3.1 showed that a random vector $\mathbf{X}$ follows a spherically symmetric distribution if and only if the distribution of $\mathbf{X}$ and that of its spherically symmetric variant $\mathbf{X}' = \|\mathbf{X}\|\mathbf{U}$ (where $\mathbf{U} \sim \text{Unif}(S^{d-1})$ and independent of $\mathbf{X}$) are identical. In this context, we have the following result, which is a direct analog of Theorem 2 of Maa, Pearl *&* Bartoszyński (1996).

Lemma 4.1. *Let $\mathbf{X}_1, \mathbf{X}_2$ be two independent realizations of $\mathbf{X} \sim P$, and $\mathbf{X}_1', \mathbf{X}_2'$ be their spherically symmetric variants. Assume that P is absolutely continuous with square integrable density function $p(\cdot)$. Also, consider a function $h(\cdot,\cdot)$ that satisfies (a) $h(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$, (b) $h(\mathbf{x}+\boldsymbol{a},\mathbf{y}+\boldsymbol{a}) = h(\mathbf{x},\mathbf{y})$ for all $\boldsymbol{a} \in \mathbb{R}^d$ and (c) the class of sets $S_t = \{\mathbf{x} | h(0,\mathbf{x}) \le t\}$ regularly shrinks towards 0 as $t \downarrow 0$. Then, we have*

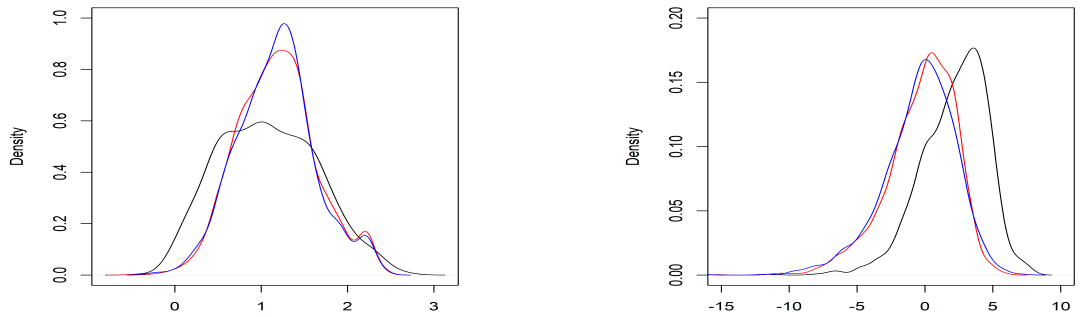
$$h(\mathbf{X}_1, \mathbf{X}_2) \overset{D}{=} h(\mathbf{X}_1, \mathbf{X}_2') \overset{D}{=} h(\mathbf{X}_1', \mathbf{X}_2') \qquad \textit{if and only if} \qquad \text{P } \textit{is spherically symmetric.}$$

We refer the reader to Chapter 7 of Wheeden *&* Zygmund (1977) for regularly shrinking sets and their consequences in the Lebesgue point theorem. In the literature, $h(\mathbf{x},\mathbf{y})$ is popularly chosen as the ℓ_2 distance between $\mathbf{x}$ and $\mathbf{y}$. Hence, we have

$$\|\mathbf{X}_1 - \mathbf{X}_2\| \overset{D}{=} \|\mathbf{X}_1 - \mathbf{X}_2'\| \overset{D}{=} \|\mathbf{X}_1' - \mathbf{X}_2'\| \quad \text{ if and only if } \quad \text{P is spherically symmetric.}$$

The left column in Figure 4.1 gives the density estimates of the logarithm of these three types of Euclidean distances when 50 observations are generated from $\mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma})$ (with $d = 10$ and $d = 100$), where $\boldsymbol{\Sigma}$ has all diagonal entries 1 and all off-diagonal entries $(d-1)/d$ (here we take the log transformation of the distances to avoid the restriction on their supports). The difference between these three distributions is evident from this figure. Note that since $\|\mathbf{X}_1\| = \|\mathbf{X}_1'\|$ and $\|\mathbf{X}_2\| = \|\mathbf{X}_2'\|$, the difference between $\|\mathbf{X}_1 - \mathbf{X}_2\|$, $\|\mathbf{X}_1 - \mathbf{X}_2'\|$ and $\|\mathbf{X}_1' - \mathbf{X}_2'\|$ mainly comes from the inner products $\mathbf{X}_1^\top \mathbf{X}_2$, $\mathbf{X}_1^\top \mathbf{X}_2'$ and $\mathbf{X}_1'^\top \mathbf{X}_2'$. However, here we have $\mathbb{E}[\mathbf{X}_1^\top \mathbf{X}_2] = \mathbb{E}[\mathbf{X}_1^\top \mathbf{X}_2'] = \mathbb{E}[\mathbf{X}_1'^\top \mathbf{X}_2'] = 0$. But, Lemma A4.2 (see Section 4.7.1) shows that as d increases, the asymptotic order of $\mathbb{E}(\mathbf{X}_1^\top \mathbf{X}_2)^2$ turns out to be higher than those of $\mathbb{E}(\mathbf{X}_1^\top \mathbf{X}_2')^2$ and $\mathbb{E}(\mathbf{X}_1'^\top \mathbf{X}_2')^2$. Figure 4.1

(a) log(Euclidean distance) for $d = 10$ (b) log(squared inner-product) for $d = 10$



(c) log(Euclidean distance) for $d = 100$ (d) log(squared inner-product) for $d = 100$

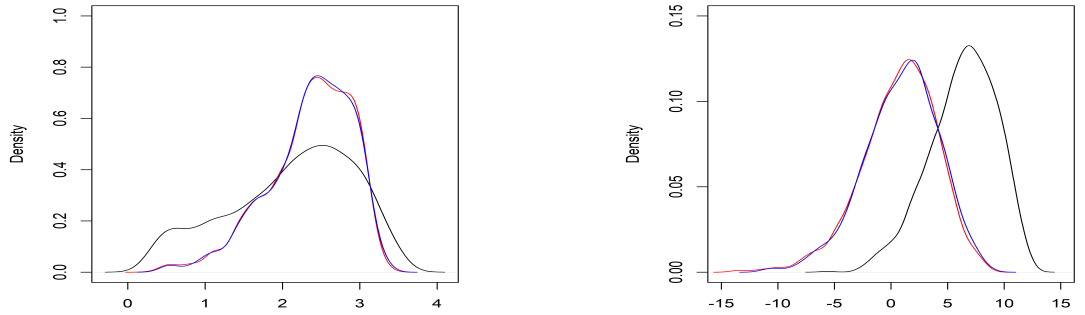


Fig. 4.1 *Densities of the logarithm of* $\|\mathbf{X}_1 - \mathbf{X}_2\|$ *(black),* $\|\mathbf{X}_1 - \mathbf{X}_2'\|$ *(blue) and* $\|\mathbf{X}_1' - \mathbf{X}_2'\|$ *(red) and those of the logarithm of* $(\mathbf{X}_1^\top \mathbf{X}_2)^2$ *(black),* $(\mathbf{X}_1^\top \mathbf{X}_2')^2$ *(blue) and* $(\mathbf{X}_1'^\top \mathbf{X}_2')^2$ *(red) when 50 observations are generated from* $\mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma})$*, where* $\mathbf{\Sigma} = ((\sigma_{ij}))$ *has* $\sigma_{ij} = 1$ *for* $i = j$ *and* $\sigma_{ij} = (d-1)/d$ *for* $i \neq j$*.*

also supports that. Clearly, this difference becomes more prominent if we look at the densities of the logarithm of squares of corresponding inner products. The right column in Figure 4.1 shows that $(\mathbf{X}_1^\top \mathbf{X}_2)^2$ is stochastically larger $(\mathbf{X}_1^\top \mathbf{X}_2')^2$ and $(\mathbf{X}_1'^\top \mathbf{X}_2')^2$. The difference is more significant in the case of $d = 100$. This phenomenon gives us the motivation to construct tests based on the squares of the inner products, which are described in the following sections.

4.1 String signs and string ranks

Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} \mathrm{P}$, and $\mathbf{X}_i' = \|\mathbf{X}_i\|\mathbf{U}_i$ is the spherically symmetric variant of $\mathbf{X}_i$ $(i = 1, 2, \ldots, n)$. Define $\boldsymbol{Z}_i = \mathbf{X}_i$ and $\boldsymbol{Z}_{n+i} = \mathbf{X}_i'$ for $i = 1, 2, \ldots, n$. To test whether P is spherically symmetric, we consider an edge-weighted undirected complete graph $\mathcal{K}_{2n}$ on the $2n$ vertices $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_{2n}$, where $\theta(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \exp\{-(\frac{1}{d}\boldsymbol{Z}_i^\top \boldsymbol{Z}_j)^2\}$ is the weight (cost) associated with the edge joining $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$ $(1 \leq i < j \leq 2n)$. We consider a path of length $n-1$ that traverses through either $\mathbf{X}_i$ or $\mathbf{X}_i'$ for each $i = 1, 2, \ldots, n$, and we call it a covering path. Clearly, there are $2^n n!$ many covering paths. But for every path, there exists another path in the reverse order. If we consider them as the same path, the number of distinct covering paths turns out to be $2^{n-1}n!$. When the observations come from a continuous distribution, each of these $2^{n-1}n!$ paths have different costs (the cost of a path is defined as the sum of the costs of its edges) with probability

one. Among these paths, we choose the one with the minimum cost, and it is called the shortest covering path $\mathcal{P}$ (see Biswas, Mukhopadhyay & Ghosh, 2015 for the use of shortest covering path in the context of one-sample location problem). Note that finding this path is equivalent to finding

$$(\boldsymbol{S}, \boldsymbol{\Pi}) = \underset{\substack{\boldsymbol{s}\in\{0,1\}^n \\ \boldsymbol{\pi}\in\mathcal{S}_n}}{\arg\min} \left[\sum_{i=1}^{n-1} \theta(\boldsymbol{Y}_{s_{\pi_i},\pi_i}, \boldsymbol{Y}_{s_{\pi_{i+1}},\pi_{i+1}}) \right], \tag{4.1}$$

where $\mathcal{S}_n$ is the set of all permutations of $\{1, 2, \ldots, n\}$ and for any $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)$, $\boldsymbol{Y}_{s_i,i} = s_i\mathbf{X}_i + (1-s_i)\mathbf{X}'_i$ (i.e., $\boldsymbol{Y}_i = \mathbf{X}_i$ if $s_i = 1$ and $\boldsymbol{Y}_i = \mathbf{X}'_i$ if $s_i = 0$) for all $i = 1, 2, \ldots, n$. Here $\boldsymbol{\Pi}$ gives us the arrangement of n observation along $\mathcal{P}$. This leads to a new notion of ranks. The position of $\mathbf{X}_i$ (or $\mathbf{X}'_i$) along $\mathcal{P}$ is called the rank of $\mathbf{X}_i$, and it is denoted by R_i ($i = 1, 2, \ldots, n$). One can notice that $\boldsymbol{\Pi}^{-1}$, the inverse permutation of $\boldsymbol{\Pi}$, gives the rank vector $\boldsymbol{R} = (R_1, R_2, \ldots, R_n)$. Similarly, $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$ can be viewed as a sign vector (instead of 1 and -1, each of its elements takes the values 1 and 0), where S_i gives the information whether $\mathbf{X}_i$ or $\mathbf{X}'_i$ lies on the path $\mathcal{P}$. Since these signs and ranks are computed along the shortest covering path $\mathcal{P}$, which can be viewed as a string joining n observations or their spherical analogs, we shall refer to them as string ranks and string signs, respectively. Note that they satisfy the following properties.

Theorem 4.1. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent realizations of a random vector $\mathbf{X}$ following a continuous distribution* P. *Also, define $\boldsymbol{S}$ and $\boldsymbol{R}$ as the vector of string signs and string ranks of $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$. Then, we have the following results.*

(a) *If* P *is spherically symmetric, $\boldsymbol{S} \sim \text{Unif}(\{0,1\}^n)$, $\boldsymbol{R} \sim \text{Unif}(\mathcal{S}_n)$ and they are independent.*

(b) *Even if* P *is not spherically symmetric, $\boldsymbol{R} \sim \text{Unif}(\mathcal{S}_n)$ and for any given $\boldsymbol{R}$ or $\boldsymbol{\Pi} = (\pi_1, \pi_2, \ldots, \pi_n)$, we have the following weak dependence structure of $\boldsymbol{S}$.*

$$S_{\pi_1} \mid S_{\pi_2}, \cdots, S_{\pi_n} \overset{D}{=} S_{\pi_1} \mid S_{\pi_2},$$
$$S_{\pi_n} \mid S_{\pi_1}, \cdots, S_{\pi_{n-1}} \overset{D}{=} S_{\pi_n} \mid S_{\pi_{n-1}},$$
$$S_{\pi_i} \mid S_{\pi_1}, \cdots, S_{\pi_{i-1}}, S_{\pi_{i+1}}, \cdots, S_{\pi_n} \overset{D}{=} S_{\pi_i} \mid S_{\pi_{i-1}}, S_{\pi_{i+1}} \text{ for } i = 2, \ldots, n-1.$$

4.2 Tests based on string signs and string ranks

From part (a) of Theorem 4.1, it is clear that under the null hypothesis of spherical symmetry, the distribution of $(\boldsymbol{S}, \boldsymbol{R})$ matches with the joint null distribution of univariate signs and ranks used in one-sample testing problem (e.g., sign test or signed rank test). So, any test statistic computed based on these string signs and sting ranks has the distribution-free property, and its null distribution matches with that of the corresponding univariate statistic based on usual signs and ranks. For instance, one can consider a test based on the sign statistic $T_S = \sum_{i=1}^n S_i$ or the runs statistic $T_R = 1 + \sum_{i=1}^{n-1} \mathrm{I}\{S_{\pi_i} \neq S_{\pi_{i+1}}\}$. Note that the values of T_S and T_R remain the same if the path $\mathcal{P}$ is traversed in the reverse order. For any linear rank statistic of the form $T_{LR} = \sum_{i=1}^n S_i a(R_i) = \sum_{i=1}^n S_{\pi_i} a(i)$, we have this property if the score function $a(\cdot)$ satisfies

$a(i) = a(n-i+1)$ for all $i = 1, 2, \ldots, n$ (taking $a(i) = 1\ \forall i$, we get T_S). In the case of the signed rank test, we have $a(i) = i$, which does not satisfy this property. So, we do not recommend this test.

Part (b) of Theorem 4.1 gives us some idea about the behavior of $\boldsymbol{S}$ and $\boldsymbol{R}$ under the alternative hypothesis (i.e., when P is not spherical). From Figure 4.1 and also from Lemma A4.2 (see Section 4.7.1), it is quite transparent that when P is not spherical, the total cost of any covering path will be small if most of the edges are of the form $(\mathbf{X}_i, \mathbf{X}_j)$. So, $\mathcal{P}$ is supposed to contain more $\mathbf{X}_i$s than $\mathbf{X}_i'$s. As a result, most of the elements of $\boldsymbol{S}$ turn out to be 1, and that leads to a higher value of T_S and a lower value of T_R. So, we can reject the null hypothesis accordingly, where the cut-offs can be obtained from the statistical tables available for the corresponding univariate nonparametric tests. To demonstrate this, let us recall the example involving 50 observations from 10-dimensional and 100-dimensional normal distributions. In each case, we repeated the experiment 1000 times. Figure 4.2 shows the bar diagram of the observed values of T_S and T_R in these 1000 cases. We can see that in all these cases, T_S took higher values than the cut-off (31) for the sign test at 5% nominal level. Similarly, T_R turned out to be smaller than the corresponding cut-off (20) in all cases. Interestingly, both for sign and runs tests, the evidence was stronger for $d = 100$.

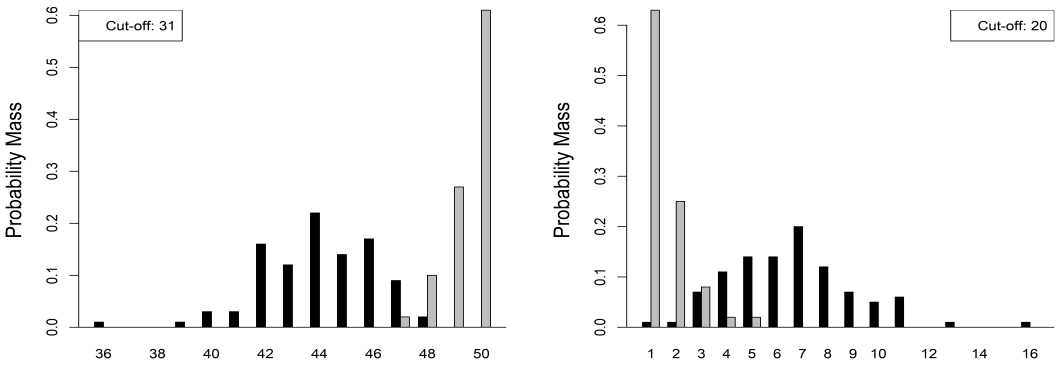


Fig. 4.2 *Distributions of T_S and T_R over 100 simulations for $d = 10$ (black bar) and $d = 100$ (grey bar) in the example involving normal distribution considered in Figure 4.1.*

4.2.1 Algorithm for finding the shortest covering path

To find the shortest covering path, one needs to consider $2^{n-1}n!$ distinct covering paths and choose the one with the minimum cost. So, unless the sample size is small, finding the shortest covering path $\mathcal{P}$ by complete enumeration becomes computationally infeasible. In fact, the optimization problem in (4.1) is equivalent to the well-known traveling salesman's problem, which is NP-hard (see Garey & Johnson, 1979). However, following Biswas, Mukhopadhyay & Ghosh (2015), we can use a heuristic method based on Prim's algorithm (Prim, 1957) for finding $\mathcal{P}$. Consider the undirected complete graph $\mathcal{K}_{2n}$ with the cost matrix $\Theta = ((\ \theta(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\ ))_{1 \le i,j \le 2n}$, where $\boldsymbol{Z}_i = \mathbf{X}_i$ and $\boldsymbol{Z}_{n+i} = \mathbf{X}_i'$ for $i = 1, 2, \ldots, n$ as defined before. First, we select the pair (i, j) (where $i \neq j$

and $|i-j| \neq n$) such that $\theta(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$ is minimum among the cost associated with such edges. We consider the edge $(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$ as a path $\mathcal{P}$ of length 1 with $\mathcal{E}_0 = \{i, j\}$ as its two end points. We define the sets $A_0 = \{i, j\}$ and $A_1 = \{k : \ k \neq \ell \text{ and } |k-\ell| \neq n \text{ for any } \ell \in A_0\}$. At the next step, we find $q \in A_1$ and $r \in \mathcal{E}_0$ such that $\theta(\boldsymbol{Z}_q, \boldsymbol{Z}_r) = \min_{k \in A_1, \ell \in \mathcal{E}_0} \theta(\boldsymbol{Z}_k, \boldsymbol{Z}_\ell)$. We join the edge $(\boldsymbol{Z}_q, \boldsymbol{Z}_r)$ to $\mathcal{P}$ to get a path $\mathcal{P} \leftarrow \mathcal{P} \cup (\boldsymbol{Z}_q, \boldsymbol{Z}_r)$ of length 2. The sets of visited nodes A_0 and the end points $\mathcal{E}_0$ of the path $\mathcal{P}$ are updated as $A_0 \leftarrow A_0 \cup \{q\}$ and $\mathcal{E}_0 \leftarrow \big(\mathcal{E}_0 \cup \{q\}\big) \setminus \{r\}$. The set A_1 is updated accordingly. We use this method repeatedly until we get $|A_0| = n$ and $|A_1| = 0$. The path $\mathcal{P}$ of length $n-1$ thus obtained is considered as the shortest covering path. Clearly, it contains either $\mathbf{X}_i$ or its spherically symmetric variant $\mathbf{X}_i'$ for every $i = 1, 2, \ldots, n$.

We use a toy example with 5 bivariate observations to demonstrate this algorithm. Figure 4.3(a) shows these 5 observations (blue dots) and their spherically symmetric variants (red dots). First, we join $\boldsymbol{Z}_3 = \mathbf{X}_3$ and $\boldsymbol{Z}_2 = \mathbf{X}_2$, the pair having the minimum cost (note that we do not consider the pairs $(\boldsymbol{Z}_i = \mathbf{X}_i, \boldsymbol{Z}_{5+i} = \mathbf{X}_i')$ for $i = 1, 2, \ldots, 5$) and consequently remove their spherical variants $\boldsymbol{Z}_8 = \mathbf{X}_3'$ and $\boldsymbol{Z}_7 = \mathbf{X}_2'$ from future considerations. So, we have $A_0 = \{2, 3\}$, $A_1 = \{1, 4, 5, 6, 9, 10\}$ and $\mathcal{E}_0 = \{3, 2\}$. At the next step, we join $\boldsymbol{Z}_4 = \mathbf{X}_4$ and $\boldsymbol{Z}_3 = \mathbf{X}_3$. This leads to $A_0 = \{2, 3, 4\}$, $A_1 = \{1, 5, 6, 10\}$ and $\mathcal{E}_0 = \{4, 2\}$ (see Figure 4.3(b)). Next, we join $\boldsymbol{Z}_5 = \mathbf{X}_5$ and $\boldsymbol{Z}_4 = \mathbf{X}_4$ to get $A_0 = \{2, 3, 4, 5\}$, $A_1 = \{1, 6\}$ and $\mathcal{E}_0 = \{5, 2\}$ (see Figure 4.3(c)). Finally, $\boldsymbol{Z}_2 = \mathbf{X}_2$ and $\boldsymbol{Z}_6 = \mathbf{X}_1'$ are joined (see Figure 4.3(d)). As a result, we get $\mathbf{X}_5 - \mathbf{X}_4 - \mathbf{X}_3 - \mathbf{X}_2 - \mathbf{X}_1'$ (or $\mathbf{X}_1' - \mathbf{X}_2 - \mathbf{X}_3 - \mathbf{X}_4 - \mathbf{X}_5$) as the shortest covering path $\mathcal{P}$. So, we have string signs $S_1 = 0, S_2 = S_3 = S_4 = S_5 = 1$ and string ranks $R_1 = 5, R_2 = 4, R_3 = 3, R_4 = 2, R_5 = 1$ (or $R_1 = 1, R_2 = 2, R_3 = 3, R_4 = 4, R_5 = 5$ if traversed in the reverse order). Therefore, the sign statistic (number of blue dots on $\mathcal{P}$) T_S turns out to be 4, and the runs statistic (the number of runs in the sequence red and blue dots on $\mathcal{P}$) T_R turns out to be 2.

Clearly, this is a heuristic algorithm, and it may lead to a sub-optimal solution of (4.1) in some cases. However, it is clear that under the null hypothesis, $\boldsymbol{S}$ and $\boldsymbol{R}$ have the same distribution irrespective of whether they are computed along the actual shortest covering path $\mathcal{P}$ or along the shortest covering path computed using the algorithm (call it $\mathcal{P}_0$). This is because of the exchangeability of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and their symmetric variants. But, in some cases, the values of T_S and T_R may differ if they are computed along these two paths. To investigate this, we generate 5 observations from $\mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\Sigma}_0 = \text{diag}(d, 1, 1, \ldots, 1)$. In this case, it is possible to find the actual $\mathcal{P}$ by complete enumeration. We compute T_S and T_R along the paths $\mathcal{P}$ and $\mathcal{P}_0$ and calculate the corresponding differences. For each value of d $(3, 30, 300, 3000)$, the experiment is repeated 1000 times, and the distributions of the differences are given by bar plots in Figure 4.4. We observe that for both T_S and T_R, the difference concentrates around zero with increasing dimensions. So, in moderate and higher dimensions, the test statistic computed along $\mathcal{P}$ matches with that computed using $\mathcal{P}_0$ in almost all cases. We have seen that in high dimensions, the actual

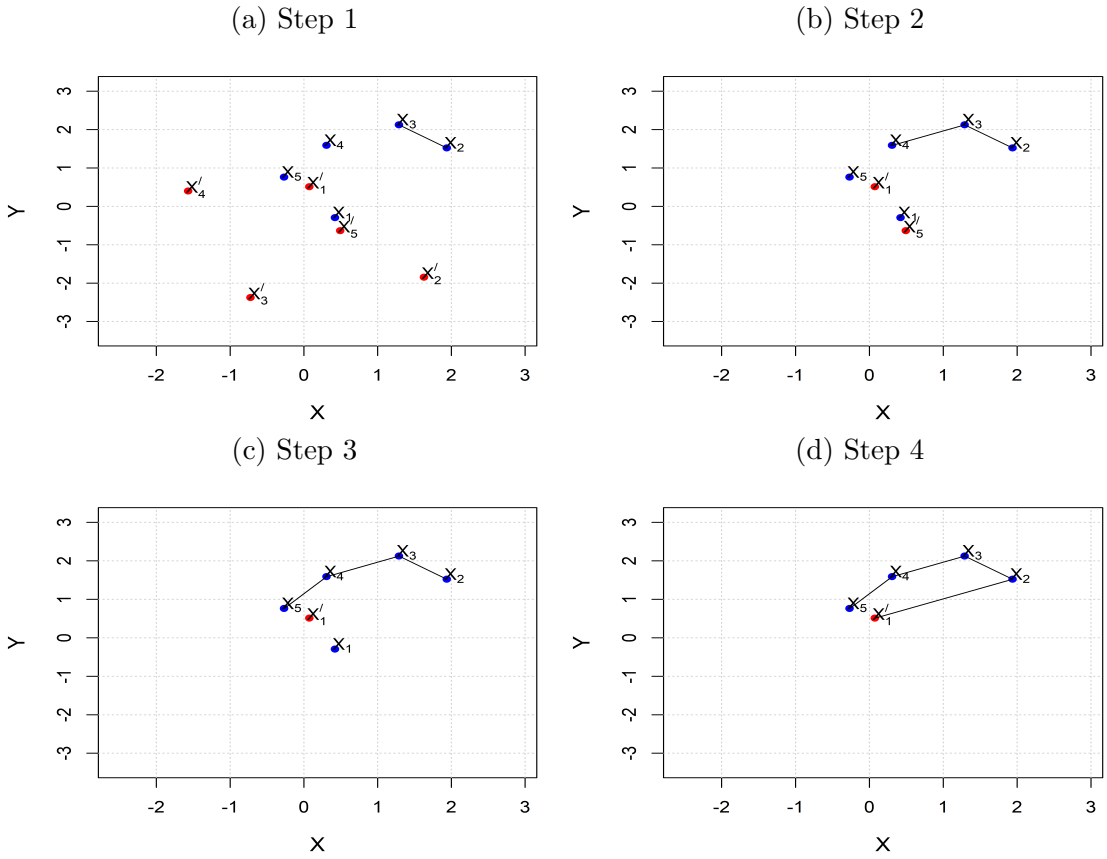


Fig. 4.3 *Algorithm for constructing the shortest covering path when* Θ *is used as the cost matrix.*

shortest covering path $\mathcal{P}$ is supposed to contain all $\mathbf{X}_i$s under the alternative. In almost all cases, $\mathcal{P}_0$ also leads to a similar path, but the arrangements of the $\mathbf{X}_i$s along these two paths differ in some cases. Therefore, though our heuristic algorithm leads to a sub-optimal solution in terms of the cost of the path, in most of the cases, $\mathcal{P}_0$ and actual $\mathcal{P}$ lead to the same values of the test statistics. This justifies the use of our heuristic algorithm for finding the shortest covering path

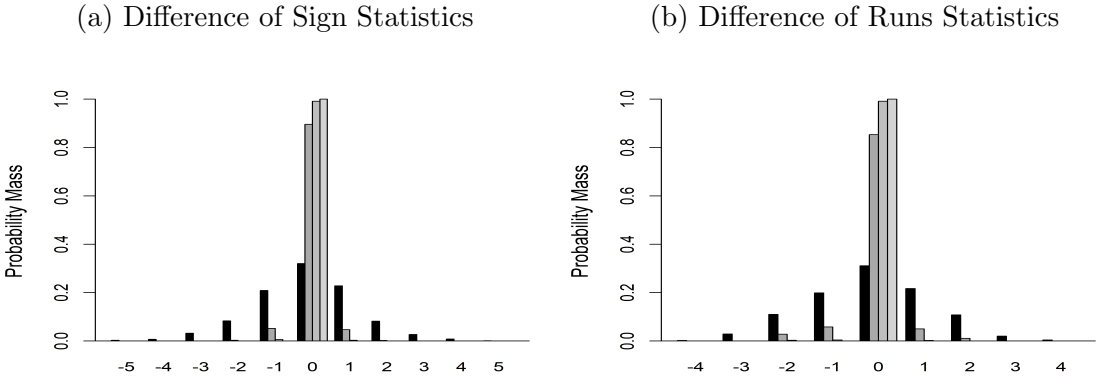


Fig. 4.4 *Barplots of the difference between (a) the sign statistic and (b) the runs statistic constructed based on* $\mathcal{P}$ *and* $\mathcal{P}_0$ *when* $\mathbf{X}_1, \ldots, \mathbf{X}_5$ *are generated independently from* $\mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}_0)$*, where* $\mathbf{\Sigma}_0 = \text{diag}(d, 1, 1, \ldots, 1)$ *for* $d = 3$ *(black), 30 (dark gray), 300 (gray), and 3000 (light gray).*

and computing the test statistics, especially for moderate or high-dimensional data.

4.2.2 Inner products vs. cosine similarities

Since $\|\mathbf{X}_i\| = \|\mathbf{X}'_i\|$ for all $i = 1, 2, \ldots, n$, instead of using the cost based on squared inner products $(\mathbf{X}_i^\top \mathbf{X}_j)^2$, $(\mathbf{X}_i^\top \mathbf{X}'_j)^2$ and $(\mathbf{X}_i^\top \mathbf{X}_j)^2$, one may be tempted to use a cost based on squared cosine similarities $C_{ij}^2 = \left(\frac{\mathbf{X}_i^\top \mathbf{X}_j}{\|\mathbf{X}_i\|\|\mathbf{X}_j\|}\right)^2$, $C_{ij'}^2 = \left(\frac{\mathbf{X}_i^\top \mathbf{X}'_j}{\|\mathbf{X}_i\|\|\mathbf{X}'_j\|}\right)^2$ and $C_{i'j'}^2 = \left(\frac{\mathbf{X}_i'^\top \mathbf{X}'_j}{\|\mathbf{X}'_i\|\|\mathbf{X}'_j\|}\right)^2$ $(1 \le i, j \le n)$. If the underlying distribution is spherically symmetric, these scaled versions of inner products $C_{ij}, C_{ij'}$ and $C_{i'j'}$ have the same distribution as ${\mathbf{U}_1}^\top \mathbf{U}_2$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are i.i.d. Unif($\mathcal{S}^{d-1}$). But, they have the same property for a sub-class of angular symmetric distribution (i.e., $\mathbf{X}/\|\mathbf{X}\| \overset{D}{=} -\mathbf{X}/\|\mathbf{X}\|$) that are not spherically symmetric. In such cases, the resulting test fails to reject the null hypothesis.

To demonstrate this, we consider a simple example involving an angular symmetric distribution. We generated 200 independent observations on a bivariate random vector $\mathbf{X} = R\mathbf{U}$ where $\mathbf{U} = (U_1, U_2)$ is uniformly distributed on the perimeter of the unit circle (i.e., $(U_1, U_2) = (\cos(\theta), \sin(\theta))$, where $\theta \sim U(0, 2\pi)$) and $R = R_1 \mathrm{I}\{U_1 U_2 > 0\} + \mathrm{I}\{U_1 U_2 \le 0\}$, for $R_1 \sim \mathrm{Unif}([1, 5])$ independent of $\mathbf{U}$. Using these observations, we computed T_S and T_R based on squared inner products and those based on squared cosine similarities. The boxplots in Figure 4.5 show the distribution of these sign and runs statistics based on 1000 repetitions of the experiment. Note that here R and $\mathbf{U}$ are not independent. So, the distribution of $\mathbf{X}$ is not spherically symmetric. However, the sign and runs statistics based on the squared cosine similarities could not figure it out. The distributions of these statistics were the same as their corresponding null distributions. But for our proposed cost function, the sign statistic had higher values, and the runs statistics had lower values leading to the rejection of the null hypothesis in most of the cases. It is clear that in this example, where $\mathbf{X}_i/\|\mathbf{X}_i\| \sim \mathrm{Unif}(\mathcal{S}^{d-1})$, any test based on $\mathbf{X}_i/\|\mathbf{X}_i\|$ $(i = 1, 2, \ldots, n)$ will fail to detect deviation from spherical symmetry. It only tests whether the distribution of $\mathbf{X}/\|\mathbf{X}\|$ is uniform but does not test for the independence between $\|\mathbf{X}\|$and $\mathbf{X}/\|\mathbf{X}\|$. The tests proposed by Zou et al. (2014) and Feng & Liu (2017) have a similar problem. This may be a reason why these

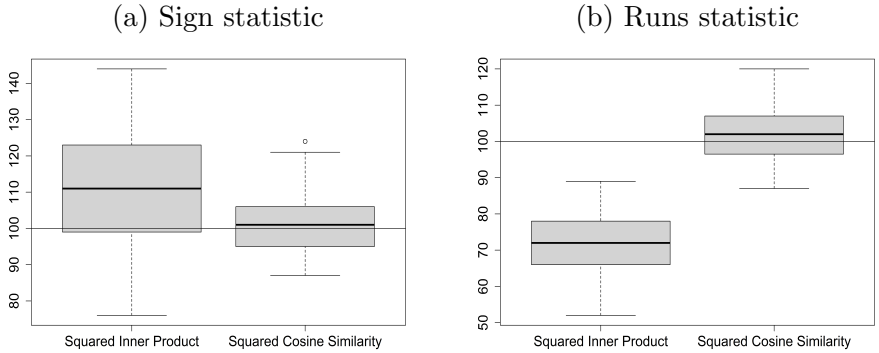


Fig. 4.5 *Boxplots showing the distributions of sign and runs statistics (based on 100 replications) when* 200 *observations are generated from angular symmetric distribution. The solid line indicates the expectation of the test statistic under spherical symmetry.*

authors proposed their tests assuming elliptic symmetry of the underlying distribution.

4.3 High dimensional behavior of the proposed tests

Let $\mathcal{D} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ be a data set consisting of n independent observations form a d-dimensional distribution P. We have already seen that if P is spherical, any function of the vectors of string signs $\boldsymbol{S}$ and string ranks $\boldsymbol{R}$ has the exact distribution-free property (see Theorem 4.1). In particular, we consider the runs statistic T_R and the sign statistic T_S. The later one can be viewed as a linear rank statistic of the form $T_{LR} = \sum_{i=1}^{n} S_{\pi_i} a(i)$, where the score function $a(\cdot)$ is given by $a(i) = 1$ or for all $i = 1, 2, \ldots, n$. From part (a) of Theorem 4.1, it is transparent that for any given score function $a(\cdot)$, the finite sample null distribution of T_{LR} is the same as the null distribution of the corresponding univariate linear rank statistic. From the description of our tests, it is also clear that they can be conveniently used for high-dimensional data even when the dimension is much larger than the sample size. In the next two sub-sections, we study the asymptotic behavior of these tests in HDLSS and HDHSS asymptotic regimes.

4.3.1 Behavior of the proposed tests for HDLSS data

Hall, Marron & Neeman (2005) showed that under certain regularity conditions on moments and weak dependence of the measurement variables, observations from a high-dimensional distribution tend to lie on the vertices of a regular simplex when the dimension diverges to infinity. Ahn et al. (2007); Jung & Marron (2009); Yata & Aoshima (2012) also provided another set of conditions based on the variance-covariance matrix for a similar geometry of high dimensional data. This geometric feature of the high-dimensional data cloud has been used extensively to study the behavior of several statistical problems. But, the HDLSS behavior of the tests of spherical symmetry is somewhat missing from the literature. There is a fundamental difficulty in this context. To understand it properly, let us first recall the following theorem by Jung & Marron (2009).

Theorem 4.2. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent copies of $\mathbf{X} \sim \mathrm{P}$, a d-dimensional distribution with the following properties.*

(A4.1) $E(\mathbf{X}) = \mathbf{0}$

(A4.2) *Let $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ be the spectral decomposition of $\boldsymbol{\Sigma} = Var(\mathbf{X})$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ is the diagonal matrix containing the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ of $\boldsymbol{\Sigma}$, and $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d]$ is the orthogonal matrix whose columns are the corresponding eigenvectors. The coordinates of $\boldsymbol{Z} = \boldsymbol{\Lambda}^{-1/2}\mathbf{U}^\top\mathbf{X}$ have uniformly bounded fourth moments and the ρ-mixing property under some permutation.*

Also, assume that $\epsilon = \left(\sum_{i=1}^{d} \lambda_i^2\right) \Big/ \left(\sum_{i=1}^{d} \lambda_i\right)^2 \to 0$ as $d \to \infty$. Then $(\sum_{i=1}^{d} \lambda_i)^{-1} S_D \xrightarrow{P} \mathbf{I}_n$, the identity matrix, as $d \to \infty$, where $S_D = ((\mathbf{X}_i^\top \mathbf{X}_j))_{1 \leq i,j \leq n}$.

The condition $\epsilon \to 0$ as $d \to \infty$ is also known as the sphericity condition. As a consequence

of Theorem 4.2, under the given conditions $(\sum_{i=1}^{d}\lambda_i)^{-1}\|\mathbf{X}_1\|^2$ and $(\sum_{i=1}^{d}\lambda_i)^{-1}(\mathbf{X}_1^\top\mathbf{X}_2)$ converges in probability to one and zero, respectively. So, the data cloud from P behaves as if they are coming from a spherical distribution. Therefore, any test of spherical symmetry based on pairwise distances or inner products has asymptotic power close to the nominal level α in high dimensions. Hence, for good performance of a test of spherical symmetry in HDLSS situations, one needs to operate outside this sphericity condition. The following theorem gives us a direction in the context.

Theorem 4.3. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be n independent copies of $\mathbf{X}$, which follows a d-dimensional non-spherical distribution* P. *Also, assume that as d diverges to infinity*

$$\mathbb{P}\left[\frac{d(\mathbf{X}_1^\top\mathbf{X}_2)^2}{\|\mathbf{X}_1\|^2\|\mathbf{X}_2\|^2} > M\right] \to 1 \quad \text{for all } M > 0. \tag{4.2}$$

Then $\boldsymbol{S}$, the sign vector, converges to $\mathbf{1}_n = (1, 1, \ldots, 1)$ in probability as d diverges to infinity.

Since $\left\{\frac{d(\mathbf{X}_1^\top\mathbf{X}_2')^2}{\|\mathbf{X}_1\|^2\|\mathbf{X}_2'\|^2}\right\}_{d\geq 1}$ and $\left\{\frac{d(\mathbf{X}_1'^\top\mathbf{X}_2')^2}{\|\mathbf{X}_1'\|^2\|\mathbf{X}_2'\|^2}\right\}_{d\geq 1}$ are two tight sequences of random variables (see the proof of Theorem 4.3), condition (4.2) ensures that $\theta(\mathbf{X}_1, \mathbf{X}_2)$ becomes smaller than $\theta(\mathbf{X}_1, \mathbf{X}_2')$ and $\theta(\mathbf{X}_1', \mathbf{X}_2')$ with probability tending to 1 as d grows to infinity. One can show that the condition (4.2) holds for the spiked covariance model considered in Jung *&* Marron (2009). So, as a corollary, we have the following result.

Corollary 4.1. *Let $\mathbf{X}_1, \mathbf{X}_2$ be two independent random variables from a d-dimensional distribution* P *satisfying (A4.1) and (A4.2) mentioned in Theorem 4.2. Also, assume that*

(a) $\lambda_1/d^\kappa \to c_1$ for some $\kappa \geq 1$ and $c_1 > 0$,

(b) $\sum_{i=2}^{d}\lambda_i^2/(\sum_{i=2}^{d}\lambda_i)^2 \to 0$ as $d \to \infty$ and $\sum_{i=2}^{d}\lambda_i = \mathcal{O}(d)$.

Then $\boldsymbol{S}$ converges to $\mathbf{1}_n = (1, 1, \ldots, 1)$ in probability as d diverges to infinity.

As a consequence of Theorem 4.3, for any given sequence of scores $\{a(i)\}_{1\leq i\leq n}$, $T_{LR} = \sum_{i=1}^{n} S_{\pi_i}a(i)$ converges to $\sum_{i=1}^{n}a(i)$ in probability as d diverges to infinity. So, if these scores $\{a(i)\}_{1\leq i\leq n}$ are non-negative, which is usually the case, T_{LR} takes its largest value with probability tending to 1 as d diverges to infinity. In the case of the sign statistic, $P(T_S = n) \to 1$ as $d \to \infty$. Now, under H_0, we have $P(T_S \geq n) = 1/2^n$. So, for any fixed level α $(0 < \alpha < 1)$, unless the sample size is very small (i.e., $2^n < 1/\alpha$), the power of the proposed sign test converges to 1 as the dimension increases. Similarly, under the condition of Theorem 4.3, the runs statistic T_R converges to 1 in probability. Now, under H_0, we have $P(T_R \leq 1) = 1/2^{n-1}$. So, if $2^{n-1} > 1/\alpha$, we have the consistency of the proposed runs test of level α in the HDLSS asymptotic regime.

Now, we consider three simple examples involving normal distributions to study the empirical performance of the proposed tests in high dimensions when the nominal levels of the tests are taken as 0.05.

Example 4.1. *We consider a d-variate normal distribution with mean $\mathbf{0}_d$ and variance covariance*

matrix $\mathbf{\Sigma} = ((\sigma_{ij}))$, *where* σ_{ij} *is* 1 *if* $i = j$ *and* 0.6 *if* $i \neq j$.

Example 4.2. *Here we consider a d-variate normal distribution with mean* $\mathbf{0}_d$ *and a diagonal variance covariance matrix* $\mathbf{\Sigma} = ((\sigma_{ij}))$, *where* $\sigma_{ii} = 1$ *for* $1 \leq i \leq [d/2]$ *and* $\sigma_{ii} = 2$ *for* $i \geq [d/2]+1$.

Example 4.3. *Here also, we deal with a d-variate normal distribution with mean* $\mathbf{0}_d$ *and a diagonal variance-covariance matrix. The matrix has the first diagonal element d and the rest equal to* 1.

For each example, we considered 10 choices of d ($d = 2^i$ for $i = 1, 2, \ldots, 10$), but a fixed value of n ($n = 50$). Each experiment was repeated 500 times, and the empirical power of a test was computed as the proportion of times it rejected H_0. The results are reported in Figure 4.6.

One can check that the normal distributions in Examples 4.1 and 4.3 satisfy conditions (a) and (b) of Corollary 4.1. So, as expected, the powers of the sign and runs tests sharply raised to 1. But in Example 4.2, we had a diametrically opposite picture, where both sign and runs tests failed to have satisfactory performance. Note that in this example, the sphericity condition (see Theorem 4.2) is satisfied, which could be the reason behind the poor performance of these tests.

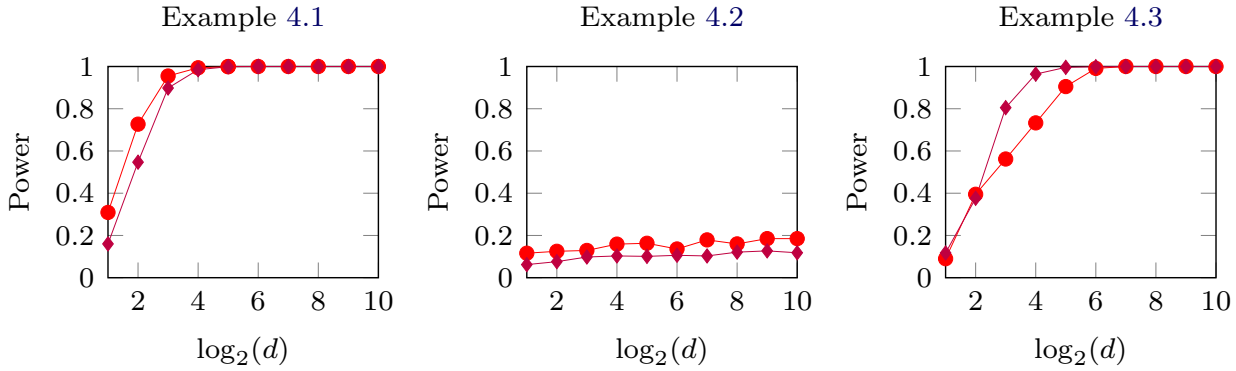


Fig. 4.6 *Powers of the sign test (●) and the runs test (◆) when* 50 *observations are generated from the d-variate normal distributions (with* $d = 2^i, i = 1, 2, \ldots, 10$*) considered in Examples 4.1-4.3*

4.3.2 Behavior of the proposed tests in HDHSS asymptotic regime

In this section, we investigate the behavior of the proposed tests for high-dimensional data sets having a large number of observations. This type of data set commonly arises in many areas of sciences, including biology, ecology, and medical sciences. Here, we study the asymptotic behavior of the tests when the dimension and the sample size grow simultaneously, but their divergence rates are arbitrary. Since the null distributions of T_{LR} (or T_S in particular) and T_R do not depend on the dimension of the data, their limiting null distributions in the HDHSS regime remain the same as they are in the classical asymptotic regime. The asymptotic null distribution of the linear rank statistic T_{LR} is given by the following theorem.

Theorem 4.4. *Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *be independent realizations of a d-dimensional random vector* $\mathbf{X} \sim \mathrm{P}$. *Assume that* P *is spherically symmetric and the sequence of scores* $\{a(i)\}_{1 \leq i \leq n}$ *satisfies*

the following conditions

$$\sum_{i=1}^{n} a^2(i) \to \infty \quad \text{and} \quad \max_{1\le i\le n} \frac{a^2(i)}{\sum_{i=1}^{n} a^2(i)} \to 0 \quad \text{as} \quad n \to \infty. \tag{4.3}$$

Then, as n and d both grow to infinity, we have

$$\frac{T_{LR} - \frac{1}{2}\sum_{i=1}^{n} a(i)}{\sqrt{\sum_{i=1}^{n} a^2(i)}} \xrightarrow{D} \mathcal{N}_1\left(0, \frac{1}{4}\right).$$

In particular we have $n^{-1/2}(T_S - n/2) \xrightarrow{D} \mathcal{N}(0, 0.25)$. Theorem 4.4 holds even when d is fixed and n diverges to infinity. So, irrespective of the value of d, when n is large, this test can be calibrated using the quantiles of the normal distribution. Now, we investigate the asymptotic behavior of T_{LR} for non-spherical distribution. In this context, we have the following result.

Theorem 4.5. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be n independent copies of a d-dimensional random vector $\mathbf{X} \sim$ P. Assume that the sequence of scores $\{a(i)\}_{1\le i\le n}$ satisfies condition (4.3) and as $n \to \infty$,*

$$\frac{\sum_{i=1}^{n-1} a(i)a(i+1)}{\sum_{i=1}^{n} a^2(i)} \to C, \tag{4.4}$$

for some $C > 0$. If P *is not spherical, then there exist finite constants σ_{11} and σ_{12}. such that*

$$\limsup_{n,d\to\infty} Var\left[\frac{T_{LR} - \mathbb{E}[T_{LR}]}{\sqrt{\sum_{i=1}^{n} a^2(i)}}\right] = \sigma_{11} + 2C\sigma_{12}.$$

Remark 4.1. *For any fixed d, one can derive the large sample distribution of T_{LR} against a sequence of contiguous alternatives (see Chapter 12 Lehmann & Romano, 2021 for contiguous alternatives) and prove its Pitman efficiency (see Section 4.7.2). This is in sharp contrast to the results in Bhattacharya (2019), where the author proved that in multivariate setup, most of the graph-based distribution-free two-sample tests turn out to be inefficient in the classical asymptotic regime.*

Now from Theorems 4.4 and 4.5, we have $[T_S - \mathbb{E}(T_S)]/\sqrt{n} = O_p(1)$ both under the null and alternative hypotheses. Hence, we have $|T_S - \mathbb{E}(T_S)|/n \xrightarrow{P} 0$. One can also show that

$$\left|\frac{1}{n}\mathbb{E}(T_S) - \mathbb{P}\left[\theta(\boldsymbol{Y}_{S_1,1}, \mathbf{X}_2) + \theta(\mathbf{X}_2, \boldsymbol{Y}_{S_3,3}) \le \theta(\boldsymbol{Y}_{S_1,1}, \mathbf{X}_2') + \theta(\mathbf{X}_2', \boldsymbol{Y}_{S_3,3})\right]\right| \to 0 \text{ as } n, d \to \infty,$$

where $\boldsymbol{Y}_i = S_i\mathbf{X}_i + (1 - S_i)\mathbf{X}_i'$ for $i = 1, 2, 3$ (see Lemma A4.3 for the proof). Let us define

$$p_S = \liminf_{d\to\infty} \mathbb{P}\left[\theta(\boldsymbol{Y}_{S_1,1}, \mathbf{X}_2) + \theta(\mathbf{X}_2, \boldsymbol{Y}_{S_3,3}) \le \theta(\boldsymbol{Y}_{S_1,1}, \mathbf{X}_2') + \theta(\mathbf{X}_2', \boldsymbol{Y}_{S_3,3})\right]$$

as the limiting value of the probability of inclusion of any $\mathbf{X}_i$ $(i = 1, 2, \ldots, n)$ in $\mathcal{P}$. Under H_0, because of the exchangeability of $\mathbf{X}_i$ and $\mathbf{X}_i'$, p_S turns out to be 0.5. However, in view of Theorem 4.3 and Lemma 4.2, under H_1, we expect $\theta(\mathbf{X}_1, \mathbf{X}_2)$ to be stochastically smaller than $\theta(\mathbf{X}_1, \mathbf{X}_2')$ and $\theta(\mathbf{X}_1', \mathbf{X}_2')$. So, p_S is expected to be higher than 0.5. Since we reject H_0 for higher values of T_S, under the condition $p_S > 0.5$, the consistency of this sign test follows from Theorem 4.5.

Since the null distribution of T_R does not depend on d, its limiting null distribution in the HDHSS setup is the same as that of the univariate runs statistic. It is given by Theorem 4.6.

Theorem 4.6. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be n independent realizations of a d-dimensional random vector $\mathbf{X} \sim \mathrm{P}$. If P is spherically symmetric, as $n, d \to \infty$, we have*

$$\frac{T_R - (n+1)/2}{\sqrt{n}} \xrightarrow{D} \mathcal{N}_1\left(0, \frac{1}{4}\right).$$

This asymptotic null distribution of T_R remains the same even for any fixed d as n grows to infinity. So, when the sample size is large, whatever be the dimension of the data, this runs test can be calibrated using the quantiles of a Gaussian distribution, and for any fixed nominal level α, the cut-off remains the same for all values of d. Now, one may be curious to know about the asymptotic behavior of T_R for non-spherical distributions when n and d both diverge to infinity. This is specified in the following theorem.

Theorem 4.7. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be n independent copies of a d-dimensional random vector $\mathbf{X} \sim \mathrm{P}$. If P is non-spherical, there exists a finite positive constant σ^2 such that*

$$\limsup_{n,d\to\infty} Var\left[\frac{T_R - \mathbb{E}[T_R]}{\sqrt{n}}\right] = \sigma^2.$$

Also, if the condition (4.2) is satisfied, T_R converges in probability to one as d diverges to infinity.

Theorem 4.7 shows that $|T_R - \mathbb{E}(T_R)|/n \xrightarrow{P} 0$ as $n, d \to \infty$. We know that under H_0, the limiting value of $\mathbb{E}(T_R)/n$ and hence that of T_R/n is 0.5 (follows from Theorem 4.6). But, as we have observed before, under the alternative, most of the string signs are expected to be one, and as a result, $p_R = \limsup_{n,d\to\infty} \mathbb{E}(T_R/n)$ is expected to be small. Since we reject H_0 for small values of T_R, when $p_R < 0.5$, the runs test turns out to be consistent in the HDHSS regime.

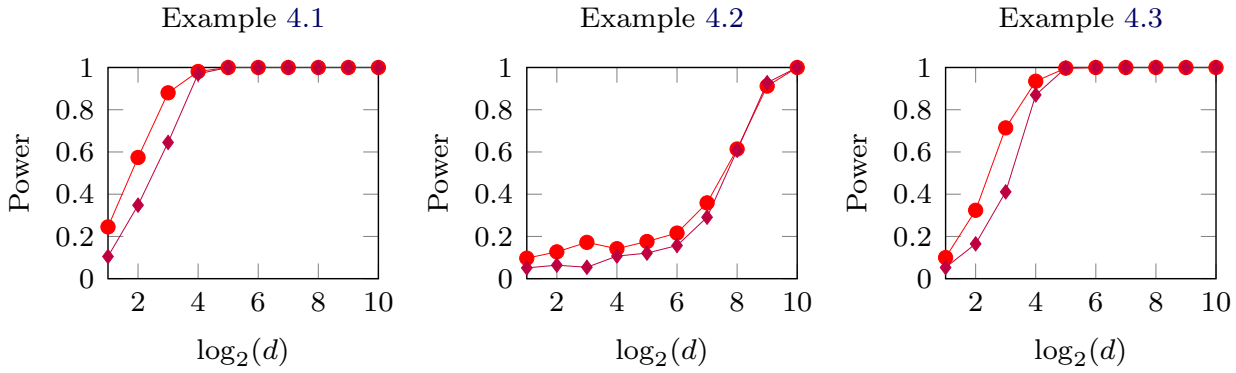


Fig. 4.7 *Powers of the sign test (●) and the runs test (◆) when $n = d + 20$ observations are generated from the d-variate normal distributions (with $d = 2^i, i = 1, 2, \ldots, 10$) considered in Examples 4.1-4.3.*

We studied the performance of sign and runs tests in Examples 4.1-4.3 for 10 values of d (i.e., $d = 2^i$ for $i = 1, 2, \ldots, 10$) as in Section 4.3.1, but this time instead of considering a fixed value of n, we took $n = d + 20$ so that it also grows with the dimension. Each example was repeated 1000 times to compute the empirical powers of the tests as before. In these three examples, we have p_S larger than 0.5 (1, 0.577 and 1, respectively) and p_R smaller than 0.5 (0, 0.422 and 0, respectively). So, as expected, in all these cases, sign and runs tests performed well for large values of n and d

(see Figure 4.7). In Examples 4.1 and 4.3, the powers of these tests raised sharply as before. But unlike what happened in the HDLSS setup, the powers of these tests increased to 1 in Example 4.2 as well. Since we get more information as the sample size increases, such results are expected.

4.4 Further modifications of the proposed tests

Recall that due to the sphericity condition in the HDLSS setup, our proposed sign and runs tests failed to have satisfactory performance in Example 4.2 even though the diagonal elements of the variance-covariance matrix were different. This motivated us to look for further modifications of these tests. Note that the performance of these tests depends on the construction of the shortest covering path using Θ as the cost matrix. One can notice that the ordering of $\theta(\mathbf{X}_1, \mathbf{X}_2)$, $\theta(\mathbf{X}_1, \mathbf{X}_2')$ and $\theta(\mathbf{X}_1', \mathbf{X}_2')$ depends on that of $(\mathbf{X}_1^\top \mathbf{X}_2)^2$, $(\mathbf{X}_1^\top \mathbf{X}_2')^2$ and $(\mathbf{X}_1'^\top \mathbf{X}_2')^2$. Now, we can break $(\mathbf{X}_1^\top \mathbf{X}_2)^2$ into two parts containing square terms and cross-product terms as

$$(\mathbf{X}_1^\top \mathbf{X}_2)^2 = \sum_{i=1}^{d} X_{1i}^2 X_{2i}^2 + \sum_{1 \le i \ne j \le d} X_{1i} X_{2i} X_{1j} X_{2j}.$$

If $\mathbf{X}_1$ and $\mathbf{X}_2$ are d-variate i.i.d. random variables with mean $\mathbf{0}_d$ and dispersion matrix $\mathbf{\Sigma} = ((\sigma_{ij}))$, expectations of these two parts are $a_d = \sum_{i=1}^{d} \sigma_{ii}^2$ and $c_d = \sum_{i=1}^{d} \sum_{j(\ne i)=1}^{d} \sigma_{ij}^2$, respectively. However, for $(\mathbf{X}_1^\top \mathbf{X}_2')^2$ (and also for $(\mathbf{X}_1'^\top \mathbf{X}_2')^2$), these two expected values are $b_d = (\sum_{i=1}^{d} \sigma_{ii})^2/d$ and 0, respectively (see Lemma A4.2). So, the differences turn out to be $v_d = a_d - b_d = \sum_{i=1}^{d} [\sigma_{ii} - (\frac{1}{d}\sum_{i=1}^{d} \sigma_{ii})]^2$ and c_d, respectively. While the first one measures the variation among the diagonal elements of $\mathbf{\Sigma}$, the second one tells us how different the off-diagonal elements of $\mathbf{\Sigma}$ are from 0. In Example 4.1, where we get a signal from the second part (i.e., $c_d > 0$), the proposed tests worked well. But, in Example 4.2, we have no signal from the second part. So, the difference in cross-products serves as a noise, and its order $\mathcal{O}_P(d^2)$ is higher than that of the signal $v_d = \mathcal{O}(d)$ obtained from the difference in square terms. Therefore, the proposed test could not have satisfactory performance. In Example 4.3 also, the difference in cross-products serves as a noise of order $\mathcal{O}_P(d^2)$, but here the signal v_d is of the order $O(d^2)$. So, the powers of sign and runs tests increased with the dimension. However, from the above discussion, it is clear that if there is no signal from the off-diagonal part (i.e., $c_d = 0$), our tests can perform better if we can get rid of this noise term involving cross-products. One possible option is to ignore the cross-product terms and construct the shortest covering path based on a different cost function. We can consider an edge-weighted complete graph $\mathcal{K}_{2n}$ on $2n$ vertices' $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_{2n}$ as before but use $\widetilde{\theta}(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \exp\{-\frac{1}{d}\sum_{q=1}^{d} Z_{iq}^2 Z_{jq}^2\}$ as the cost of the edge joining $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$ $(1 \le i < j \le 2n)$. Using this cost function, we can construct the shortest covering path as before and define the sign and rank vectors along that path as in Section 4.1. So, we look for new sign and rank vectors $\widetilde{\boldsymbol{S}}$ and $\widetilde{\boldsymbol{R}} = \widetilde{\mathbf{\Pi}}^{-1}$, where

$$(\widetilde{\boldsymbol{S}}, \widetilde{\mathbf{\Pi}}) = \underset{\substack{\boldsymbol{s} \in \{0,1\}^n \\ \boldsymbol{\pi} \in \mathcal{S}_n}}{\arg\min} \Big[\sum_{i=1}^{n-1} \widetilde{\theta}(\boldsymbol{Y}_{s_{\pi_i}, \pi_i}, \boldsymbol{Y}_{s_{\pi_{i+1}}, \pi_{i+1}}) \Big]. \tag{4.5}$$

It is easy to check that the null distribution of $(\widetilde{\boldsymbol{S}}, \widetilde{\boldsymbol{R}})$ matches with that of $(\boldsymbol{S}, \boldsymbol{R})$ (see part (a) of Theorem 4.8). So, the null distributions of the corresponding sign statistic $\widetilde{T}_S = \sum_{i=1}^n \widetilde{S}_i$ (or any linear rank statistic $\widetilde{T}_{LR} = \sum_{i=1}^n \widetilde{S}_i a(\widetilde{R}_i) = \sum_{i=1}^n \widetilde{S}_{\widetilde{\pi}_i} a(i)$) and runs statistic $\widetilde{T}_R = 1 + \sum_{i=1}^{n-1} \mathrm{I}[\widetilde{S}_{\widetilde{\pi}_i} \neq \widetilde{S}_{\widetilde{\pi}_{i+1}}]$ match with the corresponding univariate statistics, and the cut-offs can be obtained as before from the statistical tables available for the univariate sign (or linear rank) and runs tests.

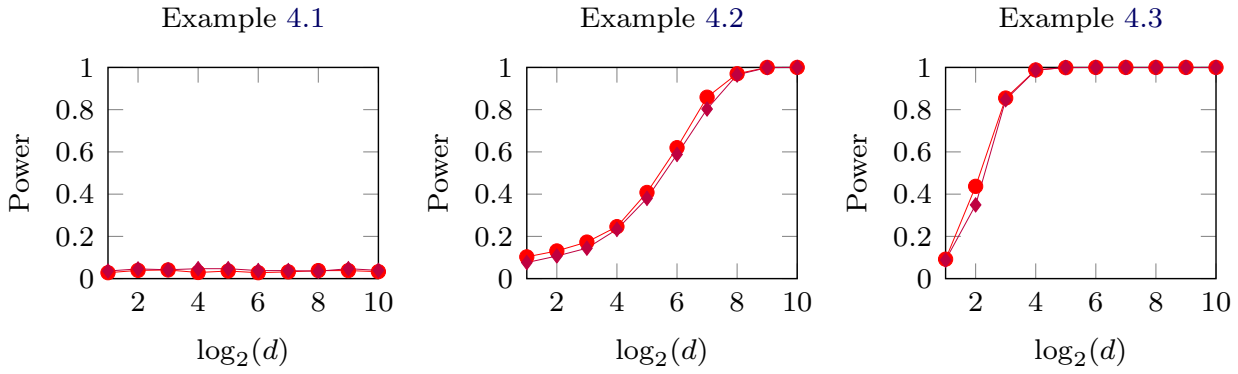


Fig. 4.8 *Powers of the tests based on* $\widetilde{T}_S$ (●) *and* $\widetilde{T}_R$ (◆) *when* 50 *observations are generated from the d-variate normal distributions (with* $d = 2^i$ *for* $i = 1, 2, \ldots, 10$*) considered in Examples 4.1-4.3.*

Figure 4.8 shows the performance of the sign and runs tests based on $\widetilde{T}_S$ and $\widetilde{T}_R$ in Examples 4.1-4.3. In Example 4.3, where the signal comes from the diagonal part, they performed better than our previous sign and runs tests. They also performed well in Example 4.2, where the previous sign and runs tests did not have satisfactory performance. However, in Example 4.1, where we do not have signals from the diagonal part, this new sign and runs tests had poor performance.

This high-dimensional behavior of the tests based on $\widetilde{T}_S$ and $\widetilde{T}_R$ can be further explained by part (b) of Theorem 4.8. But for that, we need the following technical assumption.

(A4.3) Let $\mathbf{X}_1, \mathbf{X}_2$ be two independent copies of $\mathbf{X} \sim \mathrm{P}$ and $\mathbf{X}_1', \mathbf{X}_2'$ be their spherically symmetric variants. There exists an $\gamma > 0$ such that for $W = d^{-\gamma} \sum_{i=1}^d (\mathbf{X}_1)_j^2 (\mathbf{X}_2)_j^2$, $d^{-\gamma} \sum_{i=1}^d (\mathbf{X}_1)_j^2 (\mathbf{X}_2')_j^2$ and $d^{-\gamma} \sum_{i=1}^d (\mathbf{X}_1')_j^2 (\mathbf{X}_2')_j^2$, $\big|W - \mathbb{E}[W]\big| \xrightarrow{P} 0$ as $d \to \infty$, and at least one of the limits is non-zero.

Similar assumptions were also considered by Hall, Marron *&* Neeman (2005); Jung *&* Marron (2009); Yata *&* Aoshima (2012); Sarkar *&* Ghosh (2020); Banerjee *&* Ghosh (2025); Dutta, Sarkar *&* Ghosh (2016) for studying high dimensional behaviour of different statistical methods. This assumption is satisfied in Examples 4.2 and 4.3 for $\gamma = 1$ and $\gamma = 2$, respectively. Under this assumption, we have the following result.

Theorem 4.8. *Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *be* n *independent realizations* $\mathbf{X}$ *following a* d*-dimensional continuous distribution* P*.*

(a) If P *is spherically symmetric,* $\widetilde{\boldsymbol{S}} \sim \mathrm{Unif}(\{0,1\}^n)$*,* $\widetilde{\boldsymbol{R}} \sim \mathrm{Unif}(\mathcal{S}_n)$*, and they are independent.*

(b) Let $\mathbf{\Sigma}$ *denote the covariance matrix of* $\mathbf{X}$ *and* $\mathbf{D} = \text{diag}(\mathbf{\Sigma})$. *Suppose that* P *is not spherically symmetric, and it satisfies Assumption (A4.3). If*

$$\liminf_{d\to\infty} \left\{ \frac{1}{d^{\gamma}} \text{trace}(\mathbf{D}^2) - \frac{1}{d^{1+\gamma}} \Big(\text{trace}(\mathbf{D}) \Big)^2 \right\} > 0, \tag{4.6}$$

the sign vector $\widetilde{\boldsymbol{S}}$ *converges to* $\mathbf{1}_n = (1, 1, \ldots, 1)$ *in probability as* d *diverges to infinity.*

Note that using Jensen's inequality we have $d\ \text{trace}(\mathbf{D}^2) \geq (\text{trace}(\mathbf{D}))^2$, or $\frac{1}{d^{\gamma}}\text{trace}(\mathbf{D}^2) - \frac{1}{d^{1+\gamma}} \Big(\text{trace}(\mathbf{D}) \Big)^2 \geq 0$, where the equality holds if and only if all diagonal elements of $\mathbf{D}$ are equal. So, the condition (4.6) holds when the variance among the diagonal elements of $\mathbf{D}$ remains bounded away from 0 as the dimension grows to infinity. This condition is satisfied in Examples 4.2 and 4.3 but not in Example 4.1. This explains the difference in the performance of the tests based on $\widetilde{T}_S$ and $\widetilde{T}_R$ in these three examples.

Therefore, in HDLSS setup, while the tests based on T_S and T_R may fail to detect weak signals in the diagonal part (for instance, in Example 4.2, where all diagonal elements of the variance-covariance matrix are not same but the sphericity condition holds), those based on $\widetilde{T}_S$ and $\widetilde{T}_R$ cannot detect signals present in the off-diagonal part (for instance in Example 4.1, where we have non-zero off-diagonals elements). To overcome these limitations, one can think of combining the strengths of these two types of tests. For instance, we can use $T_S^M = \max\{T_S, \widetilde{T}_S\}$ and $T_R^M = \min\{T_R, \widetilde{T}_R\}$ as test statistics to boost the performance of the sign and runs tests for a larger class of alternatives. Naturally, we reject the null hypothesis of spherical symmetry for large values of the modified sign statistic T_S^M or small values of the modified runs statistic T_R^M. The cut-offs can be computed using an appropriate resampling method (as in Chapter 3), but to keep our tests simple and computationally efficient, here we use Bonferroni's method for calibration. Note that under the null hypothesis, T_S and $\widetilde{T}_S$ (respectively, T_R and $\widetilde{T}_R$) are identically distributed. So, the cut-off for T_S^M (respectively, T_R^M) can be easily obtained from statistical tables and packages. One can use tests based on $T_S^S = T_S + \widetilde{T}_S$ and $T_R^S = T_R + \widetilde{T}_R$ as well, but in those cases, Bonferroni's

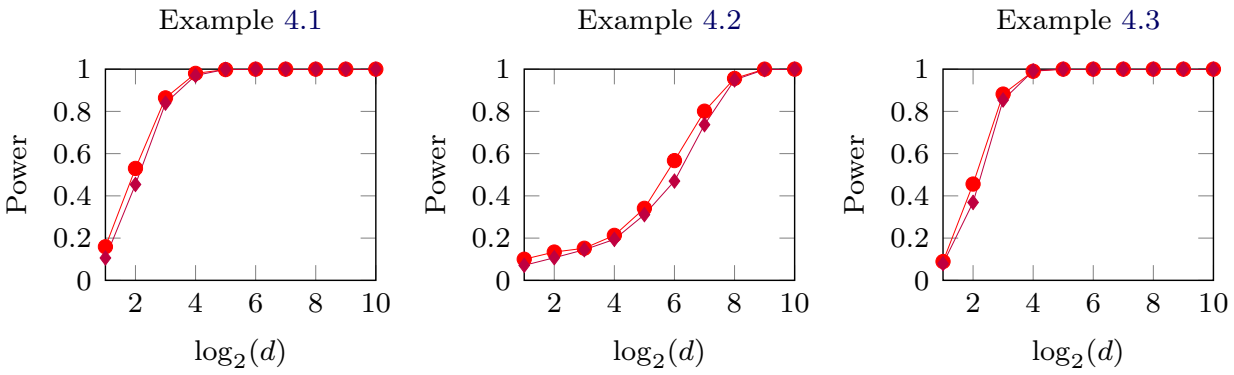


Fig. 4.9 *Powers of the modified sign test (●) and the modified runs test (◆) when* 50 *observations are generated from the* d*-variate normal distributions (with* $d = 2^i$ *for* $i = 1, 2, \ldots, 10$*) in Examples 4.1-4.3.*

method cannot be used, and the user needs to go for the resampling algorithm for calibration. So, these tests become computationally expensive, and here, we do not consider them.

Figure 4.9 shows the performance of the modified sign and runs tests in Examples 4.1-4.3. In all three examples, they had excellent performance. So, it seems reasonable to use T_S^M or T_R^M as the test statistic. The following theorem shows the consistency of these tests in the HDLSS regime.

Theorem 4.9. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be n independent realizations of $\mathbf{X}$, which follows a d-dimensional distribution* P *which is not spherically symmetric. Let $\mathbf{\Sigma}$ be the variance-covariance matrix of $\mathbf{X}$ and define $\boldsymbol{D} = \text{diag}(\mathbf{\Sigma})$. Also, assume one of the following conditions.*

(a) For all $M > 0$, $\mathbb{P}\left[\frac{d(\mathbf{X}_1^\top \mathbf{X}_2)^2}{\|\mathbf{X}_1\|^2\|\mathbf{X}_2\|^2} > M\right] \to 1$ as d diverges to infinity,

(b) Assumption (A4.3) holds and $\liminf_{d\to\infty}\left\{\frac{1}{d^\gamma}\text{trace}(\mathbf{D}^2) - \frac{1}{d^{1+\gamma}}\Big(\text{trace}(\mathbf{D})\Big)^2\right\} > 0$.

Then T_S^M converges in probability to its maximum value n and T_R^M converges in probability to its minimum value 1 as d diverges to infinity.

Remark 4.2. *Under H_0, we have $P(T_S^M \geq n) \leq P(T_S = n) + P(\widetilde{T}_S = n) = 1/2^{n-1}$. So, for any fixed level α ($0 < \alpha < 1$), if n exceeds $-\log_2(\alpha)+1$, the power of the modified sign test converges to 1 as d increases. Under H_0, we also have $P(T_R^M \leq 1) \leq P(T_R = 1) + P(\widetilde{T}_R = 1) = 1/2^{n-2}$. So, for the consistency of the modified runs test in the HDLSS asymptotic regime, we need $n > -\log_2(\alpha)+2$.*

From our discussion, it is clear that if the variance-covariance matrix of the underlying distribution differs from a scalar multiple of the identity matrix, our modified sign and runs tests can work well in the HDLSS regime. Now, one may be curious to know what happens if the underlying distribution is not spherically symmetric, but the variance-covariance matrix is a scalar multiple of the identity matrix, for instance, if the coordinate variables are i.i.d. but the distribution is not spherical. To investigate it, we consider two simple examples.

Example 4.4. *Here we deal with a uniform distribution on the d-dimensional hypercube $[-1, 1]^d$.*

Example 4.5. *The distribution of $\mathbf{X} = (X_1, \ldots, X_d)$ has i.i.d. Laplace $(0, 1)$ coordinate variables.*

For each of these examples, we consider a sample of size 50 and use six different values of d ($d = 2^i$ for $i = 1, \ldots, 6$). Each experiment was carried out 1000 times to compute the empirical powers of different tests, and they are reported in Figure 4.10.

The sign test and the runs test based on T_S and T_R had very poor performance in these examples. In Example 4.4, the modified sign test also failed, but the modified runs test had an excellent performance. However, in Example 4.5, the modified sign test performed well. The power of the modified runs test also increased with the dimension, though the rate of increment was relatively slower. This result shows that the proposed modification substantially improves the performance of the sign and runs tests in some cases.

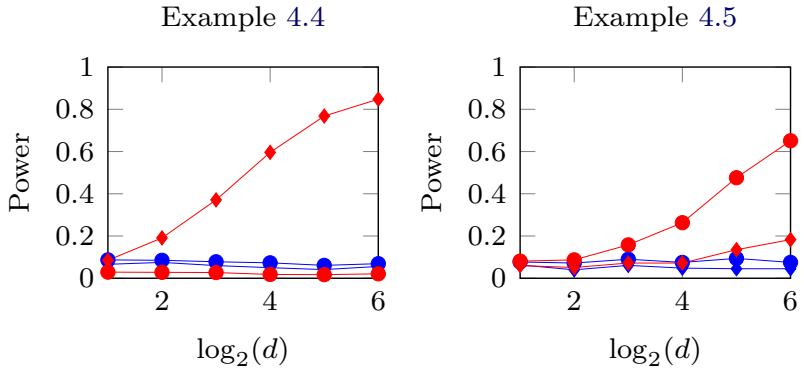


Fig. 4.10 *Powers of sign test (●), runs test, (◆), modified sign test (●), and modified runs test (◆) in Examples 4.4 and 4.5 for $n = 50$ and $d = 2^i$ for $i = 1, \ldots, 6$.*

To understand the reason behind this behavior of modified sign and runs tests, we look at the distributions of $\widetilde{\theta}(\mathbf{X}_i, \mathbf{X}_j)$, $\widetilde{\theta}(\mathbf{X}_i, \mathbf{X}'_j)$ and $\widetilde{\theta}(\mathbf{X}'_i, \mathbf{X}'_j)$s, which are shown in Figure 4.11 for $d = 1000$. From our discussions at the beginning of this section, one can show that under Assumption (A4.3), all of them converge to the same value as d increases. But Figure 4.11 shows us an interesting phenomenon. Note that the left tails of these distributions play an important role in our methods as the shortest covering path construction algorithm starts the pair of observations corresponding to the smallest value of $\widetilde{\theta}$, and then the other pairs are joined subsequently. Figure 4.11 shows that in Example 4.4, we are likely to start with a pair of the form $(\mathbf{X}'_i, \mathbf{X}'_j)$, and subsequently join more observations from the set $\{\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_n\}$. As a result, both the runs statistic and the sign statistic are expected to take smaller values. So, the test based on $\widetilde{T}_R$ and hence the modified runs test worked well, but the sign test that rejects H_0 for larger values of $\widetilde{T}_S$ performed poorly, and so did the modified sign test. But we observed an opposite picture in Example 4.5. Here, the shortest covering path is likely to start with a pair of the form $(\mathbf{X}_i, \mathbf{X}_j)$, and we are expected to

(a) Example 4.4 (b) Example 4.5

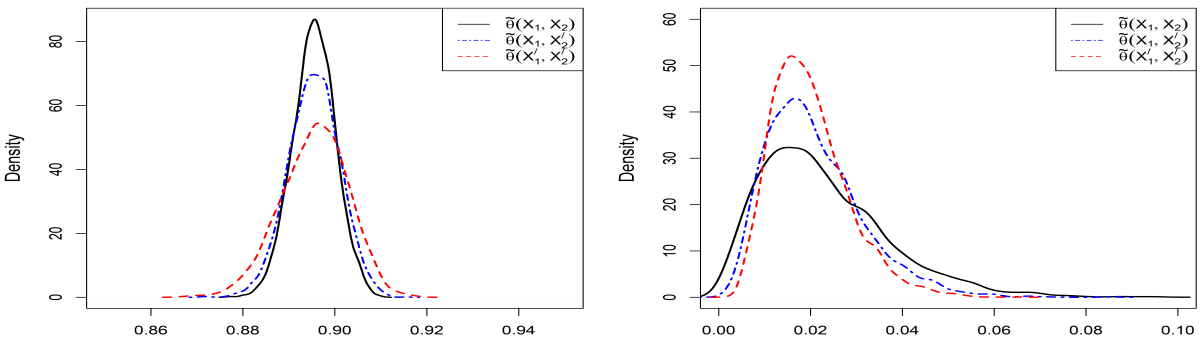


Fig. 4.11 *The density estimates of $\tilde{\theta}(.,.)$ when $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are generated independently as described in Examples 4.4 and 4.5.*

have a dominance of the original observations on the path. So, the test based on $\widetilde{T}_S$ and hence the modified sign test performed well. The power of the modified runs test also showed an increasing trend, but its performance was relatively inferior compared to the modified sign test.

However, the powers of all these tests showed increasing trends when the sample size increased with the dimension (see Figure 4.12). In both examples, modified sign and runs tests had much better performance than what we observed in Figure 4.10. We also observed the same for the tests based on T_S and T_R. We have already studied the HDHSS behavior of these two tests in Section 4.3.2. Now, we briefly investigate the HDHSS behavior of our modified tests.

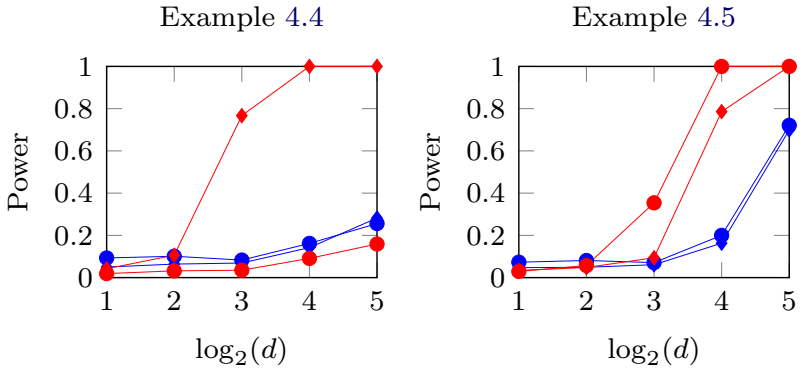


Fig. 4.12 *Powers of sign test (●), runs test, (◆), modified sign test (●) and modified runs test (◆) in Examples 4.4 and 4.5 for* $n = d^2 + 20$ *and* $d = 2^i$ *for* $i = 1, \ldots, 5$.

Note that the null distributions of $\widetilde{T}_S$ and $\widetilde{T}_R$ are identical to those of T_S and T_R, respectively. Theorem 4.4 and Theorem 4.6 give the asymptotic null distributions of T_S and T_R when n and d both tend to infinity (these results hold even when d is fixed and n diverges to infinity). One can show that $\widetilde{T}_S$ and $\widetilde{T}_R$ have the same asymptotic behavior. However, studying the asymptotic behavior of the modified tests turns out to be a bit complicated due to the dependence between these two sign statistics and that between two runs statistics. The following theorem summarizes the asymptotic null behavior of the modified test statistics.

Theorem 4.10. *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be independent realizations of a* d*-dimensional random vector* $\mathbf{X} \sim \mathrm{P}$. *If* P *is spherically symmetric, we have the following results.*

(a) Assume that $\sigma_s^2 = \lim_{d\to\infty} \mathbb{P}[S_1 = 1; \tilde{S}_1 = 1]$ *exists. Then as* n *and* d *grow to infinity, we have* $n^{-1/2}\Big(\max\{T_S, \tilde{T}_S\} - \frac{n}{2}\Big) \xrightarrow{D} \max\{Z_1, Z_2\}$.*where* (Z_1, Z_2) *follows a bivariate normal distribution with mean* $\mathbf{0}_2$*, equal variances* $\frac{1}{4}$ *and covariance* $\big(\sigma_s^2 - \frac{1}{4}\big)$.

(b) Assume that $\sigma_r^2 = \lim_{d\to\infty} \mathrm{Cov}\big(\mathrm{I}\{S_{\pi_1} \neq S_{\pi_2}\}, \mathrm{I}\{\tilde{S}_{\tilde{\pi}_1} \neq \tilde{S}_{\tilde{\pi}_2}\}\big)$ *exists. Then as* n *and* d *grow to infinity, we have* $n^{-1/2}\Big(\min\{T_R, \tilde{T}_R\} - \frac{n+1}{2}\Big) \xrightarrow{D} \min\{Z_1', Z_2'\}$, *where* (Z_1', Z_2') *also follows a bivariate normal distribution with same marginals as* (Z_1, Z_2)*, but its covariance is* $\big(\sigma_r^2 - \frac{1}{4}\big)$.

While the cut-off for the tests based on T_S or $\widetilde{T}_S$ (note that they have the same cut-off) can be computed easily, finding the cut-off for the modified sign test is difficult to obtain from this

asymptotic null distribution unless one finds a consistent estimator for the covariance. So, here also, we use Bonferroni's method for implementing the modified sign test. The same strategy is used for the modified runs test as well. From the above discussion, it is quite clear that if the test based on T_S (respectively, T_R) or that based on $\widetilde{T}_S$ (respectively, $\widetilde{T}_R$), at least one of them is consistent, the modified sign test (respectively, the modified runs test) turns out to be consistent in the HDHSS set up. This result is stated in the following theorem.

Theorem 4.11. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent copies of a d-dimensional random vector $\mathbf{X} \sim \mathrm{P}$. If P is not spherically symmetric, we have the following results as n and d both diverge to infinity.*

(a) If $\max\big\{\liminf_{n,d\to\infty} E[\frac{T_S}{n}], \liminf_{n,d\to\infty} \mathbb{E}[\frac{\tilde{T}_S}{n}]\big\} > \frac{1}{2}$, *the power of the modified sign test converges to* 1.

(b) *If* $\min\big\{\limsup_{n,d\to\infty} \mathbb{E}[\frac{T_R}{n}], \limsup_{n,d\to\infty} \mathbb{E}[\frac{\tilde{T}_R}{n}]\big\} < \frac{1}{2}$, *the power of the modified runs test converges to* 1.

4.5 Analysis of simulated and real data sets

In this section, we analyze some high-dimensional simulated and real data sets to compare the empirical performance of our proposed tests with the OT test (Huang & Sen, 2023) and the DT test (Diks & Tong, 1999). Since the PP test (Fang, Zhu & Bentler, 1993) becomes computationally prohibitive in higher dimensions, we do not consider it in this chapter. Henceforth, by sign and runs tests, we shall refer to the tests proposed in Section 4.1, and the modified versions considered in Section 4.4 will be referred to as modified sign and modified runs tests, respectively. Throughout this section, all tests are considered to have the nominal level $\alpha = 0.05$. As in Chapter 3, for the DT test, we used the bandwidth $(0.25)^2\hat{\sigma}_0^2$ in higher dimensions. Each experiment was repeated 1000 times to compute the power of the tests by the proportion of times they rejected H_0.

First, we consider some examples (see Examples 4.6-4.11) involving high-dimension, low-sample size data. In each of these examples, we generated samples of size 50 and carried out our experiment for 10 different choices d ($d = 2^i$ for $i = 1, 2, \ldots, 10$) as before. In Examples 4.6 and 4.7, we deal with elliptic distributions with equi-correlated structures. Example 4.6 is the same as Example 4.1, and in Example 4.7, we replace the normal distribution with the Cauchy distribution. Descriptions of these two examples are given below.

Example 4.6. *We consider a d-variate normal distribution with the location $\mathbf{0}_d$ and the scatter matrix $\mathbf{\Sigma} = 0.4\mathbf{I}_d + 0.6\mathbf{J}_d$.*

Example 4.7. *We deal with a d-variate Cauchy distribution with the same location and scatter matrix as in Example 4.6.*

Figure 4.13 shows that in these two examples, the OT test had much lower power than all other tests considered here. The rest of the tests had comparable performance in Example 4.6. They also had satisfactory performance in Example 4.7. However, in this example, our proposed tests performed better than the DT test in higher dimensions.

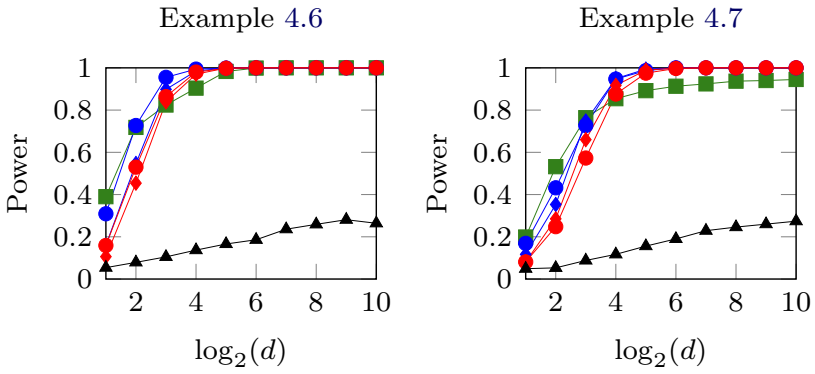


Fig. 4.13 *Powers of sign test (●), runs test (◆), modified sign test (●), modified runs test (◆), OT test (▲) and DT test (■) in Examples 4.6 and 4.7.*

Next, we consider two examples, where all off-diagonal elements of the dispersion matrix are 0, but the diagonal elements are not equal. Example 4.8 is similar to Example 4.3, but here we have a weaker signal against spherical symmetry. Example 4.9 is the same as Example 4.2, where the sign and runs tests had powers close to the nominal level, but their modified versions had much better performance. Brief descriptions of these two examples are given below.

Example 4.8. *We consider the normal distribution* $\mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma})$*, where* $\mathbf{\Sigma} = \mathrm{diag}(d^{0.3}, 1, 1, \ldots, 1)$.

Example 4.9. *Here also, we consider a normal distribution* $\mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma})$*, where* $\mathbf{\Sigma}$ *is a diagonal matrix. It has the first* $[d/2]$ *diagonal elements equal to* 1 *and the rest equal to* 2.

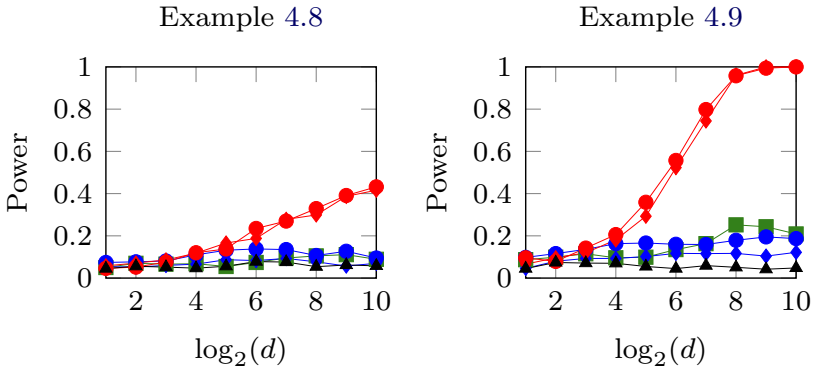


Fig. 4.14 *Powers of sign test (●) runs test (◆), modified sign test (●), modified runs test (◆), OT test (▲) and DT test (■) in Examples 4.8 and 4.9.*

In these examples, DT and OT tests had poor performance (see Figure 4.14). The sign and runs tests also performed poorly, but their modified versions worked well, especially in Example 4.9. This superiority of the modified tests was expected in view of our discussion in Section 4.4.

Now consider two examples, where the variance-covariance matrix of the underlying distribution is a constant multiple of the identity matrix, but the distribution is not spherical. Recall that we considered two such examples (see Examples 4.4 and 4.5) in Section 4.4. Here we revisit them as Examples 4.10 and 4.11, respectively.

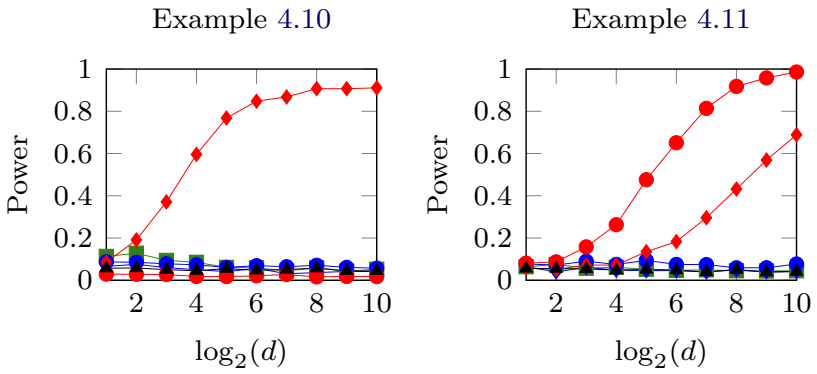


Fig. 4.15 *Powers of sign test (●), runs test (◆), modified sign test (●), modified runs test (◆), OT test (▲) and DT test (■) in Examples 4.10 and 4.11.*

Example 4.10. *We consider a d-dimensional distribution with i.i.d.* Unif$(-1,1)$ *coordinates.*

Example 4.11. *We consider a d-dimensional distribution with i.i.d Laplace(0,1) coordinates.*

In Example 4.10, while the modified runs test had an excellent performance, all other tests had powers close to the nominal level (see Figure 4.15). In Example 4.11, the modified sign test had the best performance followed by the modified runs test. All other tests performed poorly. The reasons for such performance of the modified tests has already been discussed in Section 4.4.

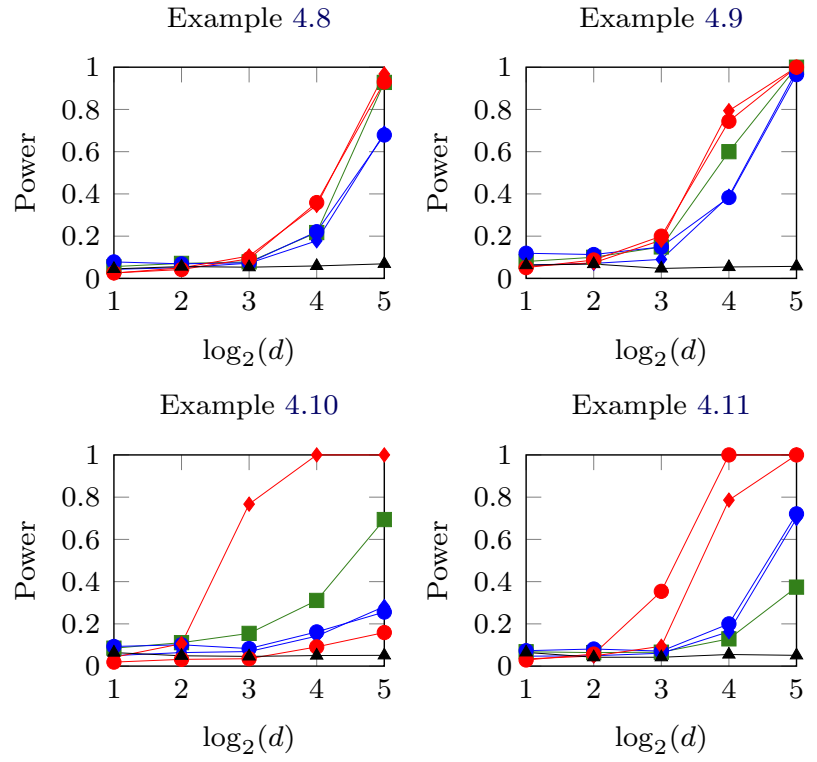


Fig. 4.16 *Power of sign test (●), runs test (◆), modified sign test (●), modified runs test (◆), OT test (▲) and DT test (■) Examples 4.8 -4.11 when the sample size $n = d^2 + 20$ increases with the dimension d.*

In Examples 4.8-4.11, though the sign and runs tests had powers close to the nominal level in high dimensions, they can have better performance if the sample size also increases with the dimension at a suitable rate. To demonstrate this, we revisit these four examples, but this time we consider the sample size $n = d^2 + 20$ that increases with the dimension d. The results are reported in Figure 4.16. Unlike before, except for the OT test, powers of all tests increased with the dimension. In Examples 4.8 and 4.9, though the modified tests had a clear edge, all these tests had satisfactory performance in high dimensions. In Example 4.10, the modified runs test outperformed all other tests as before. The DT test had the second-best performance, while the sign and runs tests performed better than the modified sign test. However, in Example 4.11, the modified sign test had the best performance followed by the modified runs test. In this example, the sign and runs tests had higher powers than the DT test in high dimensions.

4.5.1 Analysis of 'Earthquake' data

For further evaluation of the performance of our tests, we analyze the 'Earthquakes' data available at the Time Series Machine Learning website. Data were collected from the Northern California Earthquake Data Center and donated by Prof. Anthony Bagnall. Here each datum is the hourly average of readings on the Richter scale during 1967 and 2003. The single time series was then transformed into multi-dimensional objects by segmenting the time series by intervals of 512 hours. Any reading over 5 on the Richter scale is defined as a major event. However, such events are often followed by aftershocks. Hence, a segment of the time series is considered to be a positive case if there is a major event in that segment that is not preceded by another major event for at least 512 hours. Any reading below 4 that is preceded by at least 20 non-zero readings in the previous 512 hours is considered a negative case. After this initial processing, this dataset has 512 hourly readings on 368 negative cases and 93 positive cases. We consider these two groups containing (a) Positive cases and (b) Negative cases separately and test whether their underlying distributions are spherically symmetric.

However, if we use the full data set for testing, any test will either accept or reject the null hypothesis. Based on that single experiment, it is difficult to compare among different test procedures. Therefore, to compare our tests with the other methods, we adopted a sub-sampling approach, where we took a random sub-sample containing p $(0 < p < 1)$ proportion of observations and implemented the tests on those sub-samples. For each of the 5 values of p (0.2, 0.4, 0.6,0.8, and 0.95), this experiment was carried out 1000 times to compute the power of the tests by the proportion of times they rejected H_0. The results are reported in Figure 4.17.

We observed an interesting phenomenon in this dataset. For the group containing Positive cases, only our modified tests and the DT test were able to detect the deviation from spherical asymmetry. In this example, the modified sign test outperformed all other tests.For the group of Negative cases, the powers of all tests steadily increased with p. Here also, the modified sign test

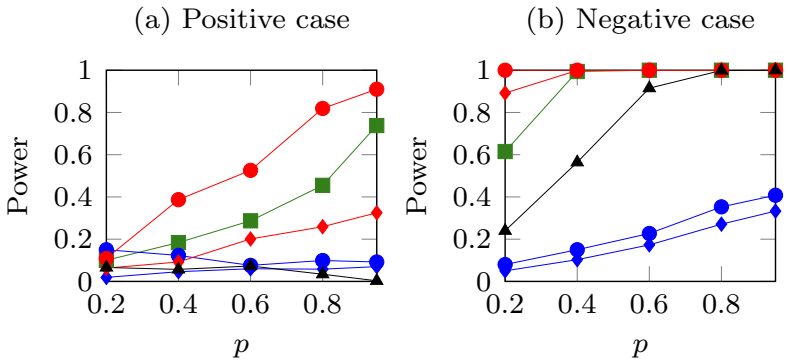


Fig. 4.17 *Powers of the sign test (●), the runs test (◆), the modified sign test (●), the modified runs test (◆), the OT test (▲) and the DT test (■) based on varying proportions of observations (p) from the positive and the negative cases in the 'Earthquakes' dataset.*

significantly outperformed all its other competitors. The modified runs test had the second-best performance, closely followed by the DT test. The OT test exhibited satisfactory performance. However, our sign test and runs test (based on T_S and T_R) had relatively low powers.

The above result indicates that the distribution of the Negative cases deviates more from spherical symmetry compared to the distribution of the Positive cases. This is confirmed by Figure 4.18, which shows the plots of the coordinate-wise mean and variance for the two groups. For the Positive cases, the mean is more or less stationary about zero, but for the Negative cases, the mean,

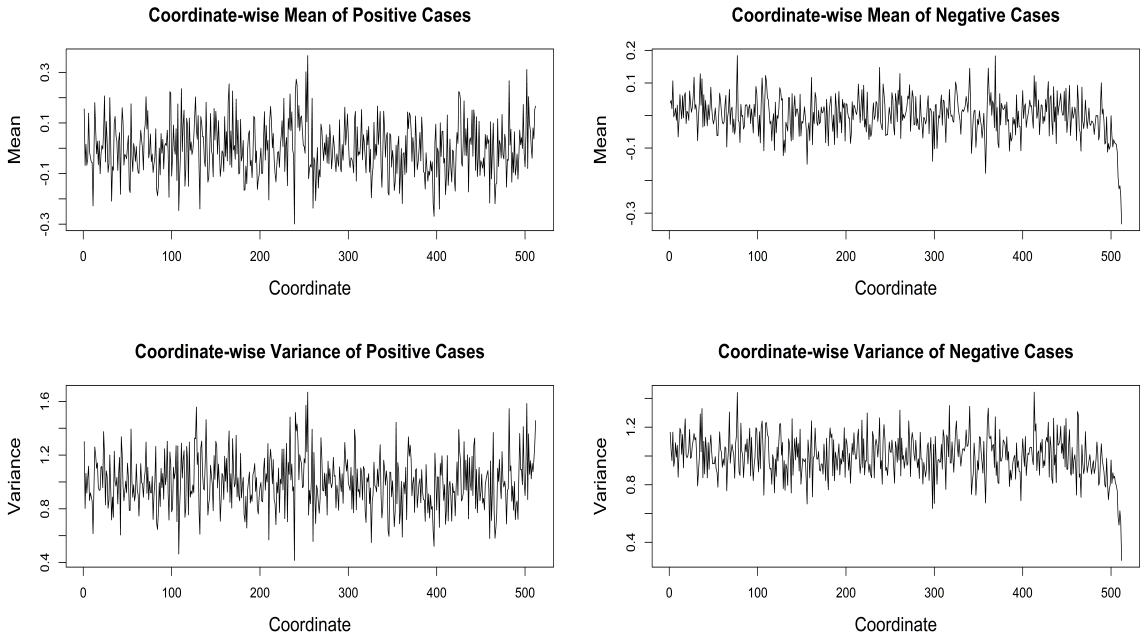


Fig. 4.18 *Coordinate-wise mean and variance of the feature vector in the 'Earthquakes' data set divided into two groups of positive and negative cases.*

as well as the variance, have a sharp drop at the right end. This sharp drop can be a potential reason behind the high powers of the tests in Figure 4.17 (b). However, this sharp drop in the mean and variance of the data could be a subjective bias at the data curation step. Therefore, to eliminate such possible bias, we truncated the feature vectors (both for Positive and Negative cases) by removing 32 features from the end and carried out our experiment with the first 480 coordinates (which corresponds to 20 days hourly readings on the Richter scale). Powers of different tests were computed based on 1000 random sub-samples as before. Our findings are reported in Figure 4.19.

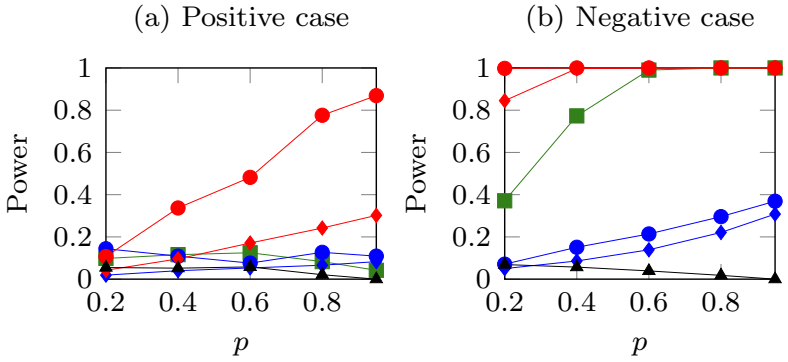


Fig. 4.19 *Powers of sign test (●), runs test (◆), modified sign test (●), modified runs test (◆), OT test (▲) and DT test (■) based on varying proportions of observations (p) from the positive and the negative cases in the truncated 'Earthquakes' dataset.*

Here also, the modified sign tests significantly outperformed their competitors both for Positive and Negative cases. For the Positive cases, unlike before, the power of the DT test did not show any increasing pattern. Here, only the modified sign and runs tests had powers increasing with p. For the Negative cases, the OT test had very poor performance. Our sign and runs tests based on T_S and T_R also failed to achieve satisfactory performance. However, the other tests showed a similar pattern as observed in Figure 4.17 (b).

4.6 Tests of spherical symmetry about an unknown center

So far, we have considered the null hypothesis that specifies the center of symmetry of the underlying distribution. Without loss of generality, the origin was taken as the specified center. However, if the null hypothesis does not specify the center, it calls for a test of spherical symmetry about an unknown center $\boldsymbol{\mu}$. In such cases, one can think of finding $\hat{\boldsymbol{\mu}}$. an estimate of $\boldsymbol{\mu}$, from the data and test for the spherical symmetry about $\hat{\boldsymbol{\mu}}$. For instance, we can use the sample mean as $\hat{\boldsymbol{\mu}}$, subtract it from the original observations for centering, and then apply our tests on the centered data. Of course, one can also use other robust estimates of $\boldsymbol{\mu}$ like the MCD estimate (see, e.g., Rousseeuw & Driessen, 1999) or the MVE estimate (see, e.g., Van Aelst & Rousseeuw, 2009). Depth-based estimates like the spatial median (see, e.g. Chaudhuri, 1996; Koltchinskii & Li, 1998) can be used as well. To evaluate the effect of centering on different tests, we consider a simple example.

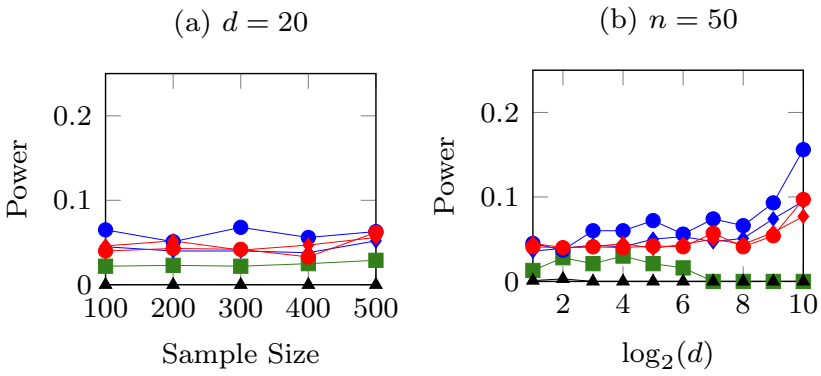


Fig. 4.20 *Type I errors of sign test (●), runs test (◆), modified sign test (●), modified runs test (◆), OT test (▲) and DT test (■) in Example 4.12 when (a) the sample size increases while the dimension is kept fixed at* 20 *and (b) the dimension increases while the sample size is kept fixed at* 50*.*

Example 4.12. *Observations are generated from* $\mathcal{N}_d(\mathbf{1}_d, \mathbf{I}_d)$*.*

We used the spatial median as $\hat{\boldsymbol{\mu}}$ and applied different tests on the centered data. First, we looked at the type I errors of different tests as functions of the sample size when the dimension was kept fixed at 20 (see Figure 4.20 (a)). In this case, type I errors of all tests became close to the nominal level $\alpha = 0.05$ as the sample size increased. With the increasing sample size, since $\hat{\boldsymbol{\mu}}$ becomes close to $\boldsymbol{\mu}$, such a phenomenon is quite expected. Next, we looked at their type I errors when the sample size was kept fixed at 50, and the dimension varied (see Figure 4.20 (b)). In higher dimensions, the type I error rates of our tests became larger than the nominal level. On the other hand, the type I error rates of DT and OT tests converged to 0 as the dimension increased. Note that if the dimension is higher compared to the sample size, $\hat{\boldsymbol{\mu}}$ may significantly deviate from the actual center $\boldsymbol{\mu}$, and this leads to the loss of the exchangeability property of the observed data points and their spherically symmetric variants (i.e., $\mathbf{X} - \hat{\boldsymbol{\mu}}$ and $\|\mathbf{X} - \hat{\boldsymbol{\mu}}\|\mathbf{U}$) , which increased the type I error of our tests. To take care of this problem, we can use an idea based on sample splitting, which is motivated by the result stated below.

Lemma 4.2. *Suppose that* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are two independent copies of* $\mathbf{X} \sim \mathrm{P}$*, which is symmetric about* $\boldsymbol{\mu}$*. Then* P *is spherically symmetric about* $\boldsymbol{\mu}$ *if and only if the distribution of* $\mathbf{X}_1 - \mathbf{X}_2$ *is spherically symmetric about the origin.*

Therefore, if the location $\boldsymbol{\mu}$ is unknown and the sample size n is even (discard one observation, if needed), we can use our tests on the transformed data $\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_{n/2}\}$, where $\boldsymbol{Z}_i = \mathbf{X}_i - \mathbf{X}_{n/2+i}$ for $i = 1, 2, \ldots, n/2$. The resulting tests will have the exact distribution-free property, and their asymptotic properties can be established using arguments similar to those in Section 4.3 and 4.4. Therefore, to avoid repetition, we omit those discussions here. To demonstrate the empirical performance of the resulting tests, we consider the following example.

Example 4.13. *We deal with* 50 *observations from* $\mathcal{N}_d(\mathbf{1}_d, 0.7\mathbf{I}_d + 0.3\mathbf{J}_d)$*.*

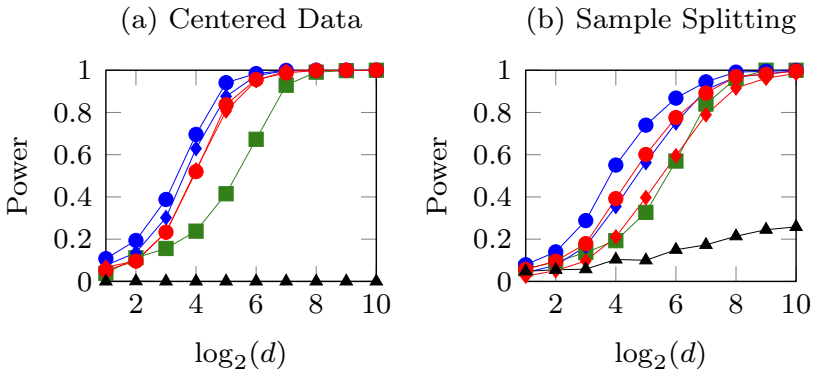


Fig. 4.21 *Powers of sign test (●), runs test (◆), modified sign test (●), modified runs test (◆), OT test (▲) and DT test (■) in Example 4.13 when (a) the samples are centered using the spatial median and (b) we use differences of the observations based on sample splitting.*

We computed the powers of the tests (a) when the observations were centered using the spatial median and (b) when the above idea-based sample splitting was used. The results are given in Figure 4.21. For the centered data, our tests had higher powers than OT and DT tests. We have seen that when we use centering based on the spatial median, the DT test becomes conservative in high dimension, while our tests have a tendency to have inflated type I error (see Figure 4.20). This may be one of the reasons for the significant difference in their powers. The OT test had a very poor performance, it had almost zero power in all dimensions. This may also be due to the conservativeness of this test, as observed in Figure 4.20. When we adopted the sample splitting idea and used the tests on the differences of the observations, the power of our tests became slightly lower than what we observed before, but still, they had an edge over the DT test. Surprisingly, this method helped the OT test to gain some power. Though its performance was inferior to other competitors, unlike before, we observed an increasing trend in its power as the dimension increased.

4.7 Proofs and mathematical details

Proof of Lemma 4.1. It is easy to see that if P is spherically symmetric $(\mathbf{X}_1, \mathbf{X}_1') \overset{D}{=} (\mathbf{X}_1', \mathbf{X}_1)$. Therefore, for any measurable function $h(\cdot,\cdot)$, $h(\mathbf{X}_1, \mathbf{X}_2) \overset{D}{=} h(\mathbf{X}_1, \mathbf{X}_2') \overset{D}{=} h(\mathbf{X}_1', \mathbf{X}_2')$ holds trivially. To prove the only if part, first note that

$$\mathbb{P}[h(\mathbf{X}_1, \mathbf{X}_2) \le t] = \int_{\{h(\mathbf{x},\mathbf{y})\le t\}} p(\mathbf{x})p(\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = \int_{\{h(\mathbf{x}-\mathbf{y},0)\le t\}} p(\mathbf{x})p(\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}$$
$$= \int_{\{h(\mathbf{v},0)\le t\}} \left[\int p(\mathbf{y}+\mathbf{v})p(\mathbf{y})\mathrm{d}\mathbf{y}\right] \mathrm{d}\mathbf{v}.$$

By the square integrability of p, we can say that

$$u(\mathbf{v}) := \int p(\mathbf{y}+\mathbf{v})p(\mathbf{y})\mathrm{d}\mathbf{y} \qquad \left[\le \left(\int p^2(\mathbf{y}+\mathbf{v})\mathrm{d}\mathbf{y}\right)^{1/2}\left(\int p^2(\mathbf{y})\mathrm{d}\mathbf{y}\right)^{1/2} = \int p^2(\mathbf{y})\mathrm{d}\mathbf{y} < \infty\right]$$

is locally integrable on $\mathbb{R}^d$ (i.e., integrable on all compact subsets of $\mathbb{R}^d$). Hence, by Theorem 7.15 in Wheeden & Zygmund (1977), we conclude that almost every point in $\mathbb{R}^d$ is a Lebesgue point of

$u(\cdot)$. Also, by our assumption (c) and Theorem 7.16 in Wheeden & Zygmund (1977), we obtain

$$\lim_{t\downarrow 0} \frac{\int_{\{h(\mathbf{v},0)\le t\}} |u(\mathbf{v}) - u(0)|\mathrm{d}\mathbf{v}}{\int_{\{h(\mathbf{v},0)\le t\}} \mathrm{d}\mathbf{v}} = 0,$$

or in other words

$$\lim_{t\downarrow 0} \frac{\mathbb{P}[h(\mathbf{X}_1,\mathbf{X}_2)\le t]}{\int_{\{h(\mathbf{v},0)\le t\}} \mathrm{d}\mathbf{v}} = \lim_{t\downarrow 0} \frac{\int_{\{h(\mathbf{v},0)\le t\}} u(\mathbf{v})\mathrm{d}\mathbf{v}}{\int_{\{h(\mathbf{v},0)\le t\}} \mathrm{d}\mathbf{v}} = u(0) = \int p^2(\mathbf{y})\mathrm{d}\mathbf{y}.$$

Now, note that the random variable $\mathbf{X}_1'$ has the density $p'(\mathbf{x}) = \int p(\mathbf{H}^\top\mathbf{x})\mathrm{d}\nu_0(\mathbf{H})$ where ν_0 is the Haar measure on the set of all $d\times d$ orthogonal matrices (see Lemma A.2 in Banerjee & Ghosh, 2024a). One can show that if p is square integrable, so is p'. Therefore, using the same argument, we can show that

$$\lim_{t\downarrow 0} \frac{\mathbb{P}[h(\mathbf{X}_1,\mathbf{X}_2')\le t]}{\int_{\{h(\mathbf{v},0)\le t\}} \mathrm{d}\mathbf{v}} = \int p(\mathbf{y})p'(\mathbf{y})\mathrm{d}\mathbf{y} \quad \text{and} \quad \lim_{t\downarrow 0} \frac{\mathbb{P}[h(\mathbf{X}_1',\mathbf{X}_2')\le t]}{\int_{\{h(\mathbf{v},0)\le t\}} \mathrm{d}\mathbf{v}} = \int p'^2(\mathbf{y})\mathrm{d}\mathbf{y}.$$

However, under the assumption $h(\mathbf{X}_1,\mathbf{X}_2) \overset{D}{=} h(\mathbf{X}_1,\mathbf{X}_2') \overset{D}{=} h(\mathbf{X}_1',\mathbf{X}_2')$, we must have

$$\mathbb{P}[h(\mathbf{X}_1,\mathbf{X}_2)\le t] = \mathbb{P}[h(\mathbf{X}_1,\mathbf{X}_2')\le t] = \mathbb{P}[h(\mathbf{X}_1',\mathbf{X}_2')\le t] \quad \forall t\in\mathbb{R}.$$

So, combining our results, we get

$$\int p^2(\mathbf{y})\mathrm{d}\mathbf{y} = \int p'^2(\mathbf{y})\mathrm{d}\mathbf{y} = \int p(\mathbf{y})p'(\mathbf{y})\mathrm{d}\mathbf{y}. \tag{4.7}$$

By Cauchy-Schwartz inequality, the equality in (4.7) holds if and only if $p = p'$ almost surely, i.e., P is spherically symmetric. ■

Proof of Theorem 4.1. Let us define $\mathcal{T}(\boldsymbol{s},\boldsymbol{\pi}) := \sum_{i=1}^{n-1} \theta(\boldsymbol{Y}_{s_{\pi_i},\pi_i}, \boldsymbol{Y}_{s_{\pi_{i+1}},\pi_{i+1}})$, for $\boldsymbol{\pi}\in\mathcal{S}_n$ and $\boldsymbol{s} = (s_1,s_2,\ldots,s_n)\in\{0,1\}^n$. Let $(\boldsymbol{S},\boldsymbol{\Pi})$ be as defined in (4.1). Then for any $\boldsymbol{\pi}_0\in\mathcal{S}_n$ and $\boldsymbol{s}_0\in\{0,1\}^n$, we have

$$\mathbb{P}[\boldsymbol{S}=\boldsymbol{s}_0;\boldsymbol{\Pi}=\boldsymbol{\pi}_0] = \mathbb{P}\left[\mathcal{T}(\boldsymbol{s}_0,\boldsymbol{\pi}_0)\le\mathcal{T}(\boldsymbol{s},\boldsymbol{\pi}) \;\; \forall \boldsymbol{s}\in\{0,1\}^n \text{ and } \boldsymbol{\pi}\in\mathcal{S}_n\right]. \tag{4.8}$$

When P is spherically symmetric, we have $(\mathbf{X}_i,\mathbf{X}_i') \overset{D}{=} (\mathbf{X}_i',\mathbf{X}_i)$ for $i=1,2,\ldots,n$. So, the random variables $\{\mathcal{T}(\boldsymbol{s},\boldsymbol{\pi})\}_{\boldsymbol{s}\in\{0,1\}^n;\boldsymbol{\pi}\in\mathcal{S}_n}$ are exchangeable. Hence, the probability on the right side of (4.8) does not depend on $\boldsymbol{s}_0$ and $\boldsymbol{\pi}_0$. Therefore, using the identity $\sum_{\boldsymbol{s}_0\in\{0,1\}^n}\sum_{\boldsymbol{\pi}_0\in\mathcal{S}_n}\mathbb{P}[\boldsymbol{S}=\boldsymbol{s}_0;\boldsymbol{\Pi}=\boldsymbol{\pi}_0]=1$, we get

$$\mathbb{P}[\boldsymbol{S}=\boldsymbol{s}_0;\boldsymbol{\Pi}=\boldsymbol{\pi}_0] = 2^{-n}(n!)^{-1}.$$

This implies $\boldsymbol{S}\sim\text{Unif}(\{0,1\}^n)$, $\boldsymbol{\Pi}\sim\text{Unif}(\mathcal{S}_n)$ and they are independent. Since, $\boldsymbol{R}$ is the inverse permutation of $\boldsymbol{\Pi}$, we have $\boldsymbol{R}\sim$Unif$(\mathcal{S}_n)$ independent of $\boldsymbol{S}$. This proves part (a) of the theorem.

If P is not spherically symmetric, $(\mathbf{X}_i,\mathbf{X}_i')$ and $(\mathbf{X}_i',\mathbf{X}_i)$ are not distributionally equal, but $\{(\mathbf{X}_i,\mathbf{X}_i')\}_{1\le i\le n}$ forms a sequence of i.i.d. random vectors. So, for any $\boldsymbol{\pi}_0\in\mathcal{S}_n$, we have

$$\mathbb{P}[\boldsymbol{\Pi}=\boldsymbol{\pi}_0] = \sum_{\boldsymbol{s}_0\in\{0,1\}^n} \mathbb{P}[\boldsymbol{S}=\boldsymbol{s}_0;\boldsymbol{\Pi}=\boldsymbol{\pi}_0] = \sum_{\boldsymbol{s}_0\in\{0,1\}^n} \mathbb{P}[\boldsymbol{\Pi}=\boldsymbol{\pi}_0 \mid \boldsymbol{S}=\boldsymbol{s}_0]\,\mathbb{P}[\boldsymbol{S}=\boldsymbol{s}_0].$$

For any given $\boldsymbol{s}_0 \in \{0,1\}^n$, the random variables $\{\mathcal{T}(\boldsymbol{s}_0, \boldsymbol{\pi})\}_{\boldsymbol{\pi} \in \mathcal{S}_n}$ are exchangeable. Hence, the probability $\mathbb{P}[\boldsymbol{\Pi} = \boldsymbol{\pi}_0 | \boldsymbol{S} = \boldsymbol{s}_0]$ is constant for every $\boldsymbol{\pi}_0 \in \mathcal{S}_n$. Using $\sum_{\boldsymbol{\pi}_0 \in \mathcal{S}_n} \mathbb{P}[\boldsymbol{\Pi} = \boldsymbol{\pi}_0 | \boldsymbol{S} = \boldsymbol{s}_0] = 1$, for any $\boldsymbol{s}_0 \in \{0,1\}^n$, we get $\mathbb{P}[\boldsymbol{\Pi} = \boldsymbol{\pi}_0 | \boldsymbol{S} = \boldsymbol{s}_0] = (n!)^{-1}$ and therefore $\mathbb{P}[\boldsymbol{\Pi} = \boldsymbol{\pi}_0] = (n!)^{-1}$. Here also, by the same argument as above, we have $\boldsymbol{R} \sim \text{Unif}(\mathcal{S}_n)$. Now, given $\boldsymbol{\Pi} = (\pi_1, \ldots, \pi_n)$ and the augmented data $\{(\mathbf{X}_i, \mathbf{X}'_i)\}_{1 \le i \le n}$ we have the following relation

$$I\{S_{\pi_1} = 1\} = \mathrm{I}\big\{\theta(\mathbf{X}_{\pi_1}, \boldsymbol{Y}_{S_{\pi_2}, \pi_2}) \le \theta(\mathbf{X}'_{\pi_1}, \boldsymbol{Y}_{S_{\pi_2}, \pi_2})\big\}.$$

This follows from the fact that the total cost of the path starting with $\mathbf{X}_{\pi_1}$ and that starting with $\mathbf{X}'_{\pi_1}$ differ only in the first term. Using similar arguments, we also get

$$\begin{aligned}
\mathrm{I}\{S_{\pi_k} = 1\} = \mathrm{I}\big\{&\theta(\boldsymbol{Y}_{S_{\pi_{(k-1)}}, \pi_{(k-1)}}, \mathbf{X}_{\pi_k}) + \theta(\mathbf{X}_{\pi_k}, \boldsymbol{Y}_{S_{\pi_{(k+1)}}, \pi_{(k+1)}}) \\
&\le \theta(\boldsymbol{Y}_{S_{\pi_{(k-1)}}, \pi_{(k-1)}}, \mathbf{X}'_{\pi_k}) + \theta(\mathbf{X}'_{\pi_k}, \boldsymbol{Y}_{S_{\pi_{(k+1)}}, \pi_{(k+1)}})\big\} \text{ for } k = 2, \ldots, n-1 \text{ and} \\
\mathrm{I}\{S_{\pi_n} = 1\} = \mathrm{I}\big\{&\theta(\boldsymbol{Y}_{S_{\pi_{n-1}}, \pi_{n-1}}, \mathbf{X}_{\pi_n}) \le \theta(\boldsymbol{Y}_{S_{\pi_{n-1}}, \pi_{n-1}}, \mathbf{X}'_{\pi_n})\big\}.
\end{aligned}$$

Now, taking conditional expectation given $\boldsymbol{\Pi}$, we have

$$\begin{aligned}
\mathbb{P}^{**}[S_{\pi_1} = 1] &= \mathbb{P}^{**}\big[\theta(\mathbf{X}_{\pi_1}, \boldsymbol{Y}_{S_{\pi_2}, \pi_2}) \le \theta(\mathbf{X}'_{\pi_1}, \boldsymbol{Y}_{S_{\pi_2}, \pi_2})\big], \\
\mathbb{P}^{**}[S_{\pi_k} = 1] &= \mathbb{P}^{**}\big[\theta(\boldsymbol{Y}_{S_{\pi_{(k-1)}}, \pi_{(k-1)}}, \mathbf{X}_{\pi_k}) + \theta(\mathbf{X}_{\pi_k}, \boldsymbol{Y}_{S_{\pi_{(k+1)}}, \pi_{(k+1)}}) \\
&\qquad \le \theta(\boldsymbol{Y}_{S_{\pi_{(k-1)}}, \pi_{(k-1)}}, \mathbf{X}'_{\pi_k}) + \theta(\mathbf{X}'_{\pi_k}, \boldsymbol{Y}_{S_{\pi_{(k+1)}}, \pi_{(k+1)}})\big] \text{ for } k = 2, \ldots, n-1 \text{ and} \\
\mathbb{P}^{**}[S_{\pi_n} = 1] &= \mathbb{P}^{**}\big[\theta(\boldsymbol{Y}_{S_{\pi_{n-1}}, \pi_{n-1}}, \mathbf{X}_{\pi_n}) \le \theta(\boldsymbol{Y}_{S_{\pi_{n-1}}, \pi_{n-1}}, \mathbf{X}'_{\pi_n})\big].
\end{aligned}$$

where $\mathbb{P}^{**}$ denotes the conditional distribution given $\boldsymbol{\Pi}$. This establishes the weak dependence structure of the string signs and completes the proof of part (b) of the theorem. ■

Proof of Theorem 4.2. For the proof of this theorem, we refer the reader to Theorem 1 in Jung & Marron (2009). ■

Proof of Theorem 4.3. Let $\mathbf{X}'_1 = \|\mathbf{X}_1\|\mathbf{U}_1$ and $\mathbf{X}'_2 = \|\mathbf{X}_2\|\mathbf{U}_2$ be the spherically symmetric variants of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Here $\mathbf{U}_1, \mathbf{U}_2$ are i.i.d. Unif($\mathcal{S}^{d-1}$) independent of $\mathbf{X}_1$ and $\mathbf{X}_2$. Then we have

$$\mathbb{P}\left[\theta(\mathbf{X}_1, \mathbf{X}_2) < \theta(\mathbf{X}_1, \mathbf{X}'_2)\right] = \mathbb{P}\left[(\mathbf{X}_1^\top \mathbf{X}_2)^2 > (\mathbf{X}_1^\top \mathbf{X}'_2)^2\right] = \mathbb{P}\left[(\mathbf{X}_1^\top \mathbf{X}_2)^2 > \|\mathbf{X}_2\|^2 (\mathbf{X}_1^\top \mathbf{U}_2)^2\right].$$

By spherical symmetry of $\mathbf{U}_2$, for any $\boldsymbol{a} \in \mathbb{R}^d$, we have $(\boldsymbol{a}^\top \mathbf{U}_2) \overset{D}{=} \|\boldsymbol{a}\| U_{21}$, where U_{21} is the first coordinate of $\mathbf{U}_2$. Therefore, conditioned on $\mathbf{X}_1$ and $\mathbf{X}_2$, we have

$$\mathbb{P}\left[\theta(\mathbf{X}_1, \mathbf{X}_2) < \theta(\mathbf{X}_1, \mathbf{X}'_2) \middle| \mathbf{X}_1, \mathbf{X}_2\right] = \mathbb{P}\left[(\mathbf{X}_1^\top \mathbf{X}_2)^2 > \|\mathbf{X}_2\|^2 \|\mathbf{X}_1\|^2 U_{21}^2 \middle| \mathbf{X}_1 \mathbf{X}_2\right].$$

Now, taking expectations with respect to $\mathbf{X}_1$ and $\mathbf{X}_2$, we get

$$\mathbb{P}\left[\theta(\mathbf{X}_1, \mathbf{X}_2) < \theta(\mathbf{X}_1, \mathbf{X}'_2)\right] = \mathbb{P}\left[\frac{(\mathbf{X}_1^\top \mathbf{X}_2)^2}{\|\mathbf{X}_2\|^2 \|\mathbf{X}_1\|^2} > U_{21}^2\right].$$

From the elementary theory of sampling distributions, we know that U_{21}^2 follows a Beta($\frac{1}{2}, \frac{d-1}{2}$)

distribution (see, e.g. Liang, Fang & Hickernell, 2008). So, we have $\mathbb{E}[U_{21}^2] = \frac{1}{d}$ and $\mathrm{Var}[U_{21}^2] = \frac{2}{d(d+2)}$. Therefore, $\{dU_{21}^2\}_{d\geq 1}$ is a tight sequence of random variables. Hence, if

$$\mathbb{P}\left[\frac{d(\mathbf{X}_1^\top \mathbf{X}_2)^2}{\|\mathbf{X}_1\|^2\|\mathbf{X}_2\|^2} > M\right] \to 1 \qquad \text{for all } M > 0,$$

we have $\mathbb{P}\Big[\frac{d(\mathbf{X}_1^\top \mathbf{X}_2)^2}{\|\mathbf{X}_1\|^2\|\mathbf{X}_2\|^2} > dU_{21}^2\Big] = \mathbb{P}\left[\theta(\mathbf{X}_1, \mathbf{X}_2) < \theta(\mathbf{X}_1, \mathbf{X}_2')\right] \to 1$ as $d \to \infty$. Similarly, we can also show that under the given condition, $\mathbb{P}\left[\theta(\mathbf{X}_1, \mathbf{X}_2) < \theta(\mathbf{X}_1', \mathbf{X}_2)\right]$ and $\mathbb{P}\left[\theta(\mathbf{X}_1, \mathbf{X}_2) < \theta(\mathbf{X}_1', \mathbf{X}_2')\right]$ also converge to one as d diverges to infinity. Now, for any $\boldsymbol{\pi} \in \mathcal{S}_n$, define

$$E_i = \Big\{\theta(\mathbf{X}_{\pi_i}, \mathbf{X}_{\pi_{i+1}}) = \min\big\{\theta(\mathbf{X}_{\pi_i}, \mathbf{X}_{\pi_{i+1}}),\ \theta(\mathbf{X}_{\pi_i}, \mathbf{X}'_{\pi_{i+1}}),\ \theta(\mathbf{X}'_{\pi_i}, \mathbf{X}_{\pi_{i+1}}),\ \theta(\mathbf{X}'_{\pi_i}, \mathbf{X}'_{\pi_{i+1}})\big\}$$

for $i = 1, \ldots, n-1$. Clearly, $\mathbb{P}(E_i \mid \boldsymbol{\Pi} = \boldsymbol{\pi}) \to 1$ for all $i = 1, \ldots, n-1$ as $d \to \infty$. Now, let $\boldsymbol{S}$ be the solution of (4.1) and define an n-dimensional vector $\mathbf{1}_n = (1, \ldots, 1)$. Note that for any $\boldsymbol{\pi} \in \mathcal{S}_n$,

$$\{\boldsymbol{S} = \mathbf{1}_n\} \supseteq E_1 \cap E_2 \cap \ldots \cap E_{n-1} \text{ and hence } \mathbb{P}(\boldsymbol{S} = \mathbf{1}_n \mid \boldsymbol{\Pi} = \boldsymbol{\pi}) \to 1 \text{ as } d \to \infty.$$

Now, the result follows by a simple application of the Dominated Convergence Theorem. ■

Proof of Corollary 4.1. Here $\boldsymbol{Z}_i$ $(i = 1, 2)$ can be viewed as a standardized version of $\mathbf{X}_i$, and we have $\mathrm{Var}(\boldsymbol{Z}_i) = \mathbf{I}_d$, the $d \times d$ identity matrix. Now,

$$(\mathbf{X}_1^\top \mathbf{X}_2) = \boldsymbol{Z}_1^\top \boldsymbol{\Lambda}_d \boldsymbol{Z}_2 = \sum_{i=1}^d \lambda_i Z_{1,i} Z_{2,i} \text{ and } (\mathbf{X}_1^\top \mathbf{X}_1) = \boldsymbol{Z}_1^\top \boldsymbol{\Lambda} \boldsymbol{Z}_1 = \sum_{i=1}^d \lambda_i Z_{1,i}^2.$$

Note that

$$\mathrm{Var}\left[\frac{\sum_{i=2}^d \lambda_i Z_{1,i} Z_{2,i}}{\sum_{i=2}^d \lambda_i}\right] = \frac{1}{\Big(\sum_{i=2}^d \lambda_i\Big)^2}\left[\sum_{i=2}^d \lambda_i^2 \mathrm{Var}\Big[Z_{1,i}Z_{2,i}\Big] + \sum_{2\leq i\neq j\leq d} \lambda_i\lambda_j \mathrm{Cov}\Big(Z_{1,i}Z_{2,i},\ Z_{1,j}Z_{2,j}\Big)\right].$$

Since $\mathrm{Cov}\Big(Z_{1,i}Z_{2,i},\ Z_{1,j}Z_{2,j}\Big) = \mathbb{E}\left[Z_{1,i}Z_{2,i}Z_{1,j}Z_{2,j}\right] = \big[\mathbb{E}\big[Z_{1,i}Z_{1,j}\big]\big]^2 = 0\ \forall i \neq j\ (i, j \geq 2)$, we have

$$\mathrm{Var}\left[\frac{\sum_{i=2}^d \lambda_i Z_{1,i} Z_{2,i}}{\sum_{i=2}^d \lambda_i}\right] = \frac{1}{\Big(\sum_{i=2}^d \lambda_i\Big)^2}\sum_{i=2}^d \lambda_i^2 \mathrm{Var}\Big[Z_{1,i}Z_{2,i}\Big] = \frac{\sum_{i=2}^d \lambda_i^2}{\Big(\sum_{i=2}^d \lambda_i\Big)^2},$$

So, using assumption (b) of the corollary, as d diverges to infinity, we get

$$\sum_{i=2}^d \lambda_i Z_{1,i} Z_{2,i} = o_P\Big(\sum_{i=2}^d \lambda_i\Big) \text{ and hence } (\mathbf{X}_1^\top \mathbf{X}_2) = \lambda_1 Z_{1,1} Z_{2,1} + o_P\Big(\sum_{i=2}^d \lambda_i\Big).$$

Also, note that under the ρ-mixing condition in (A2) and assumption (b) of the corollary,

$$\mathrm{Var}\left[\frac{\sum_{i=2}^d \lambda_i Z_{1,i}^2}{\sum_{i=2}^d \lambda_i}\right] = \frac{1}{\Big(\sum_{i=2}^d \lambda_i\Big)^2}\left[\sum_{i=2}^d \lambda_i^2 \mathrm{Var}\Big[Z_{1,i}^2\Big] + \sum_{2\leq i\neq j\leq d} \lambda_i\lambda_j \mathrm{Cov}\Big(Z_{1,i}^2, Z_{1,j}^2\Big)\right] \to 0$$

as $d \to \infty$ (follows from Theorem 1 in Jung & Marron (2009)). So, under these assumptions,

$$(\mathbf{X}_1^\top \mathbf{X}_1) = \lambda_{1,d} Z_{1,1}^2 + \sum_{i=2}^d \lambda_i + o_P\Big(\sum_{i=2}^d \lambda_i\Big). \ \Big(\text{note that } \mathbb{E}\Big[\sum_{i=2}^d Z_{1,i}^2\Big] = \sum_{i=2}^d \lambda_i\Big)$$

Now if $\kappa > 1$, using assumptions (a) and (b), as d diverges to infinity, we get

$$\frac{(\mathbf{X}_1^\top \mathbf{X}_2)^2}{\|\mathbf{X}_1\|^2\|\mathbf{X}_2\|^2} = \frac{(\mathbf{X}_1^\top \mathbf{X}_2)^2/d^{2\kappa}}{\|\mathbf{X}_1\|^2/d^\kappa \|\mathbf{X}_2\|^2/d^\kappa} \to \frac{c_1^2(Z_{1,1}Z_{2,1})^2}{c_1 Z_{1,1}^2 c_1 Z_{2,1}^2} = 1.$$

For $\kappa = 1$, as d diverges to infinity, for any $\epsilon > 0$, we have

$$\mathbb{P}\Big[\Big|\frac{\mathbf{X}_1^\top \mathbf{X}_2}{d} - c_1 Z_{1,1}Z_{2,1}\Big| > \epsilon\Big] \to 0 \text{ and } \mathbb{P}\Big[\Big|\frac{\mathbf{X}_1^\top \mathbf{X}_1}{d} - c_1 Z_{1,1}^2 - \frac{\sum_{i=2}^d \lambda_i}{d}\Big| > \epsilon\Big] \to 0.$$

Since $\sum_{i=2}^d \lambda_i = \mathcal{O}(d)$, $c_2 = \limsup_{d\to\infty} \sum_{i=2}^d \lambda_i/d$ is non-negative and finite, and for large d, we have

$$P\left(\frac{(\mathbf{X}_1^\top \mathbf{X}_2)^2}{\|\mathbf{X}_1\|^2\|\mathbf{X}_2\|^2} > t\right) \geq P\left(\frac{c_1^2 Z_{1,1}^2 Z_{2,1}^2}{(c_1 Z_{1,1}^2 + c_2)(c_1 Z_{2,1}^2 + c_2)} > t\right) \text{ for all } t \geq 0.$$

Note that on the right side we have a random variable which takes positive values with probability one. This implies both for $\kappa = 1$ and $\kappa > 1$, as d diverges to infinity, we have

$$P\Big[\frac{d(\mathbf{X}_1^\top \mathbf{X}_2)^2}{\|\mathbf{X}_1\|^2\|\mathbf{X}_2\|^2} > M\Big] \to 1 \text{ for any } M > 0.$$

Hence, under the given assumptions, the result follows from Theorem 4.3. ■

Proof of Theorem 4.4. Let $\boldsymbol{S}$ and $\boldsymbol{R}$ be the sign and rank vectors, and $\boldsymbol{\Pi}$ be the vector of anti-ranks (inverse permutation of $\boldsymbol{R}$) as defined in Section 4.1. Then by Theorem 4.1, $\boldsymbol{S} \sim \text{Unif}(\{0,1\}^n)$ and $\boldsymbol{\Pi} \sim \text{Unif}(\mathcal{S}_n)$ irrespective of the dimension d. Now, to find the distribution of the linear rank statistic $T_{LR} = \sum_{i=1}^n S_{\pi_i} a(i)$, first note that $\{a(i)(S_{\pi_i} - \frac{1}{2})\}\}_{1\leq i\leq n; n\geq 1}$ forms a triangular array of independent random variables whose distribution does not depend on the dimension of the data. Also, we have

$$\Delta_n^2 := \sum_{i=1}^n \text{Var}\big[a(i)(S_{\pi_i} - 1/2)\big] = \sum_{i=1}^n a^2(i)\text{Var}(S_{\pi_i}) = \frac{1}{4}\Big[\sum_{i=1}^n a^2(i)\Big] \text{ (since } \text{Var}[S_{\pi_i}] = 1/4 \ \forall i = 1(1)n).$$

Since the sequence of scores $\{a(i)\}$ that satisfies (4.3), as n and d both diverge to infinity, we have

$$\frac{1}{\Delta_n^3}\sum_{i=1}^n |a(i)|^3 \mathbb{E}\left[\left|S_{\pi_i} - 1/2\right|^3\right] \leq \frac{\sum_{i=1}^n |a(i)|^3}{\big(\sum_{i=1}^n a^2(i)\big)^{3/2}} \leq \max_{1\leq i\leq n} \frac{|a(i)|}{\big(\sum_{i=1}^n a^2(i)\big)^{1/2}} \to 0, \tag{4.9}$$

So, the triangular array $\{a(i)(S_{\pi_i} - 1/2)\}\}_{1\leq i\leq n; n\geq 1}$ satisfies the Lyapunov's condition. Hence, an application of the Lyapunov's Central Limit Theorem (CLT) proves the result. ■

Proof of Theorem 4.5. Let $\boldsymbol{S}$, $\boldsymbol{R}$ and $\boldsymbol{\Pi}$ be as defined in Section 4.1. Then the variance of the linear rank statistic $T_{LR} = \sum_{i=1}^n S_{\pi_i} a(i)$ is given by

$$\text{Var}\,[T_{LR}] = \sum_{i=1}^n a^2(i)\text{Var}(S_{\pi_i}) + \sum_{1\leq i\neq j\leq n} a(i)a(j)\text{Cov}(S_{\pi_i}, S_{\pi_j}).$$

From Theorem 4.1, we have $\text{Cov}(S_{\pi_i}, S_{\pi_j}) = 0 \ \forall \ |i-j| > 1$ irrespective of the dimension d. So,

$$\text{Var}[T_{LR}] = \sum_{i=1}^n a^2(i)\text{Var}(S_{\pi_i}) + \sum_{|i-j|=1} a(i)a(j)\text{Cov}(S_{\pi_i}, S_{\pi_j}).$$

Now, the uniformity of $\mathbf{\Pi}$ (see Theorem 4.1) suggests that $\text{Var}(S_{\pi_i})$ is constant for all $i = 1, 2, \ldots, n$ and $\text{Cov}(S_{\pi_i}, S_{\pi_j})$ is constant for all i, j with $|i - j| = 1$. Let us call them $\sigma_{11}^{(d)}$ and $\sigma_{12}^{(d)}$, respectively. Note that $\sigma_{11}^{(d)}$ and $\sigma_{12}^{(d)}$ are bounded and $\sum_{|i-j|=1} a(i)a(j) = 2\sum_{i=1}^{n-1} a(i)a(i+1)$. So, as n and d diverge to infinity, for a sequence of scores $\{a(i)\}_{1\le i\le n}$ satisfying (4.4), we have

$$\limsup_{n,d\to\infty} \text{Var}\left[\frac{T - \mathbb{E}[T]}{\sqrt{\sum_{i=1}^n a^2(i)}}\right] = \limsup_{n,d\to\infty}\left[\frac{\sum_{i=1}^n a^2(i)\sigma_{11}^{(d)}}{\sum_{i=1}^n a^2(i)} + 2\frac{\sum_{i=1}^{n-1} a(i)a(i+1)}{\sum_{i=1}^n a^2(i)}\sigma_{12}^{(d)}\right] \to \sigma_{11} + 2C\sigma_{12},$$

where $\sigma_{11} = \limsup_{d\to\infty} \sigma_{11}^{(d)} \le 1$ and $\sigma_{12} = \limsup_{d\to\infty} \sigma_{12}^{(d)} \le 1$ are finite constants. This gives us our desired result. ■

Proof of Theorem 4.6. Let $\boldsymbol{S}$, $\boldsymbol{R}$, and $\mathbf{\Pi}$ be vectors of string signs, string ranks, and anti-ranks as defined in Section 4.1. Recall that the runs statistic is given by

$$T_R = 1 + \sum_{i=1}^{n-1} \text{I}\{S_{\pi_i} \ne S_{\pi_{i+1}}\}.$$

When P is spherically symmetric, $\boldsymbol{S} \sim \text{Unif}(\{0,1\}^n)$, $\mathbf{\Pi} \sim \text{Unif}(\mathcal{S}_n)$ and they are independent (see Theorem 4.1). Let us define the filtration $\mathcal{F}_{n,t}^{(d)} = \mathcal{C}(S_{\pi_1}, \ldots, S_{\pi_t})$ for all $1 \le t \le n$ and $\mathcal{F}_{n,0}^{(d)}$ be the trivial σ-field. Note that the sequence $\{V_{n,i}^{(d)}\}_{0\le i\le n, n, d\ge 1}$, with $V_{n,0}^{(d)} = 0$ and $V_{n,t}^{(d)} = \sum_{i=1}^t \text{I}\{S_{\pi_i} \ne S_{\pi_{i+1}}\} - \frac{t}{2}$ for all $1 \le t \le n$, forms a triangular array of martingales adapted to the sequence of filtration $(\mathcal{F}_{n,t}^{(d)})_{1\le t\le n; n, d\ge 1}$. So, we can use the martingale central limit theorem (CLT) (see Brown, 1971) to derive the limiting null distribution of T_R.

First, let us look at the triangular array of martingale difference

$$Y_{n,i}^{(d)} = V_{n,i}^{(d)} - V_{n,i-1}^{(d)} = \text{I}\{S_{\pi_i} \ne S_{\pi_{i+1}}\} - 1/2, \quad i = 1, 2, \ldots, n.$$

One can see that $\left(\sigma_{n,i}^{(d)}\right)^2 = \mathbb{E}[\left(Y_{n,i}^{(d)}\right)^2 | \mathcal{F}_{n,i}^{(d)}] = 1/4$ for all $i = 1, 2, \ldots, n$. Therefore, $\left(\Delta_n^{(d)}\right)^2 = \sum_{i=1}^n \left(\sigma_{n,i}^{(d)}\right)^2 = n/4$ is a deterministic sequence of real numbers. Also, using $|Y_{n,i}^{(d)}| \le 1/2$, for any $\epsilon > 0$, we get

$$\frac{1}{\left(\Delta_n^{(d)}\right)^2}\sum_{j=1}^n \mathbb{E}\left[\left(Y_{n,j}^{(d)}\right)^2 \text{I}[|Y_{n,j}^{(d)}| \ge \epsilon\Delta_n^{(d)}]\right] \le \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\text{I}\left[|Y_{n,j}^{(d)}| \ge \epsilon\frac{n}{4}\right]\right] = \mathbb{P}\left[|Y_{n,1}^{(d)}| \ge \epsilon\frac{n}{4}\right] \to 0$$

as $n, d \to \infty$. So, applying the martingale CLT (Theorem 1 from Brown, 1971), we get

$$\frac{V_n^{(d)}}{\Delta_n^{(d)}} \xrightarrow{D} \mathcal{N}_1(0,1) \text{ or equivalently } \frac{V_n^{(d)}}{\sqrt{n}} \xrightarrow{D} \mathcal{N}_1\left(0, \frac{1}{4}\right), \text{as } n, d \to \infty.$$

Since $T_R = 1 + V_{n-1} + \frac{n-1}{2} = V_{n-1} + \frac{n+1}{2}$, we have $n^{-1/2}(T_R - \frac{n+1}{2}) \xrightarrow{D} \mathcal{N}_1\left(0, \frac{1}{4}\right)$ as $n, d \to \infty$. ■

Proof of Theorem 4.7. When P is not spherically symmetric, for $T_R = 1 + \sum_{i=1}^{n-1} \text{I}\{S_{\pi_i} \ne S_{\pi_{i+1}}\}$, we have

$$\text{Var}\left[\frac{T_R - \mathbb{E}[T_R]}{\sqrt{n}}\right] = \frac{1}{n}\left[\sum_{i=1}^{n-1} \text{Var}\left[\text{I}\{S_{\pi_i} \ne S_{\pi_{i+1}}\}\right] + \sum_{1\le i\ne j\le n-1} \text{Cov}\left(\text{I}\{S_{\pi_i} \ne S_{\pi_{i+1}}\}, \text{I}\{S_{\pi_j} \ne S_{\pi_{j+1}}\}\right)\right].$$

Now by Theorem 4.1, $\mathrm{Cov}\big(\mathrm{I}\{S_{\pi_i} \neq S_{\pi_{i+1}}\}, \mathrm{I}\{S_{\pi_j} \neq S_{\pi_{j+1}}\}\big) = 0$ for all $|i-j| > 2$.. So, as $n, d \to \infty$,

$$\limsup_{n,d\to\infty} \mathrm{Var}\left[\frac{T_R - \mathbb{E}[T_R]}{\sqrt{n}}\right]$$
$$= \limsup_{n,d\to\infty} \frac{1}{n}\left[\sum_{i=1}^{n-1} \mathrm{Var}\big[\mathrm{I}\{S_{\pi_i} \neq S_{\pi_{i+1}}\}\big] + \sum_{|i-j|\leq 2} \mathrm{Cov}\big(\mathrm{I}\{S_{\pi_i} \neq S_{\pi_{i+1}}\}, \mathrm{I}\{S_{\pi_j} \neq S_{\pi_{j+1}}\}\big)\right] \to \sigma^2,$$

where $\sigma^2 = \limsup_{d\to\infty}\Big[\mathrm{Var}\big[\mathrm{I}\{S_{\pi_1} \neq S_{\pi_2}\}\big] + \mathrm{Cov}\big(\mathrm{I}\{S_{\pi_1} \neq S_{\pi_2}\}, \mathrm{I}\{S_{\pi_2} \neq S_{\pi_3}\}\big) + \mathrm{Cov}\big(\mathrm{I}\{S_{\pi_1} \neq S_{\pi_2}\}, \mathrm{I}\{S_{\pi_3} \neq S_{\pi_4}\}\big)\Big]$ which is a non-negative finite constant. This gives us our desired result. ■

Proof of Theorem 4.8. Part (a) of the theorem can be proved using arguments identical to the proof of Theorem 4.1. Therefore, to avoid repetition, we omit it and give a detailed proof of part (b) only. Note that

$$\mathbb{E}\left[\frac{1}{d^\gamma}\sum_{j=1}^{d}(\mathbf{X}_1)_j^2(\mathbf{X}_2)_j^2\right] = \frac{1}{d^\gamma}\mathrm{trace}(\mathbf{D}^2) \text{ and}$$
$$\mathbb{E}\left[\frac{1}{d^\gamma}\sum_{j=1}^{d}(\mathbf{X}_1)_j^2(\mathbf{X}_2')_j^2\right] = \mathbb{E}\left[\frac{1}{d^\gamma}\|\mathbf{X}_2\|^2\sum_{j=1}^{d}(\mathbf{X}_1)_j^2(\mathbf{U}_2)_j^2\right] = \frac{1}{d^\gamma}\mathbb{E}\left[\|\mathbf{X}_2\|^2\right]\mathbb{E}\left[\sum_{j=1}^{d}(\mathbf{X}_1)_j^2(\mathbf{U}_2)_j^2\right]$$
$$= \frac{1}{d^\gamma}\mathrm{trace}(\mathbf{D})\left[\frac{1}{d}\sum_{j=1}^{d}\mathbb{E}(\mathbf{X}_1)_j^2\right] = \frac{1}{d^{1+\gamma}}\Big(\mathrm{trace}(\mathbf{D})\Big)^2.$$

Similarly one can also show that $\mathbb{E}\left[\frac{1}{d^\gamma}\sum_{j=1}^{d}(\mathbf{X}_1')_j^2(\mathbf{X}_2')_j^2\right] = \frac{1}{d^{1+\gamma}}\Big(\mathrm{trace}(\mathbf{D})\Big)^2$. Now, by Assumption (A4.3), $\left|\frac{1}{d^\gamma}\sum_{j=1}^{d}(\mathbf{X}_1)_j^2(\mathbf{X}_2)_j^2 - \frac{1}{d^\gamma}\mathrm{trace}(\mathbf{D}^2)\right|$, $\left|\frac{1}{d^\gamma}\sum_{j=1}^{d}(\mathbf{X}_1)_j^2(\mathbf{X}_2')_j^2 - \frac{1}{d^{1+\gamma}}\big(\mathrm{trace}(D)\big)^2\right|$ and $\left|\frac{1}{d^\gamma}\sum_{j=1}^{d}(\mathbf{X}_1')_j^2(\mathbf{X}_2')_j^2 - \frac{1}{d^{1+\gamma}}\big(\mathrm{trace}(D)\big)^2\right|$ converge to 0 in probability as d diverges to infinity. Therefore, if Assumption (A4.3) holds and

$$\liminf_{d\to\infty}\left\{\frac{1}{d^\gamma}\mathrm{trace}(D^2) - \frac{1}{d^{1+\gamma}}\Big(\mathrm{trace}(D)\Big)^2\right\} > 0,$$

we have $\mathbb{P}\left[\tilde{\theta}(\mathbf{X}_1, \mathbf{X}_2) < \tilde{\theta}(\mathbf{X}_1, \mathbf{X}_2')\right] \to 1$ as $d \to \infty$. Similarly, $\mathbb{P}\left[\tilde{\theta}(\mathbf{X}_1, \mathbf{X}_2) < \tilde{\theta}(\mathbf{X}_1', \mathbf{X}_2)\right] \to 1$ and $\mathbb{P}\left[\tilde{\theta}(\mathbf{X}_1, \mathbf{X}_2) < \tilde{\theta}(\mathbf{X}_1', \mathbf{X}_2')\right] \to 1$ as d diverges to infinity. Now, the result can be proved using the same argument as used in the proof of Theorem 4.3. ■

Proof of Theorem 4.9. First note that the tests statistics T_S and $\widetilde{T}_S$ take values in $\{0, 1, \ldots, n\}$ and $T_R, \widetilde{T}_R$ take values in $\{1, 2, \ldots, n\}$. Now, if condition (a) holds, then by Theorem 4.3, we get $T_S \xrightarrow{P} n$ as $d \to \infty$, and by Theorem 4.6, we get $T_R \xrightarrow{P} 1$ as $d \to \infty$. On the other hand, if condition (b) holds, then by Theorem 4.8 (b), we get $\widetilde{T}_S \to n$ and $\widetilde{T}_R \to 1$ in probability as d diverges to infinity. So, if either of conditions (a) or (b) holds, then $T_S^M = \max\{T_S, \widetilde{T}_S\} \xrightarrow{P} n$ and $T_R^M = \min\{T_R, \widetilde{T}_R\} \xrightarrow{P} 1$ as d diverges to infinity. ■

Proof of Theorem 4.10. Let $\boldsymbol{S}$ and $\boldsymbol{\Pi}$ be as defined in (4.1) and $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{\Pi}}$ be as defined in (4.5). Under spherical symmetry of P, it is easy to see that the elements of the sequence $\{(S_i, \tilde{S}_i)\}_{1\le i\le n}$ are mutually independent. However, for each $i = 1, \ldots, n$, S_i and $\tilde{S}_i$ may be dependent.

Now, to find the joint limiting distribution of $(T_S, \tilde{T}_S) = \sum_{i=1}^n (S_i, \tilde{S}_i)$, we first find the joint distribution of $\frac{1}{\sqrt{n}} \sum_{i=1}^n \big(t_1(S_i - \frac{1}{2}) + t_2(\tilde{S}_i - \frac{1}{2})\big)$ for $t_1, t_2 \in \mathbb{R}$. Note that the sequence $\{W_i^{(d)} := \big(t_1(S_i - \frac{1}{2}) + t_2(\tilde{S}_i - \frac{1}{2})\big)\}_{1\le i\le n}$ forms a triangular array of row wise i.i.d. bounded random variables, where $\mathbb{E}[W_1^{(d)}] = 0$ and

$$\mathrm{Var}[W_1^{(d)}] = t_1^2 \mathrm{Var}[S_1] + t_2^2 \mathrm{Var}[\tilde{S}_1] + 2\ t_1\ t_2\ \mathrm{Cov}(S_1, \tilde{S}_1) = t_1^2/4 + t_2^2/4 + 2\ t_1\ t_2\ \mathrm{Cov}(S_1, \tilde{S}_1).$$

We know that for a row-wise i.i.d. triangular array of bounded random variables, Lyapunov's condition holds trivially. Therefore, by Lyapunov's CLT, we get

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n W_i^{(d)} \xrightarrow{D} \mathcal{N}_1(0, \sigma^2)$$

as n and d diverge to infinity, where $\sigma^2 = \lim_{d\to\infty} \mathrm{Var}[W_1^{(d)}] = t_1^2/4 + t_2^2/4 + 2\ t_1\ t_2\ (\sigma_s^2 - \frac{1}{4})$. Since this distributional convergence holds irrespective of $t_1, t_2 \in \mathbb{R}$, using Cramer-Wold device, we get

$$\frac{1}{\sqrt{n}} \left\{ (T_S, \tilde{T}_S) - \Big(\frac{n}{2}, \frac{n}{2}\Big) \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \Big(S_i, \tilde{S}_i\Big) - \Big(\frac{1}{2}, \frac{1}{2}\Big) \right\} \xrightarrow{D} \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_S),$$

where the diagonal elements of $\boldsymbol{\Sigma}_S$ are $\frac{1}{4}$ and the off-diagonal element is $\big(\sigma_s^2 - \frac{1}{4}\big)$. Now, applying the continuous mapping theorem, we get

$$\frac{1}{\sqrt{n}} \left\{ \max(T_S, \tilde{T}_S) - \frac{n}{2} \right\} \xrightarrow{D} \max\{Z_1, Z_2\},$$

where $(Z_1, Z_2) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_S)$. This completes part (a) of the theorem.

For finding the joint limiting distribution of $(T_R, \tilde{T}_R)$, note that when P is a spherically symmetric, $\{M_{n,i}^{(d)} : t_1\big(\mathrm{I}\{S_{\pi(i)} \ne S_{\pi(i+1)}\} - \frac{1}{2}\big) + t_2\big(\mathrm{I}\{\tilde{S}_{\tilde{\pi}(i)} \ne \tilde{S}_{\tilde{\pi}(i+1)}\} - \frac{1}{2}\big)\}_{1\le i\le n}$ forms a triangular array of martingale differences w.r.t. the filtration $\{\mathcal{F}_{n,t}^{(d)} = \mathcal{C}(\{S_{\pi(i)}, \tilde{S}_{\tilde{\pi}(i)}\}_{1\le i\le t})\}_{1\le t\le n, n\ge 1, d\ge 1}$. Here, for all $1 \le i \le n$,

$$\begin{aligned}
\sigma_{n,i}^2 &= \mathbb{E}\big[\big(M_{n,i}^{(d)}\big)^2 | \mathcal{F}_{n,i}\big] = t_1^2/4 + t_2^2/4 + 2\ t_1 t_2\ \mathrm{Cov}\big(\mathrm{I}\{S_{\pi(i)} \ne S_{\pi(i+1)}\}\mathrm{I}\{\tilde{S}_{\tilde{\pi}(i)} \ne \tilde{S}_{\tilde{\pi}(i+1)}\}\big) \\
&= t_1^2/4 + t_2^2/4 + 2\ t_1 t_2\ \mathrm{Cov}\big(\mathrm{I}\{S_{\pi(1)} \ne S_{\pi(2)}\}, \mathrm{I}\{\tilde{S}_{\tilde{\pi}(1)} \ne \tilde{S}_{\tilde{\pi}(2)}\}\big) \qquad \text{(by Theorem 4.1 and 4.8)}
\end{aligned}$$

Now using the same arguments as in Theorem 4.6 and martingale CLT, as $n, d \to \infty$, we get

$$\frac{1}{\sqrt{n}} \Big(\sum_{i=1}^n M_{n,i}^{(d)} - \frac{n+1}{2} \Big) \xrightarrow{D} \mathcal{N}_1(0, \sigma^2), \text{ where } \sigma^2 = \frac{t_1^2}{4} + \frac{t_2^2}{4} + 2t_1 t_2 \big(\sigma_r^2 - \frac{1}{4}\big).$$

Therefore, applying the Cramer-Wold device and continuous mapping theorem, we get

$$\frac{1}{\sqrt{n}} \left\{ \min(T_R, \tilde{T}_R) - \frac{n+1}{2} \right\} \xrightarrow{D} \min\{Z_1', Z_2'\},$$

where $(Z_1', Z_2') \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_R)$ with $\boldsymbol{\Sigma}_R$ having all diagonal elements equal to $\frac{1}{4}$ and all off-diagonal elements equal to $\big(\sigma_r^2 - \frac{1}{4}\big)$. This completes part (b) of the theorem. ■

Proof of Theorem 4.11. If the underlying distribution is spherically symmetric, as n and d diverge to infinity, $T_S^M/n = \max\{T_S, \tilde{T}_S\}/n \xrightarrow{P} 1/2$ and $T_R^M/n = \min\{T_R, \tilde{T}_R\}/n \xrightarrow{P} 1/2$ (follows from Theorem 4.10). So, the cut-offs of the modified tests converge to 0.5. Now, from Theorems 4.5 and 4.7, we have $|T_S - \mathbb{E}(T_S)|/n \xrightarrow{P} 0$ and $|T_R - \mathbb{E}(T_R)|/n \xrightarrow{P} 0$ as n, d diverge to infinity. Following the same idea, one can prove this property for $\widetilde{T}_S$ and $\widetilde{T}_R$ as well.

(a) Therefore, when $\liminf_{n,d\to\infty} \mathbb{E}[T_S/n] > 0.5$ or $\liminf_{n,d\to\infty} \mathbb{E}[\tilde{T}_S/n] > 0.5$, T_S^M takes value bigger than 0.5 with probability converging to one. This implies that the power of the modified sign test based on T_S^M converges to one as n and d diverge to infinity.

(b) Similarly, as d and n grow to infinity, we have the consistency of modified runs test based on T_R^M when $\limsup_{n,d\to\infty} \mathbb{E}[T_R/n] < 0.5$ or $\limsup_{n,d\to\infty} \mathbb{E}[\tilde{T}_R/n] < 0.5$. ■

Proof of Lemma 4.2. Since $\mathbf{X}_1, \mathbf{X}_2$ are symmetric about $\boldsymbol{\mu}$, the characteristic function of $\mathbf{X}_1$ is of the form $\varphi(\boldsymbol{t}) = \exp\{i\langle \boldsymbol{t}, \boldsymbol{\mu}\rangle\} g(\boldsymbol{t})$, where $g(\cdot)$ is some real-valued function with $g(\boldsymbol{t}) = g(-\boldsymbol{t})$ for all $\boldsymbol{t}$. Note that if $\mathbf{X}_1, \mathbf{X}_2$ are spherically symmetric about $\boldsymbol{\mu}$, then it is trivial to show that $\mathbf{X}_1 - \mathbf{X}_2$ is spherically symmetric about zero. Therefore, we only prove the if part.

If $\mathbf{X}_1 - \mathbf{X}_2$ is spherically symmetric about zero, then its characteristic function is of the form $f(\|\boldsymbol{t}\|)$ where $f(\cdot)$ is some real-valued function. Also, note that

$$\varphi_{\mathbf{X}_1 - \mathbf{X}_2}(\boldsymbol{t}) = \varphi_{\mathbf{X}_1}(\boldsymbol{t})\varphi_{-\mathbf{X}_2}(\boldsymbol{t}) = \varphi_{\mathbf{X}_1}(\boldsymbol{t})\varphi_{\mathbf{X}_1}(-\boldsymbol{t}) = g^2(\boldsymbol{t}).$$

Hence, $f(\cdot)$ is non-negative and $g^2(\boldsymbol{t}) = f(\|\boldsymbol{t}\|)\ \forall \boldsymbol{t} \in \mathbb{R}^d$. Therefore, $\varphi_{\mathbf{X}_1}(\boldsymbol{t}) = \exp\{i\langle \boldsymbol{t}, \boldsymbol{\mu}\rangle\} h(\|\boldsymbol{t}\|)$, where $|h(\|\boldsymbol{t}\|)| = f^{1/2}(\|\boldsymbol{t}\|)$. This gives us the desired result. ■

4.7.1 Some additional mathematical details

Lemma A4.1. *If $\mathbf{X}_1$ and $\mathbf{X}_2$ are two independent realizations of a d-dimensional random vector $\mathbf{X} \sim \mathrm{P}$, and $\mathbf{X}_1' = \|\mathbf{X}_1\|\mathbf{U}_1, \mathbf{X}_2' = \|\mathbf{X}_2\|\mathbf{U}_2$ (where $\mathbf{U}_1, \mathbf{U}_2 \overset{iid}{\sim} \mathrm{Unif}(S^{d-1})$) are their respective spherically symmetric variants, then $\langle \mathbf{X}_1, \mathbf{X}_2'\rangle \overset{D}{=} \langle \mathbf{X}_1', \mathbf{X}_2'\rangle$.*

Proof. We shall prove this result using the fact that for any $\boldsymbol{a} \in \mathbb{R}^d$, $\boldsymbol{a}^\top \mathbf{U}_i \overset{D}{=} \|\boldsymbol{a}\| U_{i,1}$ $(i = 1, 2)$, where $U_{i,1}$ is the first component of $\mathbf{U}_i$. Now, note that for any $t \in \mathbb{R}$,

$$\mathbb{E}[\exp\{it\langle \mathbf{X}_1, \mathbf{X}_2'\rangle\}] = \mathbb{E}[\exp\{it\|\mathbf{X}_2\|\langle \mathbf{X}_1, \mathbf{U}_2\rangle\}] = \mathbb{E}[\exp\{it\|\mathbf{X}_2\|\|\mathbf{X}_1\|U_{2,1}\}] \quad \text{and}$$
$$\mathbb{E}[\exp\{it\langle \mathbf{X}_1', \mathbf{X}_2'\rangle\}] = \mathbb{E}[\exp\{it\|\mathbf{X}_1\|\|\mathbf{X}_2\|\langle \mathbf{U}_1, \mathbf{U}_2\rangle\}] = \mathbb{E}[\exp\{it\|\mathbf{X}_1\|\|\mathbf{X}_2\|U_{2,1}\}].$$

The equality of these two characteristic functions proves the result. ■

Lemma A4.2. *Let $\mathbf{X}_1, \mathbf{X}_2$ be independent copies of $\mathbf{X} \sim \mathrm{P}$, where $\mathbb{E}(\mathbf{X}) = \mathbf{0}_d$ and $Var(\mathbf{X}) = \boldsymbol{\Sigma}$. Let $\mathbf{X}_1'$ and $\mathbf{X}_2'$ be the spherically symmetric variants of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Then*

$$\mathbb{E}\left\{\frac{1}{d}(\mathbf{X}_1^\top \mathbf{X}_2)^2\right\} = \frac{1}{d}\mathrm{trace}(\boldsymbol{\Sigma}^2) \text{ and } \mathbb{E}\left\{\frac{1}{d}(\mathbf{X}_1^\top \mathbf{X}_2')^2\right\} = \mathbb{E}\left\{\frac{1}{d}(\mathbf{X}_1'^\top \mathbf{X}_2')^2\right\} = \left(\frac{1}{d}\mathrm{trace}(\boldsymbol{\Sigma})\right)^2.$$

Proof. Note that $(\mathbf{X}_1^\top \mathbf{X}_2)^2 = \text{trace}(\mathbf{X}_2\mathbf{X}_2^\top \mathbf{X}_1\mathbf{X}_1^\top)$, and hence we have

$$\mathbb{E}\left\{\frac{1}{d}(\mathbf{X}_1^\top \mathbf{X}_2^\top)^2\right\} = \mathbb{E}\left\{\frac{1}{d}\text{trace}(\mathbf{X}_2\mathbf{X}_2^\top \mathbf{X}_1\mathbf{X}_1^\top)\right\} = \frac{1}{d}\text{trace}\big(\mathbb{E}\{\mathbf{X}_2\mathbf{X}_2^\top\}\mathbb{E}\{\mathbf{X}_1\mathbf{X}_1^\top\}\big) = \frac{1}{d}\text{trace}(\boldsymbol{\Sigma}^2).$$

Now, $(\mathbf{X}_1^\top \mathbf{X}_2')^2 = \|\mathbf{X}_2\|^2(\mathbf{X}_1^\top \mathbf{U}_2)^2 = \|\mathbf{X}_2\|^2\text{trace}(\mathbf{U}_2\mathbf{U}_2^\top \mathbf{X}_1\mathbf{X}_1^\top)$, where $\mathbf{U}_2 \sim \text{Unif}(S^{d-1})$ is independent of $\mathbf{X}_1$ and $\mathbf{X}_2$. Also note that $\mathbb{E}\{\mathbf{U}_2\mathbf{U}_2^\top\} = \frac{1}{d}\mathbf{I}_d$ and $\mathbb{E}\{\|\mathbf{X}_2\|^2\} = \sum_{i=1}^d \mathbb{E}\{X_{2i}^2\} = \text{trace}(\boldsymbol{\Sigma})$. So, taking expectations, we get

$$\begin{aligned}\mathbb{E}\left\{\frac{1}{d}(\mathbf{X}_1^\top \mathbf{X}_2')^2\right\} = \mathbb{E}\|\mathbf{X}_2\|^2\mathbb{E}\left\{\frac{1}{d}\text{trace}(\mathbf{U}_2\mathbf{U}_2^\top \mathbf{X}_1\mathbf{X}_1^\top)\right\} &= \text{trace}(\boldsymbol{\Sigma})\ \frac{1}{d}\text{trace}\big(\mathbb{E}\{\mathbf{U}_2\mathbf{U}_2^\top\}\mathbb{E}\{\mathbf{X}_1\mathbf{X}_1^\top\}\big) \\ &= \text{trace}(\boldsymbol{\Sigma})\ \frac{1}{d}\text{trace}(\frac{1}{d}\mathbf{I}_d\boldsymbol{\Sigma}) = \left(\frac{1}{d}\text{trace}(\boldsymbol{\Sigma})\right)^2.\end{aligned}$$

Now, the proof follows from the fact that $\mathbf{X}_1^\top \mathbf{X}_2' \overset{D}{=} \boldsymbol{X_1'}^\top \mathbf{X}_2'$. (see Lemma A4.1). ■

Lemma A4.3. *Let $\boldsymbol{X}_1, \ldots, \mathbf{X}_n$ be independent copies of $\mathbf{X} \sim \text{P}$ and $\mathbf{X}_1', \ldots, \mathbf{X}_n'$ be their respective spherically symmetric variants. For $s \in \{0,1\}^n$, define $\boldsymbol{Y}_{s,i} = s_i\mathbf{X}_i + (1-s_i)\mathbf{X}_i'$ for each $i = 1, \ldots, n$. Then for the sign statistic T_S based on $\mathbf{X}_1, \ldots, \mathbf{X}_n$, we have*

$$\left|\frac{1}{n}\mathbb{E}(T_S) - \mathbb{P}\left[\theta(\boldsymbol{Y}_{S_1,1}, \mathbf{X}_2) + \theta(\mathbf{X}_2, \boldsymbol{Y}_{S_3,3}) \le \theta(\boldsymbol{Y}_{S_1,1}, \mathbf{X}_2') + \theta(\mathbf{X}_2', \boldsymbol{Y}_{S_3,3})\right]\right| \to 0 \text{ as } n, d \to \infty,$$

where S_1, S_3 are i.i.d. $\text{Unif}(\{0,1\})$.

Proof. Recall the vectors of string signs $\boldsymbol{S} = (S_1, \ldots, S_n)$, string ranks $\boldsymbol{R} = (R_1, \ldots, R_n)$ and anti-ranks $\boldsymbol{\Pi} = (\pi_1, \ldots, \pi_n)$ as defined in Section 4.1. Now,

$$\mathbb{E}\left[\frac{T_S}{n}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n S_i\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n S_{\pi_i}\right] = \frac{1}{n}\Big(\mathbb{P}[S_{\pi_1} = 1] + \sum_{i=2}^{n-1}\mathbb{P}[S_{\pi_i} = 1] + \mathbb{P}[\pi_n = 1]\Big). \quad (4.10)$$

From the proof of Theorem 4.1, we have

$$\begin{aligned}\mathbb{P}[S_{\pi_1} = 1 \mid \boldsymbol{S}_{-\pi_1}] &= \mathbb{P}\big[\theta(\mathbf{X}_{\pi_1}, \boldsymbol{Y}_{S_{\pi_2},\pi_2}) \le \theta(\mathbf{X}_{\pi_1}', \boldsymbol{Y}_{S_{\pi_2},\pi_2})\big|S_{\pi_2}\big], \\ \mathbb{P}[S_{\pi_k} = 1 \mid \boldsymbol{S}_{-\pi_k}] &= \mathbb{P}\big[\theta(\boldsymbol{Y}_{S_{\pi_{k-1}},\pi_{k-1}}, \mathbf{X}_{\pi_k}) + \theta(\mathbf{X}_{\pi_k}, \boldsymbol{Y}_{S_{\pi_{k+1}},\pi_{k+1}}) \\ &\qquad \le \theta(\boldsymbol{Y}_{S_{\pi_{k-1}},\pi_{k-1}}, \mathbf{X}_{\pi_k}') + \theta(\mathbf{X}_{\pi_k}', \boldsymbol{Y}_{S_{\pi_{k+1}},\pi_{k+1}})\big|S_{\pi_{k-1}}, S_{\pi_{k+1}}\big], \\ \mathbb{P}[S_{\pi_n} = 1 \mid \boldsymbol{S}_{-\pi_n}] &= \mathbb{P}\big[\theta(\boldsymbol{Y}_{S_{\pi_{n-1}},\pi_{n-1}}, \mathbf{X}_{\pi_n}) \le \theta(\boldsymbol{Y}_{S_{\pi_{n-1}},\pi_{n-1}}, \mathbf{X}_{\pi_n}')\big|S_{\pi_{n-1}}\big].\end{aligned}$$

The distribution of $\boldsymbol{\Pi}$ does not depend on the distribution P, and it follows $\text{Unif}(\mathcal{S}_n)$. Therefore,

$$\begin{aligned}\mathbb{P}[S_{\pi_1} = 1] &= \mathbb{E}\left[\mathbb{P}[S_{\pi_1} = 1 \mid S_{-\pi_1}]\right] \\ &= \mathbb{E}\left[\mathbb{P}[\theta(\mathbf{X}_{\pi_1}, \boldsymbol{Y}_{S_{\pi_2},\pi_2}) \le \theta(\mathbf{X}_{\pi_1}', \boldsymbol{Y}_{S_{\pi_2},\pi_2}) \mid S_{\pi_2}]\right] \\ &= \mathbb{P}[S_{\pi_2} = 1]\ \mathbb{P}[\theta(\mathbf{X}_{\pi_1}, \boldsymbol{Y}_{1,\pi_2}) \le \theta(\mathbf{X}_{\pi_1}', \boldsymbol{Y}_{1,\pi_2})] \\ &\qquad + \mathbb{P}[S_{\pi_2} = 0]\ \mathbb{P}[\theta(\mathbf{X}_{\pi_1}, \boldsymbol{Y}_{0,\pi_2}) \le \theta(\mathbf{X}_{\pi_1}', \boldsymbol{Y}_{0,\pi_2})] \\ &= \mathbb{P}[S_2 = 1]\ \mathbb{P}[\theta(\mathbf{X}_1, \boldsymbol{Y}_{1,2}) \le \theta(\mathbf{X}_1', \boldsymbol{Y}_{1,2})] + \mathbb{P}[S_2 = 0]\ \mathbb{P}[\theta(\mathbf{X}_1, \boldsymbol{Y}_{0,2}) \le \theta(\mathbf{X}_1', \boldsymbol{Y}_{0,2})] \\ &= \mathbb{P}[\theta(\mathbf{X}_1, \boldsymbol{Y}_{S_2,2}) \le \theta(\mathbf{X}_1', \boldsymbol{Y}_{S_2,2})].\end{aligned}$$

Similarly, we can also show that

$$\mathbb{P}[S_{\pi_k}=1]=\mathbb{P}\big[\theta(\boldsymbol{Y}_{S_1,1},\mathbf{X}_2)+\theta(\mathbf{X}_2,\boldsymbol{Y}_{S_3,3})\le\theta(\boldsymbol{Y}_{S_1,1},\mathbf{X}_2')+\theta(\mathbf{X}_2',\boldsymbol{Y}_{S_3,3})\big],\ \text{for } k=2,\ldots,n-1$$
$$\mathbb{P}[S_n=1]=\mathbb{P}\big[\theta(\boldsymbol{Y}_{S_{n-1},n-1},\mathbf{X}_n)\le\theta(\boldsymbol{Y}_{S_{n-1},n-1},\mathbf{X}_n')\big]. \tag{4.11}$$

Therefore, combining (4.10) and (4.11), we get

$$\begin{aligned}&\left|\mathbb{E}\left[\frac{T_S}{n}\right]-\mathbb{P}\big[\theta(\boldsymbol{Y}_{S_1,1},\mathbf{X}_2)+\theta(\mathbf{X}_2,\boldsymbol{Y}_{S_3,3})\le\theta(\boldsymbol{Y}_{S_1,1},\mathbf{X}_2')+\theta(\mathbf{X}_2',\boldsymbol{Y}_{S_3,3})\big]\right|\\&=\left|\frac{1}{n}\big(\mathbb{P}[S_{\pi_1}=1]+\mathbb{P}[S_{\pi_n}=1]\big)\right|\to 0\end{aligned}$$

as n and d diverge to infinity simultaneously. This completes the proof. ■

4.7.2 Pitman efficiency of the linear rank statistic

We know that the linear rank tests for univariate data are Pitman efficient (see Hájek, Sidák & Sen, 1999). So, one may wonder whether the linear rank tests defined in Section 4.1 have the same property. In the following theorem, we address this issue for finite-dimensional data.

Theorem A4.1. *Let* P *be a spherically symmetric distribution and* Q *be a non-spherical distribution, which are mutually absolutely continuous and have densities $p(\cdot)$ and $q(\cdot)$, respectively, such that $\int|q(\mathbf{x})/p(\mathbf{x})-1|^3p(\mathbf{x})\mathrm{d}\mathbf{x}$ is finite. For any $\delta>0$, consider the contamination model*

$$F_n=\left(1-\frac{\delta}{\sqrt{n}}\right)\mathrm{P}+\frac{\delta}{\sqrt{n}}\mathrm{Q},$$

as a local asymptotically normal contiguous alternative (see Proposition 3.1 Banerjee & Ghosh, 2024a). Let $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ be n independent realizations of $\mathbf{X}\sim F_n$ and $\mathbf{X}_i'=\|\mathbf{X}_i\|\mathbf{U}_i$ for $i=1,2,\ldots,n$, where $\mathbf{U}_1,\mathbf{U}_2,\ldots,\mathbf{U}_n$ are independent $\mathrm{Unif}(S^{d-1})$ *random variables. Consider a sequence of uniformly bounded scores $\{a(i)\}_{1\le i\le n}$ with the following properties*

$$\frac{1}{n}\sum_{i=1}^n a^2(i)\to\sigma^2\ (\sigma^2>0),\quad \frac{1}{n}\sum_{i=1}^n a(i)\to\tau\quad\text{and}\quad\max_{1\le i\le n}\frac{a^2(i)}{\sum_{i=1}^n a^2(i)}\to 0\quad\text{as } n\to\infty. \tag{4.12}$$

Then as $n\to\infty$, we have

$$\frac{T_{LR}-\frac{1}{2}\sum_{i=1}^n a(i)}{\sqrt{\sum_{i=1}^n a^2(i)}}\xrightarrow{D}\mathcal{N}_1\left(\delta\frac{\tau}{\sigma}\Big(p-\frac{1}{2}\Big),\frac{1}{4}\right),$$

where $p=\mathbb{P}\left[\theta(\mathbf{X}_1,\boldsymbol{V})+\theta(\boldsymbol{V},\mathbf{X}_2)<\theta(\mathbf{X}_1,\boldsymbol{V}')+\theta(\boldsymbol{V}',\mathbf{X}_2)\right]$ for $\mathbf{X}_1,\mathbf{X}_2\overset{iid}{\sim}\mathrm{P}$, $\boldsymbol{V}\sim\mathrm{Q}$ and $\boldsymbol{V}'$ is a spherically symmetric variant of $\boldsymbol{V}$.

Proof. By Proposition 3.1 in Banerjee & Ghosh (2024a), we have

$$\log\left\{\prod_{i=1}^n\Big(1+\frac{\delta}{\sqrt{n}}\Big\{\frac{q(\mathbf{X}_i)}{p(\mathbf{X}_i)}-1\Big\}\Big)\right\}=\frac{\delta}{\sqrt{n}}\sum_{i=1}^n\left\{\frac{q(\mathbf{X}_i)}{p(\mathbf{X}_i)}-1\right\}-\frac{\delta^2}{2}\mathbb{E}\left\{\frac{q(\mathbf{X}_1)}{p(\mathbf{X}_1)}-1\right\}^2+o_P(1).$$

Note that by Jensen's inequality, the finiteness of $\int|q(\mathbf{u})/p(\mathbf{u})-1|^2p(\mathbf{u})\mathrm{d}\mathbf{u}$ follows from the finiteness

of $\int |q(\mathbf{u})/p(\mathbf{u}) - 1|^3 p(\mathbf{u})\mathrm{d}\mathbf{u}$. Now, let us look at the triangular array

$$\left\{ W_{ni} = t_1 \frac{a(i)(S_{\pi_i} - \frac{1}{2})}{\sqrt{\sum_{i=1}^n a^2(i)}} + t_2 \frac{\delta}{\sqrt{n}} \left\{ \frac{q(\mathbf{X}_{\pi_i})}{p(\mathbf{X}_{\pi_i})} - 1 \right\} \right\}_{1 \le i \le n, n \ge 1}.$$

Under H_0, this is an array of i.i.d. random variables with finite third moments. Now under the assumptions $\frac{1}{n}\sum_{i=1}^n a^2(i) \to \sigma^2$ and $\frac{1}{n}\sum_{i=1}^n a(i) \to \tau$, we have

$$\begin{aligned}
s_n^2 = \sum_{i=1}^n \mathrm{Var}(W_{ni}) &= \frac{t_1^2}{4} + t_2^2\delta^2\ \mathbb{E}\left\{ \frac{q(\mathbf{X}_1)}{p(\mathbf{X}_1)} - 1 \right\}^2 + 2t_1t_2\ \frac{\delta}{\sqrt{n}} \sum_{i=1}^n \mathbb{E}\left\{ \frac{a(i)(S_{\pi_i} - \frac{1}{2})}{\sqrt{\sum_{i=1}^n a^2(i)}} \left\{ \frac{q(\mathbf{X}_{\pi_i})}{p(\mathbf{X}_{\pi_i})} - 1 \right\} \right\} \\
&= \frac{t_1^2}{4} + t_2^2\delta^2\ \mathbb{E}\left\{ \frac{q(\mathbf{X}_1)}{p(\mathbf{X}_1)} - 1 \right\}^2 + 2t_1t_2\ \frac{\delta}{n} \sum_{i=1}^n \mathbb{E}\left\{ \frac{a(i)(S_{\pi_i} - \frac{1}{2})}{\sqrt{\frac{1}{n}\sum_{i=1}^n a^2(i)}} \left\{ \frac{q(\mathbf{X}_{\pi_i})}{p(\mathbf{X}_{\pi_i})} - 1 \right\} \right\} \\
&= \frac{t_1^2}{4} + t_2^2\delta^2\ \mathbb{E}\left\{ \frac{q(\mathbf{X}_1)}{p(\mathbf{X}_1)} - 1 \right\}^2 + 2t_1t_2\ \frac{\delta}{n} \sum_{i=2}^{n-1} \mathbb{E}\left\{ \frac{a(i)(S_{\pi_i})}{\sigma} \left\{ \frac{q(\mathbf{X}_{\pi_i})}{p(\mathbf{X}_{\pi_i})} - 1 \right\} \right\} + o(1) \\
&= \frac{t_1^2}{4} + t_2^2\delta^2\ \mathbb{E}\left\{ \frac{q(\mathbf{X}_1)}{p(\mathbf{X}_1)} - 1 \right\}^2 + 2t_1t_2\ \frac{\delta}{n} \sum_{i=2}^{n-1} \frac{a(i)}{\sigma} \mathbb{E}\left\{ S_{\pi_2} \left\{ \frac{q(\mathbf{X}_{\pi_2})}{p(\mathbf{X}_{\pi_2})} - 1 \right\} \right\} + o(1) \\
&= \frac{t_1^2}{4} + t_2^2\delta^2\ \mathbb{E}\left\{ \frac{q(\mathbf{X}_1)}{p(\mathbf{X}_1)} - 1 \right\}^2 + 2t_1t_2\ \delta\frac{\tau}{\sigma} \mathbb{E}\left\{ S_{\pi_2} \left\{ \frac{q(\mathbf{X}_{\pi_2})}{p(\mathbf{X}_{\pi_2})} - 1 \right\} \right\} + o(1).
\end{aligned}$$

Now $\mathrm{I}\Big\{S_{\pi_2} = 1\Big\} = \mathrm{I}\Big\{\theta(\boldsymbol{Y}_{S_{\pi_1},\pi_1}, \mathbf{X}_{\pi_2}) + \theta(\mathbf{X}_{\pi_2}, \boldsymbol{Y}_{S_{\pi_3},\pi_3}) \le \theta(\boldsymbol{Y}_{S_{\pi_1},\pi_1}, \mathbf{X}'_{\pi_2}) + \theta(\mathbf{X}'_{\pi_2}, \boldsymbol{Y}_{S_{\pi_3},\pi_3})\Big\}$, where $\boldsymbol{Y}_{S,i} = S\mathbf{X}_i + (1-S)\mathbf{X}'_i$ $(i = \pi_1, \pi_2, \pi_3)$. Clearly, $\mathbb{E}[S_{\pi_2}] = 1/2$ under spherical symmetry and

$$\begin{aligned}
&\mathbb{E}\left\{ S_{\pi_2} \left\{ \frac{q(\mathbf{X}_{\pi_2})}{p(\mathbf{X}_{\pi_2})} \right\} \middle| S_{\pi_1}, S_{\pi_3}, \pi_1, \pi_3 \right\} \\
&= \mathbb{P}\Big[\theta(\boldsymbol{Y}_{S_{\pi_1},\pi_1}, \boldsymbol{V}) + \theta(\boldsymbol{V}, \boldsymbol{Y}_{S_{\pi_3},\pi_3}) \le \theta(\boldsymbol{Y}_{S_{\pi_1},\pi_1}, \boldsymbol{V}') + \theta(\boldsymbol{V}', \boldsymbol{Y}_{S_{\pi_3},\pi_3}) \Big| S_{\pi_1}, S_{\pi_3}, \pi_1, \pi_3 \Big],
\end{aligned}$$

where $\boldsymbol{V} \sim \mathrm{Q}$ and $\boldsymbol{V}' = \|\mathbf{X}\|\mathbf{U}$ for $\mathbf{U} \sim \mathrm{Unif}(S^{d-1})$ independent of $\boldsymbol{V}$. Now, under H_0, S_{π_1}, S_{π_3} are i.i.d. $\mathrm{Unif}(\{0,1\})$ and π_1, π_3 are simple random samples without replacement from $\{1, 2, \ldots, n\}$ independent of $\boldsymbol{S}$. So, taking expectations with respect to S_{π_1}, S_{π_3} and π_1, π_3 we get

$$\begin{aligned}
&\mathbb{E}\left\{ S_{\pi_2} \left\{ \frac{q(\mathbf{X}_{\pi_2})}{p(\mathbf{X}_{\pi_2})} \right\} \right\} \\
&= \sum_{S_1,S_2 \in \{0,1\}} \sum_{\pi_1,\pi_3 \in \mathcal{S}_n} \frac{1}{2^2} \frac{1}{n(n-1)} \mathbb{P}\Big[\theta(\boldsymbol{Y}_{S_{\pi_1},\pi_1}, \boldsymbol{V}) + \theta(\boldsymbol{V}, \boldsymbol{Y}_{S_{\pi_3},\pi_3}) \le \theta(\boldsymbol{Y}_{S_{\pi_1},\pi_1}, \boldsymbol{V}') + \theta(\boldsymbol{V}', \boldsymbol{Y}_{S_{\pi_3},\pi_3})\Big] \\
&= \frac{1}{4} \sum_{S_1,S_2 \in \{0,1\}} \mathbb{P}\Big[\theta(\boldsymbol{Y}_{S_1,1}, \boldsymbol{V}) + \theta(\boldsymbol{V}, \boldsymbol{Y}_{S_2,2}) \le \theta(\boldsymbol{Y}_{S_1,1}, \boldsymbol{V}') + \theta(\boldsymbol{V}', \boldsymbol{Y}_{S_2,2})\Big] = p. \qquad (4.13)
\end{aligned}$$

However, since $\mathbf{X}_1$ and $\mathbf{X}_2$ are spherically symmetric under H_0, the probabilities on the right-hand side of (4.13) are all equal. Hence, the last equality follows. Therefore, under H_0,

$$s_n^2 \to \frac{t_1^2}{4} + t_2^2\delta^2\ \mathbb{E}\left\{ \frac{q(\mathbf{X}_1)}{p(\mathbf{X}_1)} - 1 \right\}^2 + 2t_1t_2\delta\frac{\tau}{\sigma}\big(p - \frac{1}{2}\big) \text{ as } n \to \infty.$$

Also, note that

$$\frac{1}{s_n^3}\sum_{i=1}^{n}\mathbb{E}\Big[\Big|W_{ni}-\mathbb{E}[W_{ni}]\Big|^3\Big]=\frac{1}{s_n^3}\sum_{i=1}^{n}\mathbb{E}\big[\Big|t_1\Big(\frac{a(i)(S_{\pi_i}-\frac{1}{2})}{\sqrt{\sum_{i=1}^n a^2(i)}}\Big)+t_2\Big(\frac{\delta}{\sqrt{n}}\Big\{\frac{q(\mathbf{X}_{\pi_i})}{p(\mathbf{X}_{\pi_i})}-1\Big\}\Big)\Big|^3\Big]$$

$$\leq 2^2|t_1|^3\frac{1}{s_n^3}\sum_{i=1}^{n}\mathbb{E}\Big[\Big|\Big(\frac{a(i)(S_{\pi_i}-\frac{1}{2})}{\sqrt{\sum_{i=1}^n a^2(i)}}\Big)\Big|^3\Big]+2^2|t_2|^3\frac{1}{s_n^3}\sum_{i=1}^{n}\mathbb{E}\Big[\Big|\Big(\frac{\delta}{\sqrt{n}}\Big\{\frac{q(\mathbf{X}_i)}{p(\mathbf{X}_i)}-1\Big\}\Big)\Big|^3\Big], \quad (4.14)$$

where the last inequality follows using $|a+b|^p \leq 2^{p-1}(|a|^p+|b|^p)$ for $p=3$. The first term in (4.14) goes to zero using (4.9). The second term in (4.14) is of the order $O(n^{-1/2})$ by assumption $\int \big|q(\mathbf{x})/p(\mathbf{x})-1\big|^3 \mathrm{d}\mathbf{x}<\infty$. Therefore, using Lyapunov's CLT, we also have $\sum_{i=1}^n W_{ni} \xrightarrow{D} \mathcal{N}_1(0,\sigma^2)$, where $\sigma^2=\frac{t_1^2}{4}+t_2^2\delta^2\,\mathbb{E}\left\{\frac{q(\mathbf{X}_1)}{p(\mathbf{X}_1)}-1\right\}^2+2t_1t_2\delta\frac{\tau}{\sigma}\big(p-\frac{1}{2}\big)$. Now, applying the Cramer-Wold device, we can conclude

$$\Big(\frac{\sum_{i=1}^n a(i)(S_{\pi_i}-\frac{1}{2})}{\sqrt{\sum_{i=1}^n a^2(i)}},\frac{\delta}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{q(\mathbf{X}_i)}{p(\mathbf{X}_i)}-1\Big\}\Big)\xrightarrow{D}\mathcal{N}_2(0,\mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is a 2×2 matrix with diagonal entries $1/4$ and $\delta^2\,\mathbb{E}\left\{\frac{q(\mathbf{X}_1)}{p(\mathbf{X}_1)}-1\right\}^2$, and the off-diagonal entry $\delta\frac{\tau}{\sigma}\big(p-\frac{1}{2}\big)$. Now using Le Cam's third lemma (see Van der Vaart, 1998), we get

$$\frac{\sum_{i=1}^n a(i)(S_{\pi_i}-\frac{1}{2})}{\sqrt{\sum_{i=1}^n a^2(i)}}\xrightarrow{D}\mathcal{N}\Big(\delta\frac{\tau}{\sigma}\Big(p-\frac{1}{2}\Big),\frac{1}{4}\Big)$$

as n diverges to infinity. This gives us our desired result. ■

Theorem A4.1 establishes that for a sequence of contiguous alternatives of the form $\{F_n:\ n\geq 1\}$ for which with $p\neq 0.5$, the limiting distribution of the linear rank test introduced in Section 4.1 is a non-centered normal distribution. Therefore, if the score functions satisfy assumption (4.12) and $p\neq 0.5$, the corresponding tests are Pitman efficient. See Section 4.5 for examples of distributions that satisfy this assumption.

Chapter 5

Two-Sample Test for Functional Data

In a two-sample problem, we test for the equality of two distributions F and G based on two sets of independent observations $\mathcal{X} = \{X_1, \ldots, X_n\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_m\}$ on $X \sim F$ and $Y \sim G$, respectively. We discussed this problem in Chapter 2, where F and G were two multivariate distributions. In this chapter, we assume X and Y to be two independent functional random variables lying in an infinite dimensional separable Hilbert space $\mathcal{H}$ with inner product $\langle ., . \rangle$. Let $\mathcal{B}(\mathcal{H})$ be the Borel σ-field on $\mathcal{H}$. Now, consider a random variable Z that takes values on $\mathcal{H}$. We know that (i) Z is $\mathcal{B}(\mathcal{H})$-measurable if and only if $\langle Z, f \rangle$ is measurable for all $f \in \mathcal{H}$ and (ii) the distribution of Z is uniquely determined by the distributions of $\langle Z, f \rangle$ over $f \in \mathcal{H}$ (see, e.g., Theorem 7.1.2 in Hsing & Eubank, 2015). So, two $\mathcal{H}$-valued random variables X and Y have the same distribution if and only if $\langle X, f \rangle$ and $\langle Y, f \rangle$ are identically distributed for all $f \in \mathcal{H}$. Now, consider any measure of difference $T(\cdot, \cdot)$ between two univariate distributions, which is non-negative and takes the value zero if and only if the two distributions are equal. One can use it to measure the difference between F^f and G^f, the distributions of $\langle X, f \rangle$ and $\langle Y, f \rangle$, and aggregate them over $f \in \mathcal{H}$ to construct a measure of dissimilarity between F and G. It can be expressed as

$$\eta^{\nu'}(F, G) = \int_{\mathcal{H}} T(F^f, G^f)\, d\nu'(f),$$

where ν' is a probability measure on $\mathcal{H}$. Note that $F = G$ implies $\eta^{\nu'}(F, G) = 0$. But $\eta^{\nu'}(F, G) = 0$ only implies $F^f = G^f$ almost everywhere w.r.t. ν', which does not necessarily imply $F = G$. In the multivariate case, if T is chosen as the squared L_2-distance between F^f and G^f and ν' is chosen as the uniform distribution over the surface of the unit sphere in $\mathbb{R}^d$, $\eta^{\nu'}(F, G)$ turns out to be the energy distance between F and G (Baringhaus & Franz, 2004), and in that case, $\eta^{\nu'}(F, G)$ has the characterization property, i.e., $\eta^{\nu'}(F, G) = 0$ implies $F = G$. If T is the Cramer-von-Mises distance between F^f and G^f, the same choice of ν' leads to the two-sample test statistic proposed in Kim, Balakrishnan & Wasserman (2020). It also has the characterization property. In these two cases, ν' being the uniform distribution, has support over the entire surface of the unit ball and hence considers all possible directions for projection. Keeping that in mind, we can consider a probability measure ν', whose support contains the unit sphere centered at the origin of the Hilbert space. In that case, $\eta^{\nu'}(\cdot, \cdot)$ has the characterization property, as shown in the following theorem.

Theorem 5.1. *If* $\text{supp}\{\nu'\}$ *contains the unit sphere in* $\mathcal{H}$*, then* $\eta^{\nu'}(F, G) = 0$ *if and only if* $F = G$.

However, note that the f's, which are orthogonal to $\text{supp}\{F\} \cup \text{supp}\{G\}$, do not contribute to $\eta^{\nu'}(F, G)$ even when the two random variables X and Y are highly separated. Therefore, it seems reasonable to discard those directions and work with $\nu' = (F + G)/2$, an equal mixture of F and G. It turns out that the characterization property of $\eta^{\nu'}(F, G)$ holds for this choice of ν' as well. This is formally stated in the following theorem.

Theorem 5.2. *If $\nu' = (F + G)/2$, then, $\eta^{\nu'}(F, G) = 0$ if and only if $F = G$.*

Throughout this chapter, we use $\nu' = (F + G)/2$ while T is taken as the measure proposed in Baringhaus & Franz (2010), which is defined as

$$T_\phi(\mathcal{L}_1, \mathcal{L}_2) = 2\mathbb{E}\ \phi\big(|U - V|^2\big) - \mathbb{E}\ \phi\big(|U - U'|^2\big) - \mathbb{E}\ \phi\big(|V - V'|^2\big),$$

where $U, U' \overset{iid}{\sim} \mathcal{L}_1$ and $V, V' \overset{iid}{\sim} \mathcal{L}_2$ are independent random variables, $\phi : [0, \infty) \to [0, \infty)$ is continuous, monotinically increasing function with $\phi(0) = 0$, and it has non-constant completely monotone derivative on $(0, \infty)$ with $\mathbb{E}\ \phi(|U|^2)$ and $\mathbb{E}\ \phi(|V|^2)$ being finite. For this choice of T, the measure of dissimilarity between F and G is given by

$$\eta_\phi(F, G) := \frac{1}{2}\int_{\mathcal{H}} T_\phi(F^f, G^f)dF(f) + \frac{1}{2}\int_{\mathcal{H}} T_\phi(F^f, G^f)dG(f).$$

Since $\eta_\phi(F, G)$ is obtained by aggregating the Baringhaus-Franz statistic T_ϕ computed along different projection directions, we call it the projected BF (pBF) criterion. It has a closed-form expression given by

$$\begin{aligned}\eta_\phi(F, G) = \mathbb{E}\phi\big(|\langle X_1, X_3\rangle - \langle Y_1, X_3\rangle|^2\big) - \frac{1}{2}\mathbb{E}\phi\big(|\langle X_1, X_3\rangle - \langle X_2, X_3\rangle|^2\big)\\ - \frac{1}{2}\mathbb{E}\phi\big(|\langle Y_1, X_3\rangle - \langle Y_2, X_3\rangle|^2\big) + \mathbb{E}\phi\big(|\langle X_1, Y_3\rangle - \langle Y_1, Y_3\rangle|^2\big)\\ - \frac{1}{2}\mathbb{E}\phi\big(|\langle X_1, Y_3\rangle - \langle X_2, Y_3\rangle|^2\big) - \frac{1}{2}\mathbb{E}\phi\big(|\langle Y_1, Y_3\rangle - \langle Y_2, Y_3\rangle|^2\big),\end{aligned}$$

where $X_i \overset{iid}{\sim} F\,(i = 1, 2, 3)$, $Y_i \overset{iid}{\sim} G\,(i = 1, 2, 3)$ are independent and $\mathbb{E}\ \phi(|\langle X_1, X_2\rangle|^2)$, $\mathbb{E}\ \phi(|\langle X_1, Y_1\rangle|^2)$ and $\mathbb{E}\ \phi(|\langle Y_1, Y_2\rangle|^2)$ are finite. The measure $\eta_\phi(F, G)$ has some nice theoretical properties as mentioned in the following proposition.

Proposition 5.1. *Suppose that $\phi : [0, \infty) \to [0, \infty)$ is a continuous, monotinically increasing function with $\phi(0) = 0$, and it has non-constant completely monotone derivative on $(0, \infty)$. Also assume that $\mathbb{E}\ \phi(|U|^2)$ and $\mathbb{E}\ \phi(|V|^2)$ are finite for all $f \in \mathcal{H}$, where $U \sim F^f$ and $V \sim G^f$. Then $\eta_\phi(F, G)$ has the following properties.*

(a) $\eta_\phi(F, G) = \mathbb{E}\{g(X_1, X_2, X_3; Y_1, Y_2, Y_3)\}$, *where*

$$\begin{aligned}g(X_1, X_2, X_3; Y_1, Y_2, Y_3) = \frac{1}{2}\Big\{2\phi\big(|\langle X_2, X_1\rangle - \langle Y_1, X_1\rangle|^2\big) - \phi\big(|\langle X_2, X_1\rangle - \langle X_3, X_1\rangle|^2\big)\\ - \phi\big(|\langle Y_2, X_1\rangle - \langle Y_3, X_1\rangle|^2\big) + 2\phi\big(|\langle X_1, Y_1\rangle - \langle Y_2, Y_1\rangle|^2\big)\\ - \phi\big(|\langle X_2, Y_1\rangle - \langle X_3, Y_1\rangle|^2\big) - \phi\big(|\langle Y_2, Y_1\rangle - \langle Y_3, Y_1\rangle|^2\big)\Big\}. \quad (5.1)\end{aligned}$$

(b) $\eta_\phi(F,G)$ *has the characterization property, i.e.,* $\eta_\phi(F,G)=0$ *if and only if* $F=G$.

(c) $\eta_\phi(F,G)$ *is invariant under unitary operations on* X *and* Y*, i.e., if* $U:\mathcal{H}\to\mathcal{H}_0$ *is an unitary operator, then* $\eta_\phi(F\circ U^{-1},G\circ U^{-1})=\eta_\phi(F,G)$.

(d) *Let* $\{X_n:n\geq 1\}$ *and* $\{Y_n:n\geq 1\}$ *be independent sequences of Hilbertian random variables such that* $X_n\xrightarrow{D}X$ *and* $Y_n\xrightarrow{D}Y$*. Then* $\lim_{n\to\infty}\eta_\phi(\mathcal{L}(X_n),\mathcal{L}(Y_n))=\eta_\phi(\mathcal{L}(X),\mathcal{L}(Y))$*, where* $\mathcal{L}(Z)$ *denotes the distribution of a random function* Z.

Remark 5.1. Proposition 5.1(c) implies that η_ϕ only depends on the inner product defined on the Hilbert space, but not on the space used for modeling the random variables. For example, modeling the two samples as random variables in $L_2[0,1]$ and in $L_2[0,10]$ leads to the same value of $\eta_\phi(F,G)$.

5.1 Estimation of pBF and construction of the two-sample test

Suppose $\hat{F}_n$ and $\hat{G}_m$ are the empirical distribution functions based on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Replacing F by $\hat{F}_n$ and G by $\hat{G}_m$, we get an estimator of $\eta_\phi(F,G)$, which can be expressed as

$$\begin{aligned}\hat{\eta}^{\phi}_{n,m}=&\frac{1}{n^2m}\sum_{1\leq i,j\leq n}\sum_{k=1}^{m}\phi\big(|\langle X_j,X_i\rangle-\langle Y_k,X_i\rangle|^2\big)-\frac{1}{2n^3}\sum_{1\leq i,j,k\leq n}\phi\big(|\langle X_j,X_i\rangle-\langle X_k,X_i\rangle|^2\big)\\&-\frac{1}{2nm^2}\sum_{i=1}^{n}\sum_{1\leq j,k\leq m}\phi\big(|\langle Y_j,X_i\rangle-\langle Y_k,X_i\rangle|^2\big)+\frac{1}{nm^2}\sum_{1\leq i,k\leq m}\sum_{j=1}^{n}\phi\big(|\langle X_j,Y_i\rangle-\langle Y_k,Y_i\rangle|^2\big)\\&-\frac{1}{2n^2m}\sum_{i=1}^{m}\sum_{1\leq j,k\leq n}\phi\big(|\langle X_j,Y_i\rangle-\langle X_k,Y_i\rangle|^2\big)-\frac{1}{2m^3}\sum_{1\leq i,j,k\leq m}\phi\big(|\langle Y_j,Y_i\rangle-\langle Y_k,Y_i\rangle|^2\big).\end{aligned}\tag{5.2}$$

Clearly, $\hat{\eta}^{\phi}_{n,m}$ can be viewed as a two-sample V-statistic with the core function

$$g^*(X_1,X_2,X_3;Y_1,Y_2,Y_3)=\frac{1}{3!3!}\sum_{\pi_1,\pi_2\in\mathcal{S}_3}g(X_{\pi_1(1)},X_{\pi_1(2)},X_{\pi_1(3)};Y_{\pi_2(1)},Y_{\pi_2(2)},Y_{\pi_2(3)}),\tag{5.3}$$

where g is as in equation (5.1). The large sample distributions of $\eta^{\phi}_{n,m}$ under null and alternative hypotheses follow from Section 4.2 of Lee (1990). These results are stated below as Theorem 5.3. Alternative derivations of these results are given in Section 5.3 as the proof of the theorem.

Theorem 5.3. *Let* $X_1,\ldots,X_n\overset{iid}{\sim}F$ *and* $Y_1,\ldots,Y_m\overset{iid}{\sim}G$ *be independent random functions and* $\lim n/(n+m)=\lambda\in[0,1]$*. Then for any* ϕ *satisfying the properties mentioned in Proposition 5.1, as* $\min\{n,m\}\to\infty$*, we have the following results.*

(a) *Under the alternative hypothesis* $H_1:F\neq G$, $\sqrt{nm/(n+m)}(\hat{\eta}^{\phi}_{n,m}-\eta_\phi(F,G))\xrightarrow{D}\mathcal{N}(0,\sigma^2)$*, where* $\sigma^2=(1-\lambda)\,Var(h_1^*(X_1))+\lambda\,Var(h_2^*(Y_1))$*, while* h_1^* *and* h_2^* *are the first order Hoeffding's projection of the core function defined in equation (5.3).*

(b) *Under the null hypothesis* $H_0:F=G$, $nm/(n+m)\hat{\eta}^{\phi}_{n,m}\xrightarrow{D}\sum_{k=1}^{\infty}\lambda_kZ_k^2$*, where the* Z_k*'s are i.i.d.* $\mathcal{N}_1(0,1)$ *and the* λ_k*'s are the eigenvalues of the integral operator* $T(g)(v)=\int h(u,v)g(u)dF(u)$ *and* $h(u,v)=\mathbb{E}\Big\{\phi\big(|\langle u,X_1\rangle,\langle X_2,X_1\rangle|^2\big)\Big\}+\mathbb{E}\Big\{\phi\big(|\langle v,X_1\rangle,\langle X_2,X_1\rangle|^2\big)\Big\}-\mathbb{E}\Big\{\phi\big(|\langle u,X_1\rangle-\langle v,X_1\rangle|^2\big)\Big\}-\mathbb{E}\Big\{\phi\big(|\langle X_2,X_1\rangle-\langle X_3,X_1\rangle|^2\big)\Big\}$.

As a consequence of Theorem 5.3, we get the probability convergence of $\hat{\eta}^{\phi}_{n,m}$ to its population counterpart $\eta_\phi(F,G)$. This is formally stated as a corollary.

Corollary 5.1. *If $X_1,\ldots,X_n \overset{iid}{\sim} F$ and $Y_1,\ldots,Y_m \overset{iid}{\sim} G$ are independent, $\hat{\eta}^{\phi}_{n,m}$ converges in probability to $\eta_\phi(F,G)$ as $\min\{n,m\}\to\infty$. This holds even when $n/(n+m)\to 0$ or 1.*

Hence even in the extremely unbalanced scenario (i.e., when $n/(n+m)\to 0$ or 1), our estimator can detect the distributional difference between the two samples $\mathcal{X}$ and $\mathcal{Y}$.

5.1.1 Two-sample test based on $\hat{\eta}^{\phi}_{n,m}$

Proposition 5.1(b) shows that for suitable choices of ϕ, we have $\eta_\phi(F,G)\geq 0$, where the equality holds if and only if $F=G$. Since $\hat{\eta}^{\phi}_{n,m}$ is a consistent estimator of $\eta(F,G)$, we can reject $H_0: F=G$ if $\hat{\eta}^{\phi}_{n,m}$ is large. Theorem 5.3 (b) gives us the limiting null distribution of the test statistics $\hat{\eta}^{\phi}_{n,m}$, but it involves some unknown quantities which are quite difficult to estimate. Hence, for a nominal level α $(0<\alpha<1)$, the cut-off is computed using the permutation method as described below.

- Let $\mathcal{U}^\pi = \{U_{\pi(1)},\ldots,U_{\pi(N)}\}$ denote a permutation of the pooled sample $\mathcal{U}$ based on the permutation π of $\{1,\ldots,N\}$.
- Partition $\mathcal{U}^\pi$ into $\mathcal{X}^\pi_n = \{U_{\pi(1)},\ldots,U_{\pi(n)}\}$ and $\mathcal{Y}^\pi_m = \{U_{\pi(n+1)},\ldots,U_{\pi(n+m)}\}$ and compute the statistic $\hat{\eta}^{\phi,\pi}_{n,m}$ (permutation analog of $\hat{\eta}^{\phi}_{n,m}$) based on them.
- Compute the critical value $c^{\phi}_{1-\alpha}$ as
$$c^{\phi}_{1-\alpha} = \inf\{t\in\mathbb{R} : \frac{1}{N!}\sum_{\pi\in\mathcal{S}_N} \mathrm{I}[\hat{\eta}^{\phi,\pi}_{n,m}\leq t]\geq 1-\alpha\},$$
where $\mathcal{S}_N$ is the set of all permutations of $\{1,\ldots,N\}$.

The proposed test rejects H_0 if $\hat{\eta}^{\phi}_{n,m}$ is larger than $c^{\phi}_{1-\alpha}$ or equivalently the corresponding p-value
$$p_{n,m} := \frac{1}{N!}\sum_{\pi\in\mathcal{S}_N} \mathrm{I}[\hat{\eta}^{\phi,\pi}_{n,m}\geq \hat{\eta}^{\phi}_{n,m}]$$
is smaller than α. Here, the cut-off $c^{\phi}_{1-\alpha}$ is a random quantity, but using the following theorem, one can prove that it converges to zero in probability as $\min\{n,m\}$ diverges to infinity.

Theorem 5.4. *If ϕ satisfies the conditions mentioned in Proposition 5.1, as $\min\{n,m\}\to\infty$ and $n/(n+m)\to\lambda$, $nm/(n+m)\hat{\eta}^{\phi,\pi}_{n,m}$ converges in distribution to $\sum_{k=1}^{\infty}\lambda_k Z_k^2$, where $\{Z_k\}$ is a sequence of independent $\mathcal{N}_1(0,1)$ variables and the λ_k's are the eigenvalues of the integral operator $T(g)(v) = \int h(u,v)g(u)dF(u)$ where $h(u,v) = \mathbb{E}\Big\{\phi\big(|\langle u,U_1\rangle,\langle U_2,U_1\rangle|^2\big)\Big\} + \mathbb{E}\Big\{\phi\big(|\langle v,U_1\rangle,\langle U_2,U_1\rangle|^2\big)\Big\} - \mathbb{E}\Big\{\phi\big(|\langle u,U_1\rangle - \langle v,U_1\rangle|^2\big)\Big\} - \mathbb{E}\Big\{\phi\big(|\langle U_2,U_1\rangle - \langle U_3,U_1\rangle|^2\big)\Big\}$ with $U_1,U_2,U_3 \overset{iid}{\sim} \lambda F + (1-\lambda)G$.*

In particular, under H_0, the permuted test statistic $nm/(n+m)\hat{\eta}^{\phi,\pi}_{n,m}$ and $nm/(n+m)\hat{\eta}^{\phi}_{n,m}$ attain the same limiting distribution as $\min\{n,m\}\to\infty$ and $n/(n+m)\to\lambda$. Hence, the permutation test asymptotically attains the significance level α, and it turns out to be consistent for any fixed alternative. Therefore, when $\min\{n,m\}$ is sufficiently large, the permutation test is close to the oracle test, which assumes the knowledge of the data-generating distribution.

However, in practice, it is not computationally feasible to consider all permutations even when N is moderately large. In such scenario, we generate random permutations $\pi_1, \ldots, \pi_B$ of the set $\{1, \ldots, N\}$ and obtain a randomized p-value

$$p_{n,m,B} = \frac{1}{B+1}\Big\{\sum_{i=1}^{B} \mathrm{I}[\hat{\eta}_{n,m}^{\phi,\pi_i} \geq \hat{\eta}_{n,m}^{\phi}] + 1\Big\}.$$

We have seen that using all $N!$ permutations leads to the p-value $p_{n,m}$. Naturally, one would expect $p_{n,m,B}$ and $p_{n,m}$ to be close as the number of random permutations B grows to infinity. This is asserted by the following proposition.

Proposition 5.2. *For any given $\mathcal{U}$, $p_{n,m,B}$ converges almost surely to $p_{n,m}$ as B grows to infinity.*

So, when B and $\min\{n, m\}$ are sufficiently large, the randomized permutation test (which is used in practice) also approximates the oracle test. However, the oracle test is never available, while the randomized permutation test is applicable in general. This strongly advocates the use of the randomized permutation test in practice.

5.1.2 Local asymptotic behaviour of the test

In this section, we construct a locally asymptotically normal sequence of contiguous alternatives and study the behaviour of our test under such alternatives. Suppose that $Z_1, \ldots, Z_N$ are i.i.d. functional random variables with distribution F. Define $F^{(N)} = (1 - \delta_N)F + \delta_N L$, where F and L are two probability distributions on $\mathcal{H}$, and $\{\delta_N\}$ is a sequence in $(0, 1)$ that converges to zero as N grows to infinity. Clearly, the total variation distance between $F^{(N)}$ and F converges to zero as N diverges to infinity. Hence, $F^{(N)}$ and F are mutually contiguous for any distribution L and a sequence $\{\delta_N\}$ in $(0, 1)$ that converges to zero as N increases. Now, for studying the local behavior of our test, we assume that

(A5.1) L is absolutely continuous with respect to F with square integrable density $\ell(\cdot)$.

Note that Assumption (A5.1), in particular, implies that the second-order central moment of $\ell(Z)$ is finite, and $F^{(N)}$ is absolutely continuous with respect to F. For $\delta_N = \delta/\sqrt{N}$, we have the following result on the local asymptotic normality for functional random variables.

Theorem 5.5. *Under Assumption (A5.1) and $\delta_N = \delta/\sqrt{N}$, the Radon-Nikodym derivative of $F^{(N)}$ with respect to F is $\Big(1 + \frac{\delta}{\sqrt{N}}\big(\ell(z) - 1\big)\Big)$, and as N goes to infinity, we have*

$$\left|\log\Big\{\prod_{i=1}^{N} \frac{dF^{(N)}}{dF}(Z_i)\Big\} - \frac{\delta}{\sqrt{N}}\sum_{i=1}^{N}\Big(\ell(Z_i) - 1\Big) + \frac{\delta^2}{2}\mathbb{E}\Big\{\ell(Z_1) - 1\Big\}^2\right| \xrightarrow{P} 0,$$

Since for functional random variables, there is no universally accepted dominating measure (such as the Lebesgue measure), the quadratic mean differentiability assumption (see Chapter 7, Van der Vaart, 1998) is difficult to formulate. But contiguity through contamination alternatives is naturally extendable in such cases. Let $F^{(n)} = F$ and $G^{(m)} = (1 - \delta/\sqrt{m})F + \delta/\sqrt{m}L$ for

some probability distribution L satisfying (A5.1) and a positive number δ. Clearly, the alternative $(F^{(n)}, G^{(m)})$ is contiguous with the null (F, F). The next theorem shows that under $(F^{(n)}, G^{(m)})$, $nm/(n+m)\hat{\eta}^{\phi}_{n,m}$ converges in distribution to a tight random variable.

Theorem 5.6. *Under $(F^{(n)}, G^{(m)})$, as $\min\{n, m\}$ grows to infinity, $nm/(n+m)\hat{\eta}^{\phi}_{n,m}$ converges in distribution to $\sum_{k=1}^{\infty} \lambda_k \big(Z_k - \sqrt{\lambda}\delta \int \varphi_k dL\big)^2$, where $\lim n/(n+m) = \lambda \in [0,1]$, $\{Z_k\}$ is a sequence of i.i.d. $\mathcal{N}_1(0,1)$ random variables and $\{\lambda_k\}$ and $\{\varphi_k\}$ are the eigenvalues and eigenfunctions of the integral equation $\int h(u,v)\gamma(v)dF(v) = \lambda\gamma(u)$, where for $U_1, U_2, U_3 \overset{iid}{\sim} F$,*

$$\begin{aligned} h(u,v) =& \mathbb{E}\{\phi\big(|\langle u, U_1\rangle, \langle U_2, U_1\rangle|^2\big)\} + \mathbb{E}\{\phi\big(|\langle v, U_1\rangle, \langle U_2, U_1\rangle|^2\big)\} \\ & - \mathbb{E}\{\phi\big(|\langle u, U_1\rangle - \langle v, U_1\rangle|^2\big)\} - \mathbb{E}\{\phi\big(|\langle U_2, U_1\rangle - \langle U_3, U_1\rangle|^2\big)\Big\}. \end{aligned}$$

To investigate the asymptotic behavior of the test, we also need to study the convergence of $\hat{\eta}^{\phi,\pi}_{n,m}$ under the sequence of alternatives $(F^{(n)}, G^{(m)})$. Intuitively, $\hat{\eta}^{\phi,\pi}_{n,m}$ has the same behavior as $\hat{\eta}^{\phi}_{n,m}$ when the samples are generated from the mixture distribution $\lambda_n F + (1-\lambda_n)G$, where $\lambda_n = n/(n+m)$. Therefore, under the contiguous alternative $(F^{(n)}, G^{(m)})$, $\hat{\eta}^{\phi,\pi}_{n,m}$ should have the same behavior as $\hat{\eta}^{\phi}_{n,m}$ when the sample is generated from $\lambda_n F^{(n)} + (1-\lambda_n)G^{(m)}$, which converges weakly to F as n and m diverge to infinity. So, it is expected that asymptotically, the limiting behavior of $\hat{\eta}^{\phi,\pi}_{n,m}$ should be the same as it is under H_0. This is formalized in the following theorem.

Theorem 5.7. *If ϕ satisfies the conditions of Proposition 5.1, under the contiguous alternative $(F^{(n)}, G^{(m)})$, as $\min\{n, m\}$ grows to infinity and $\lim n/(n+m) = \lambda \in (0,1)$, $nm/(n+m)\hat{\eta}^{\phi,\pi}_{n,m}$ converges in distribution to $\sum_{k=1}^{\infty} \lambda_k Z_k^2$ where $\{Z_k\}$ and $\{\lambda_k\}$ are as in Theorem 5.3.*

Theorems 5.6 and 5.7 together show that for a suitable choice of ϕ, under $(F^{(n)}, G^{(m)})$ the power of our test converges to a non-trivial limit which is a function of δ. This shows that the proposed test is statistically efficient in the Pitman sense. It can be easily verified that as δ diverges to infinity, the asymptotic power will be one, and it will be equal to the level α when δ shrinks to zero. The exact expression of the limit is not analytically tractable. In the next section, we compute the empirical power and efficiency of our test and compare them with some of the state-of-the-art methods through simulations.

5.2 Empirical performance of the proposed test

In this section, we evaluate the empirical performance of the proposed test by carrying out some simulated experiments and analyzing a real dataset. Note that in practice, one needs to choose a suitable function ϕ to implement the test. There are several choices of ϕ available in the literature, but here we take the functions $\phi_1(z) = \sqrt{z}$, $\phi_2(z) = 1 - \exp(-z/2)$ and $\phi_3(z) = \log(1+z)$ to construct our test, which we further refer to as pBF-ℓ_2, pBF-exp and pBF-log tests, respectively.

First, we consider some examples for studying the level property of our tests and then we compare their powers with the powers of the tests proposed in Pomann, Staicu & Ghosh

(2016), Wynne & Duncan (2020) and Pan et al. (2018), which are referred to as the FAD (Functional Anderson-Darling) test, the WD (Wynne-Duncan) test, and the BD (Ball Divergence) test, respectively. Throughout this section, all tests are considered to have a 5% nominal level. In our simulated experiments, the random functions are observed on an equispaced grid of size 101 on the interval $[0, 1]$, and the randomized p-values of the permutation tests are computed based on 500 random permutations. Each experiment was repeated 1000 times to estimate the power of a test by the proportion of times it rejected H_0.

5.2.1 Analysis of simulated data sets

First, we study the level properties of our tests. For this purpose, we generated $\mathcal{X}$ and $\mathcal{Y}$ from the same distribution. Here we consider three examples (Example 5.1-5.3) and compute the power (which is the same as the level when $F = G$) for different sample sizes ($n = m = 20, 30, 40$ and 50).

Example 5.1. X and Y are independent Wiener process W on $[0, 1]$.

Example 5.2. X and Y are independently distributed as $\mu + W$ on $[0, 1]$ where $\mu(t) = t$ and W is the Wiener process.

Example 5.3. X and Y are independent random functions defined as $\sum_{i=1}^{9} \frac{1}{i^{2.5}} \xi_i \psi_i(t)$, where the ξ_i's are i.i.d. $\mathcal{N}_1(0, 1)$ random variables, and $\{\psi_i\}$ is the trigonometric basis on $L_2([0, 1])$.

Figure 5.1 shows that the observed levels of our tests were approximately 0.05 in all three examples, which we expect in view of the theoretical results stated in the previous sections.

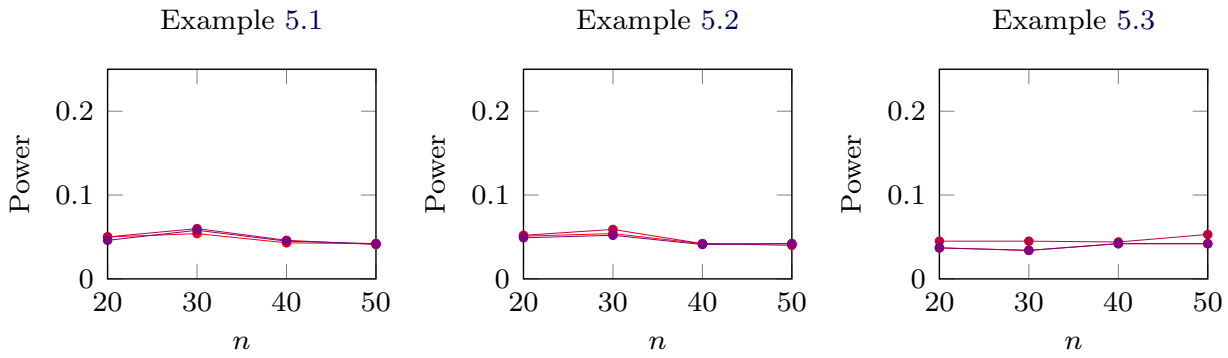


Fig. 5.1 *Powers of pBF-ℓ_2 (•), pBF-exp (•) and pBF-log (•) test in Examples 5.1-5.3.*

Next, we consider some location and scale alternatives (Examples 5.4 and 5.5) to compare the power of pBF-ℓ_2, pBF-exp, and pBF-log tests with FAD, WD, and BD tests.

Example 5.4. X is the Wiener process W on $[0, 1]$, while Y is distributed as $\mu + W$ and is independent of X. We consider two choices of μ, (i) $\mu(t) = rt^2$ and (ii) $\mu(t) = re^t$.

We carried out our experiment for different choices of $r \in [0, 1]$ as shown in Figure 5.2. We generated 50 observations on each of X and Y. Here X is pure noise, while Y has a non-zero

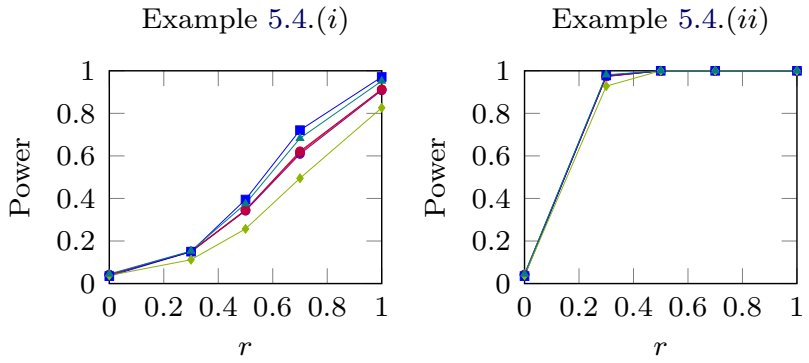


Fig. 5.2 *Powers of pBF-ℓ_2 (●), pBF-exp (●), pBF-log (●), FAD (■), BD (♦) and WD (▲) tests in Examples 5.4 (i) and (ii).*

signal μ. The location difference between the two distributions is an increasing function of r. So, as expected, the powers of all tests increased with r (see Figure 5.2). In both cases, FAD, WD, and pBF tests had competitive performance, but the performance of the BD test was relatively poor.

Example 5.5. X and Y are independent random functions as in Example 5.3 with respective coefficients denoted as ξ_i^X and ξ_i^Y for each $i \in \{1, \ldots, 9\}$. Here, we consider two scale problems: (i) ξ_i^Xs are i.i.d $\mathcal{N}_1(0,1)$, while ξ_i^Ys are i.i.d. $\mathcal{N}_1(0,\sigma^2)$; (ii) ξ_i^Xs are independent standard Cauchy variables and ξ_i^Ys are the centered Cauchy random variables with scale parameters $\sigma > 0$.

Here also we generated 50 observations on each of the random functions X and Y. Note that as σ deviates from 1, since the scale difference between the two distributions increases, the power of a test is also expected to increase. Figure 5.3 shows the powers of different tests. Here the BD test had the best performance. In Example 5.5 (i), WD, pBF-exp, and pBF-log tests had similar performance, but in Example 5.5 (ii), pBF-exp and pBF-log tests outperformed the WD test. The pBF-ℓ_2 test and the FAD test had a relatively poor performance in this example.

In Example 5.5(ii), the poor performance of pBF-ℓ_2 can be attributed to the fact that the moment conditions required for this measure (as mentioned in Proposition 5.1) are not satisfied. However, for pBF-exp and pBF-log tests, the respective moment conditions were satisfied, and they exhibited excellent performances. Note that the pBF-ℓ_2 test is not robust against heavy-tailed distributions, while the bounded ϕ function used in the pBF-exp test makes it more robust against outliers and contaminating observations.

Next, we consider a couple of examples (Examples 5.6 and 5.7), where the projection-based tests have superior performance compared to the distance-based tests.

Example 5.6. Define $X(t) = \sum_{i=1}^{d} \frac{1}{\sqrt{d}} \xi_i \sqrt{2} \sin(2\pi i t)$ and $Y(t) = \sum_{i=1}^{d} \frac{1}{\sqrt{d}} \eta_i \sqrt{2} \cos(2\pi i t)$, where the ξ_is and the η_is are independent mean zero random variables with the same variance. Here we consider two cases: (i) ξ_i's and η_i's are i.i.d. $\mathcal{N}_1(0,2)$ random variables, (ii) ξ_i's are i.i.d $\mathcal{N}_1(0,2)$, but η_is follow the standard t-distribution with 4 degrees of freedom.

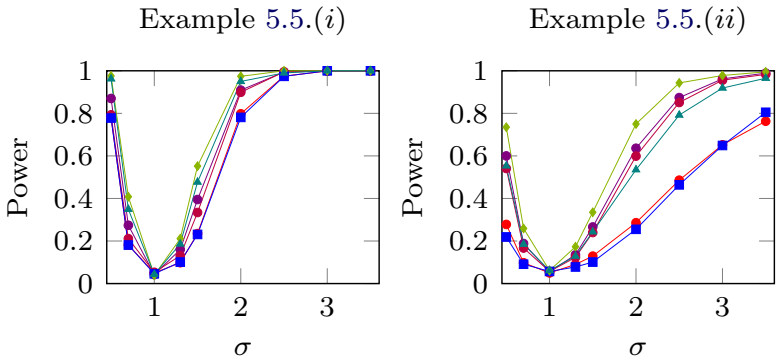


Fig. 5.3 *Powers of pBF-ℓ_2 (●), pBF-exp (●), pBF-log (●), FAD (■), BD (◆) and WD (▲) tests in Examples 5.5 (i) and (ii).*

We generated 30 observations on both X and Y and compute the power of the tests for $d = 3^i$ with $i = 1, \ldots, 4$. Figure 5.4 provides the sample paths of X and Y observations for Examples 5.6 (i) and 5.6 (ii) when $d = 9$. We observe that $X(0.5)$ is identically zero, but the $Y(0.5)$ is random. Therefore, these random functions have a visible distributional difference. In Figure 5.5, we observe that the powers of pBF-ℓ_2, pBF-exp, and pBF-log tests remained one for all choices of d. The FAD test also performed well, but our tests had an edge for large d. However, BD and WD tests had a decaying performance with increasing d.

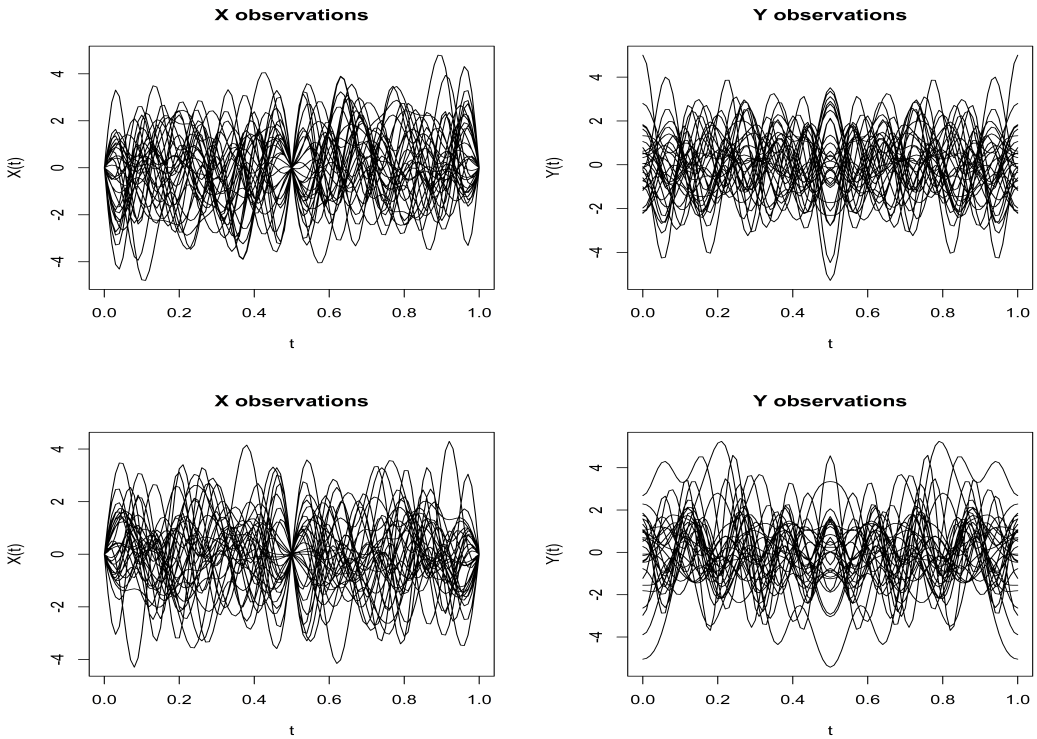


Fig. 5.4 *Sample paths of $X(t)$ and $Y(t)$ for $0 \leq t \leq 1$. The top row corresponds to Example 5.6 (i), and the bottom row corresponds to Example 5.6 (ii).*

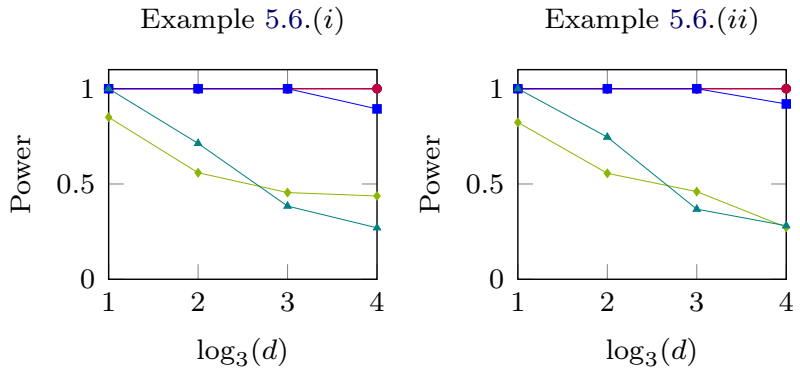


Fig. 5.5 *Powers of pBF-ℓ_2 (●), pBF-exp (●), pBF-log (●), FAD (■), BD (♦) and WD (▲) tests in Examples 5.6 (i) and (ii).*

Note that in this example, the squared pairwise distances are $\|X_1 - Y_1\|^2 = \sum_{i=1}^d \xi_i^2/d + \sum_{i=1}^d \eta_i^2/d$, $\|X_1 - X_2\|^2 = \sum_{i=1}^d (\xi_{1i} - \xi_{2i})^2/d$ and $\|Y_1 - Y_2\|^2 = \sum_{i=1}^d (\eta_{1i} - \eta_{2i})^2/d$ (where the ξ_{ji}'s are associated with the X_j's and the η_{ji}'s are associated with the Y_j's, $j = 1, 2$ and $i = 1, \ldots, d$). One can show that with increasing d, the pairwise distances converge to the same limit. Here the distributions of $\|X_1 - Y_1\|, \|X_1 - X_2\|$ and $\|Y_1 - Y_2\|$ are non-degenerate, and they are expected to be close for large d. On the other hand, we have $\langle X_1, X_2 \rangle = \sum_{i=1}^d \xi_{1i}\xi_{2i}/d$, $\langle Y_1, Y_2 \rangle = \sum_{i=1}^d \eta_{1i}\eta_{2i}/d$, but $\langle X_1, Y_1 \rangle = 0$. The pBF tests successfully discriminate between non-degenerate and degenerate distributions (i.e., the distributions of $\langle X_1, X_2 \rangle$ and $\langle Y_1, X_2 \rangle$ or the distributions of $\langle Y_1, Y_2 \rangle$ and $\langle X_1, Y_2 \rangle$) and aggregate them in a suitable way. As a result, they had superior performance compared to the distance-based methods. From the above discussion, one would expect these tests to exhibit a similar behaviour if $\phi_{0,i}(t) = \sqrt{2}\sin(2\pi it)$ and $\psi_{0,i}(t) = \sqrt{2}\cos(2\pi it)$ are replaced by some other orthonormal sequences $\{\phi_i(t)\}$ and $\{\psi_i(t)\}$, which are orthogonal among themselves.

Example 5.7. Define $U(t) = \sum_{i=1}^{81} \frac{1}{\sqrt{81}} \xi_i \sqrt{2}\sin(2\pi it)$ and $V(t) = \sum_{i=1}^{81} \frac{1}{\sqrt{81}} \eta_i \sqrt{2}\cos(2\pi it)$ for $t \in [0, 1]$, where the ξ_is and the η_is are i.i.d. $\mathcal{N}_1(0, 2)$. We generate independent observations $X_1, \ldots, X_n$ from the distribution of U and $Y_1, \ldots, Y_n$ from the mixture distribution of U and V with mixing proportion $1-\delta/\sqrt{n}$ and $\delta/\sqrt{n}$, respectively. We consider three choices of δ (1,2 and 4) and compute the powers of different tests for different sample sizes as reported in Figure 5.6.

Here, the pBF-ℓ_2 test had an overwhelming performance for all choices of δ. The pBF-log test and the pBF-exp test had the next best performance followed by the FAD test. BD and WD tests had poor performance in all cases.

Next, we generate $X_1, \ldots, X_n$ independently from F, and $Y_1, \ldots, Y_n$ independently from a contiguous alternative $(1 - \delta/\sqrt{n})F + \delta/\sqrt{n}G$ with $\delta > 0$, where G differs from F either in their locations (Example 5.8) or in their scales (Example 5.9). We use these examples to compare the efficiencies of different tests.

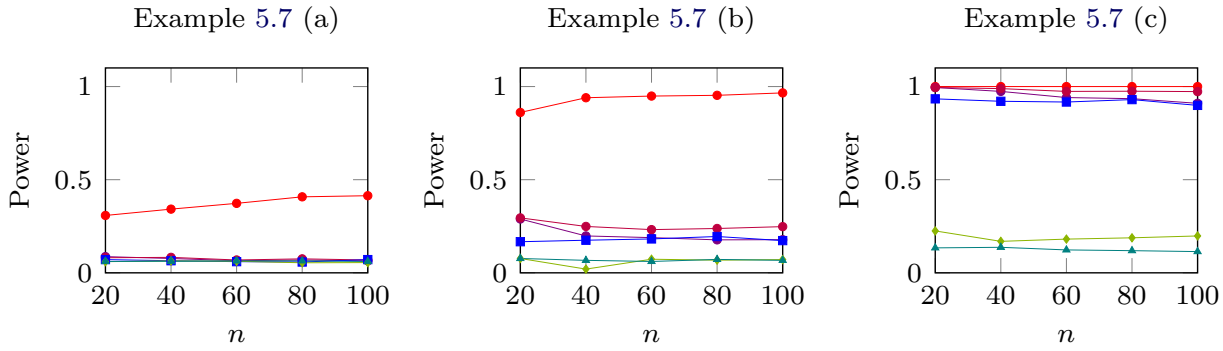


Fig. 5.6 *Powers of pBF-ℓ_2 (●), pBF-exp (●), pBF-log (●), FAD (■), BD (♦) and WD (▲) tests in Example 5.7.*

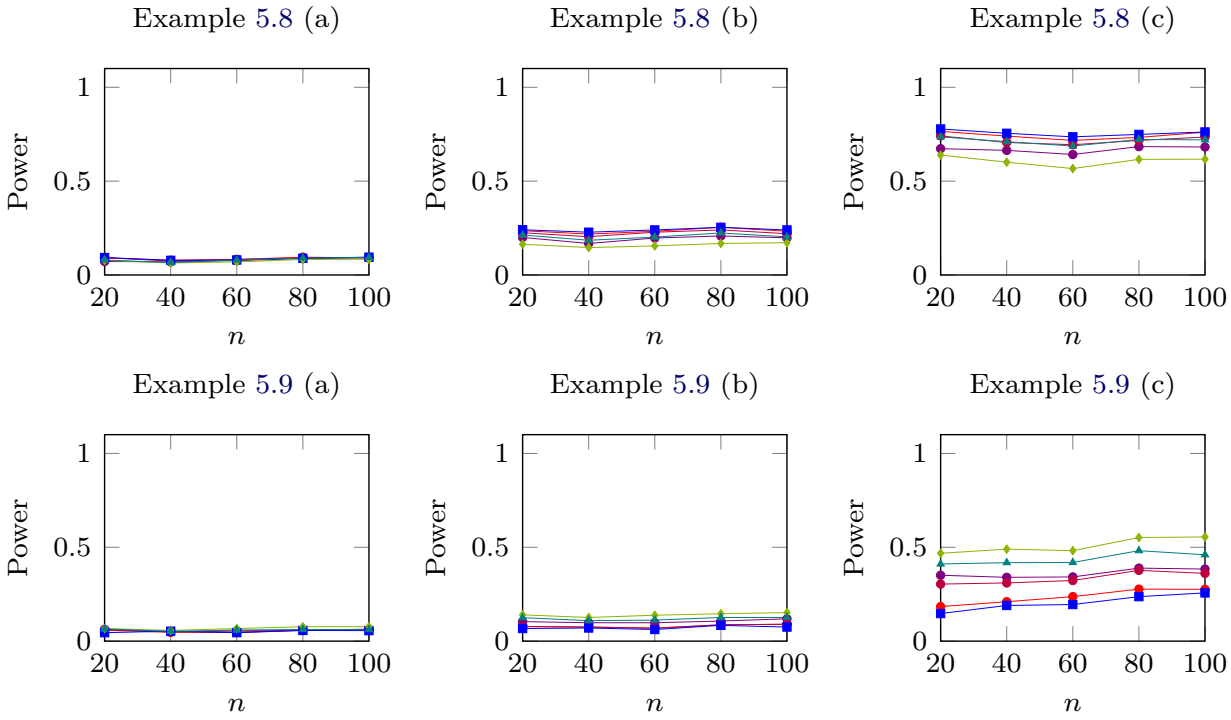


Fig. 5.7 *Powers of pBF-ℓ_2 (●), pBF-exp (●), pBF-log (●), FAD (■), BD (♦) and WD (▲) tests in Examples 5.8 and 5.9.*

Example 5.8. We generate $X_1, \ldots, X_n$ independently from F, the distribution of $\sum_{i=1}^{9} \frac{1}{i^{2.5}} \xi_i \psi_i(t)$, where the ξ_i's are i.i.d. $\mathcal{N}_1(0,1)$ and $\{\psi_i\}$ is the trigonometric basis of $L_2[0,1]$. $Y_1, \ldots, Y_n$ are generated independently from a contiguous alternative $(1-\delta/\sqrt{n})F + \delta/\sqrt{n}G$, where $\delta > 0$ and G is the distribution of $\sum_{i=1}^{9} \frac{1}{i^{2.5}} \eta_i \psi_i(t)$ with the η_i's being i.i.d. $\mathcal{N}_1(1,1)$.

Example 5.9. We generate X and Y observations similar to Example 5.8. Here we consider the ξ_i's to be i.i.d. $\mathcal{N}_1(0,1)$ and the η_i's to be i.i.d. $\mathcal{N}_1(0,2)$ random variables.

Figure 5.7 displays the power of different tests for different values of δ and sample size n. Note that when n is large, the power of a test serves as an approximation of the efficiency of that

test. In Example 5.8, the FAD test turned out to be the most efficient, closely followed by the pBF-ℓ_2 test. The rest of the tests can be arranged as pBF-log, WD, pBF-exp, and BD tests in decreasing order of efficiency. In Example 5.9, the BD test turned out to be the most efficient, followed by WD, pBF-exp, pBF-log, pBF-ℓ_2, and FAD tests arranged in the same order.

Finally, we consider two imbalanced problems (Examples 10 and 11), which are imbalanced versions of Examples 5.4 and 5.5, respectively.

Example 5.10. We generate 50 observations on X from the Wiener process W on $[0, 1]$ and 20 observations on Y, which follows the distribution of $\mu + W$, where (i) $\mu(t) = rt^2$ and (ii) $\mu(t) = re^t$ on $[0, 1]$ as in Example 5.4.

The power of the tests for different values of r is reported in Figure 5.8. In this example, the pBF tests and the FAD test had a competitive performance. In Example 5.10 (i), the FAD test had an edge over the proposed tests, whereas, in Example 5.10 (ii), our tests had an edge over the FAD test. The BD test also exhibited good performance in Examples 5.10 (i) and 5.10 (ii). However, the WD test had a very poor performance.

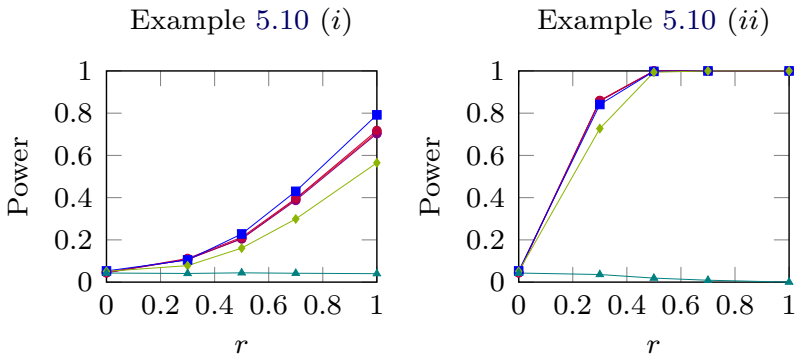


Fig. 5.8 *Powers of pBF-ℓ_2 (●), pBF-exp (●), pBF-log (●), FAD (■), BD (♦) and WD (▲) tests in Examples 5.10 (i) and (ii).*

Example 5.11. We generate observations on X and Y following the model considered in Example 5.5, where X and Y differ only in the scales. But this time we generate 50 observations on X and 20 observations on Y.

Figure 5.9 displays the powers of the tests for this example, which are computed for various choices of the scale parameter σ. In Example 5.11(i), the BD test had the best performance as in Example 5.5. But contrarily, the WD test was highly affected by the imbalanced nature of the problem and exhibited a decaying performance with increasing σ. The pBF tests were competitive with the BD test. The FAD test had a relatively poor performance in this example. In Example 5.11(ii), the BD test, the pBF-exp test, and the pBF-log test exhibited a similar performance, with the pBF tests having an edge over the BD test. Here, pBF-ℓ_2 and FAD tests had a relatively poor performance, which can be explained as in the case of Example 5.5.

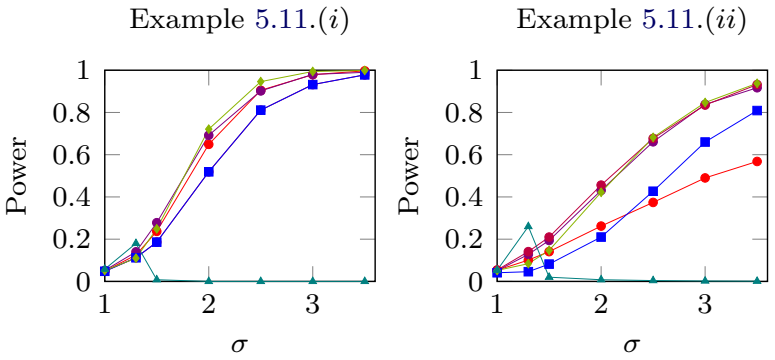


Fig. 5.9 *Powers of pBF-ℓ_2 (●), pBF-exp (●), pBF-log (●), FAD (■), BD (♦) and WD (▲) tests in Examples 5.11 (i) and (ii).*

5.2.2 Analysis of DTI data

For further evaluation of the performance of our test, we analyze the DTI dataset available in the R package 'refund'. The DTI data were collected at Johns Hopkins University and the Kennedy-Krieger Institute. Diffusion tensor imaging (DTI) is a magnetic resonance imaging technology that traces water diffusivity in the brain and helps to create an image of the white matter tract. This dataset has been studied in Goldsmith et al. (2011, 2012) in the context of functional regression. Several measurements of water diffusion are provided by DTI, but here we work with the fractional anisotropy (FA) tract profiles recorded at 93 different locations of the corpus callosum. The dataset contains measurements on 100 'Multiple Sclerosis patients (MS)' and 42 'healthy controls (HC)'. While the number of visits for MS patients ranges between 2 and 8, each healthy person visits

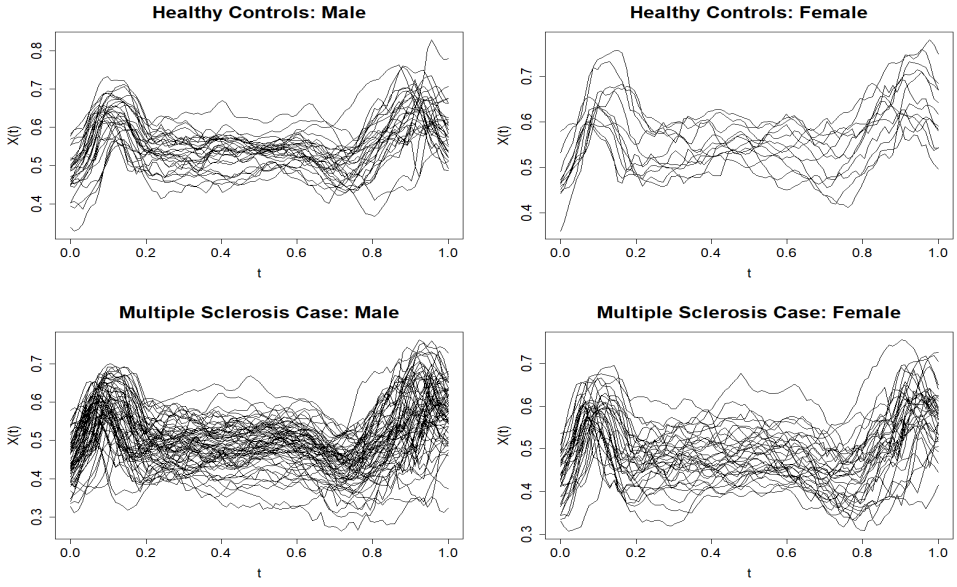


Fig. 5.10 *The fractional anisotropy (FA) tract profiles on the first visit of the patients (divided according to health status and gender) reported in the DTI dataset available in the R package 'refund'.*

just once. So, we consider the FA tract profiles only for the first visit of the subjects. We deleted the subject with ID '2017', which had some missing values, and worked with the remaining 99 MS patients and 42 healthy controls. Among the subjects, there are both males and females. To test whether the 'health status' or the 'gender' of the subject affects the FA tract profile, we divided the dataset into four groups, each corresponding to a particular combination of gender and health status. Figure 5.10 displays the FA tracts of different individuals divided into these four groups.

Taking one pair of groups at a time, we tested for the distributional difference. So, we considered $\binom{4}{2} = 6$ cases: (C1) HC males vs. HC females, (C2) HC males vs. MS males, (C3) HC males vs. MS females, (C4) HC females vs. MS males, (C5) HC females vs. MS females and (C6) MS males vs. MS females. The p-values of pBF-ℓ_2, pBF-exp, pBF-log, BD, WD tests, and Bonferonni corrected p-value of FAD test are reported in Table 5.1. Here, the randomized p-values were computed based on 10,000 random permutations. In many cases, the WD test failed to detect the distributional difference when the others rejected the null hypothesis at 5% level. Our analysis suggests that the distributional difference among the males and females is statistically insignificant, but the distribution of the FA tract profile differs significantly depending on the health status.

Table 5.1 *p-values of pBF-ℓ_2, pB-log, pBF-exp, BD, WD tests, and the Bonferroni corrected p-value of the FAD test for the DTI dataset divided according to health status and gender of the patients ((C1) HC males vs. HC females, (C2) HC males vs. MS males, (C3) HC males vs. MS females, (C4) HC females vs. MS males, (C5) HC females vs. MS females and (C6) MS males vs. MS females).*

Case	pBF-ℓ_2	pBF-exp	pBF-log	BD	WD	FAD
(C1)	0.577	0.366	0.377	0.803	0.369	0.449
(C2)	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	0.096	7×10^{-6}
(C3)	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	8×10^{-6}
(C4)	4×10^{-4}	2×10^{-4}	2×10^{-4}	0.001	0.221	0.002
(C5)	2×10^{-4}	2×10^{-4}	2×10^{-4}	4×10^{-4}	0.542	5×10^{-4}
(C6)	0.574	0.476	0.475	0.638	0.639	0.316

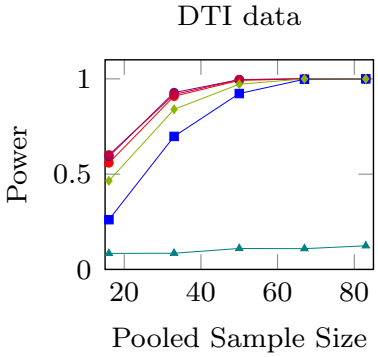


Fig. 5.11 *Powers of pBF-ℓ_2 (●), pBF-exp (●), pBF-log (●), FAD (■), BD (♦) and WD (▲) tests in the DTI data set, where the first sample consists of fractional anisotropy (FA) tract profiles of multiple sclerosis patients and the second sample consisted of those for healthy patients.*

Hence, we can merge the data sets corresponding to males and females and look into the DTI data divided based on health status only. Using this, we compared the performance of the tests by generating random sub-samples, keeping the sample proportions from the two distributions approximately the same as they were in the original data. The sub-sampling procedure was repeated 1000 times to estimate the power of the tests by the proportion of times they rejected H_0. Figure 5.11 shows that for this data set, the pBF tests had comparable performance among themselves, and they had higher powers than all the other tests. BD and FAD tests also had relatively good performance, whereas the WD test performed poorly in this data set.

5.3 Proofs and mathematical details

Lemma A5.1. *Let $\{X_n\}$ be a sequence of i.i.d. random variables from the distribution P defined on the measurable space $(\mathbb{X}, \mathcal{A})$. Let $h : \mathbb{X}^k \to \mathbb{R}$ be a symmetric function such that $\mathbb{E}\, h^2(X_1, \ldots, X_k)$, $\mathbb{E}\, h^2(X_1, \ldots, X_1)$ are finite and $\mathbb{E}\, h(x_1, \ldots, x_{k-1}, X_K) = 0$ almost surely. Then the random variable $\int h(x_1, \ldots, x_k) \prod_{i=1}^k d\hat{G}_P(x_i) \xrightarrow{D} \int h(x_1, \ldots, x_k) \prod_{i=1}^k d\mathbb{G}_P(x_i)$ as $n \to \infty$.*

Proof. Let $f_0 = 1, f_1, f_2, \ldots$ be an orthonormal basis of $L_2(\mathbb{X}, \mathcal{A}, P)$. Since, $h \in L_2(\mathbb{X}^k, \mathcal{A}^k, P^k)$ and is degenerate, we can write $h = \sum_{i_1,\ldots,i_k} \langle h, f_{i_1} \times \cdots \times f_{i_k} \rangle f_{i_1} \times \cdots \times f_{i_k}$, where $i_1, \ldots, i_k \geq 1$. Define, $V_n(f) := \int h(x_1, \ldots, x_k) \prod_{i=1}^k d\hat{G}_P(x_i)$. Then it is easy to see that for any such $h \in L_2(\mathbb{X}^k, \mathcal{A}^k, P^k)$,

$$\mathrm{Var}\big(V_n(h)\big) \leq C \int \tilde{h}^2(x_1, \ldots, x_k) \prod_{i=1}^k dP(x_i), \tag{5.4}$$

where C is a universal constant that depends on k, $V_n(h)$ is linear in h and $\mathbb{E}\big\{V_n(h)\big\} = \mathbb{E}\big\{\tilde{h}(X_1, \ldots, X_1)\big\}$, where

$$\tilde{h}(u_1, \ldots, u_k) = \int h(x_1, \ldots, x_k) \prod_{i=1}^k d(\delta_{u_i} - P)(x_i). \tag{5.5}$$

Since, $\mathbb{E}\, h(x_1, \ldots, x_{k-1}, X_K) = 0$ almost surely, we have $\tilde{h} = h$. Since $\mathbb{E}\, h^2(X_1, \ldots, X_k)$ and $\mathbb{E}\, h^2(X_1, \ldots, X_1)$ are finite, $h_\ell = \sum_{i_1=1}^{\ell} \cdots \sum_{i_k=1}^{\ell} \langle h, f_{i_1} \times \cdots \times f_{i_k} \rangle f_{i_1} \times \cdots \times f_{i_k}$ converges to h in $L_2(P^k)$ as $\ell \to \infty$. Hence, the series $\sum_{i_1,\ldots,i_k \geq 1} \langle h, f_{i_1} \times \cdots \times f_{i_k} \rangle V_n(f_{i_1} \times \cdots \times f_{i_k})$ converges in $L_2(P^n)$ (since $\mathrm{Var}\big(V_n(h - h_\ell)\big)$ and $|\mathbb{E}V_n(h - h_\ell)|$ both converges to zero as ℓ diverges to infinity).

Also note that the finite-dimensional distributions of $\{V_n(f_{i_1} \times \cdots \times f_{i_k}) \mid (i_1, \ldots, i_k) \in \mathbb{N}^k\}$ converges to those of $\{\int f_{i_1}(u_1) \cdots f_{i_k}(u_2) d\mathbb{G}_P(u_1) \cdots d\mathbb{G}_P(u_k) \mid (i_1, \ldots, i_k) \in \mathbb{N}^k\}$. Since the stochastic integrals are taken in the Wiener sense, the limiting class of random variables can also be written as $\{X_{i_1} \cdots X_{i_k} \mid (i_1, \ldots, i_k) \in \mathbb{N}^k\}$. Hence, by continuous mapping theorem, we can say that $\sum_{i_1,\ldots,i_k \geq 1} \langle h_\ell, f_{i_1} \times \cdots \times f_{i_k} \rangle V_n(f_{i_1} \times \cdots \times f_{i_k})$ converges in distribution to $\sum_{i_1,\ldots,i_k \geq 1} \langle h_\ell, f_{i_1} \times \cdots \times f_{i_k} \rangle \prod_{j=1}^k \int f_{i_j}(u) d\mathbb{G}_P(u) = \int h_\ell(x_1, \ldots, x_k) \prod_{i=1}^k d\mathbb{G}_P(x_i)$ as n grows to infinity.

Now note that $\{\int f_i(u)d\mathbb{G}_P(u)\}$ is a sequence of i.i.d. $\mathcal{N}_1(0,1)$ random variables. Define, $[\ell] := \{1, 2, \ldots, \ell\}$. Then for any $\ell_1 < \ell_2$, we have

$$\mathbb{E}\Big|\int h_{\ell_2}(x_1,\ldots,x_k)\prod_{i=1}^k d\mathbb{G}_P(x_i) - \int h_{\ell_2}(x_1,\ldots,x_k)\prod_{i=1}^k d\mathbb{G}_P(x_i)\Big|^2$$

$$= \mathbb{E}\Big|\sum_{(i_1,\ldots,i_k)\in[\ell_2]^k\setminus[\ell_1]^k}\langle h, f_{i_1}\times\cdots\times f_{i_k}\rangle\prod_{j=1}^k\int f_{i_j}(u)d\mathbb{G}_P(u)\Big|^2 \le C_{2k}\sum_{(i_1,\ldots,i_k)\in[\ell_2]^k\setminus[\ell_1]^k}\langle h, f_{i_1}\times\cdots\times f_{i_k}\rangle^2,$$

where C_{2k} is a constant that depends on the even moments of the $\mathcal{N}_1(0,1)$ distribution up to order $2k$. Since $h_\ell \to h$ in $L_2(P^k)$, it is also a Cauchy sequence, and hence for any $\epsilon > 0$, we get an $N \in \mathbb{N}$ such that $\sum_{(i_1,\ldots,i_k)\in[\ell_2]^k\setminus[\ell_1]^k}\langle h, f_{i_1}\times\cdots\times f_{i_k}\rangle^2 < \epsilon/C_{2k}$ for any $\ell_1, \ell_2 \ge N$. Using it, we get

$$\mathbb{E}\Big|\int h_{\ell_2}(x_1,\ldots,x_k)\prod_{i=1}^k d\mathbb{G}_P(x_i) - \int h_{\ell_2}(x_1,\ldots,x_k)\prod_{i=1}^k d\mathbb{G}_P(x_i)\Big|^2 < \epsilon.$$

Hence, the sequence $\int h_\ell(x_1,\ldots,x_k)\prod_{i=1}^k d\mathbb{G}_P(x_i)$ is Cauchy in L_2. It is easy to see that the limit of this sequence of random variables is $\int h(x_1,\ldots,x_k)\prod_{i=1}^k d\mathbb{G}_P(x_i)$.

Let $\varphi_{n\ell}(t)$ and $\varphi_n(t)$ be the characteristic functions of $\int h_\ell(x_1,\ldots,x_k)\prod_{i=1}^k d\hat{G}_P(x_i)$ and $\int h(x_1,\ldots,x_k)\prod_{i=1}^k d\hat{G}_P(x_i)$, respectively. Also, define $\varphi_\ell(t)$ and $\varphi(t)$ as the characteristic functions of $\int h_\ell(x_1,\ldots,x_k)\prod_{i=1}^k d\mathbb{G}_P(x_i)$ and $\int h(x_1,\ldots,x_k)\prod_{i=1}^k d\mathbb{G}_P(x_i)$, respectively. Then by the previous arguments, for any $\epsilon > 0$ and any $t \in \mathbb{R}$, we can find $K \in \mathbb{N}$ such that $|\varphi_\ell(t) - \varphi(t)| < \epsilon$ for all $\ell \ge K$. So, it is easy to see that for all $t \in \mathbb{R}$, and for all $\ell \ge K$,

$$\begin{aligned}
&\lim_{n\to\infty}|\varphi_n(t) - \varphi(t)| \\
&\le \lim_{n\to\infty}|\varphi_{n\ell}(t) - \varphi_n(t)| + \lim_{n\to\infty}|\varphi_{n\ell}(t) - \varphi_\ell(t)| + |\varphi_\ell(t) - \varphi(t)| \\
&= \lim_{n\to\infty}|\varphi_{n\ell}(t) - \varphi_n(t)| + \epsilon \\
&\le |t|\lim_{n\to\infty}\left\{\mathbb{E}\Big\{\Big|\int h_\ell(x_1,\ldots,x_k)\prod_{i=1}^k d\hat{G}_P(x_i) - \int h(x_1,\ldots,x_k)\prod_{i=1}^k d\hat{G}_P(x_i)\Big|^2\Big\}\right\}^{1/2} + \epsilon \\
&= |t|\Big\{\mathbb{E}\big\{(h-h_\ell)^2(X_1,\ldots,X_k)\big\} + \big\{\mathbb{E}\{(h-h_\ell)(X_1,\ldots,X_1)\}\big\}^2\Big\}^{1/2} + \epsilon.
\end{aligned}$$

For the second last inequality, we have used the fact that $\mathbb{E}\{|e^{itX} - 1|\} \le |t|\mathbb{E}|X| \le |t|\{\mathbb{E}|X|^2\}^{1/2}$. The last equality follows from the fact $\mathbb{E}(X^2) = \mathrm{Var}(X) + \big(\mathbb{E}(X)\big)^2$ and equation (5.4). Now as ℓ grows to infinity, we have $\lim_{n\to\infty}|\varphi_n(t) - \varphi(t)| < \epsilon$ for any arbitrary $\epsilon > 0$. Hence, we can say $\lim_{n\to\infty}|\varphi_n(t) - \varphi(t)| = 0$. So, $\int h(x_1,\ldots,x_k)\prod_{i=1}^k d\hat{G}_P(x_i) \xrightarrow{D} \int h(x_1,\ldots,x_k)\prod_{i=1}^k d\mathbb{G}_P(x_i)$. ■

Lemma A5.2. *Let $\{X_n\}$ and $\{Y_m\}$ be two independent sequences of i.i.d. random variables from distributions P and Q, respectively, defined on the measurable space $(\mathbb{X}, \mathcal{A})$. Let $h : \mathbb{X}^2 \to \mathbb{R}$ be a measurable function (not necessarily symmetric) such that $\mathbb{E}\ h^2(X_1, Y_1) < \infty$. Assume that $\mathbb{E}\ h(x_1, Y_1) = 0$ and $\mathbb{E}\ h(X_1, y_1) = 0$ almost surely. Then the random variable $\sqrt{nm}\int h(x_1, y_1)d\hat{G}_P(x_1)d\hat{G}_Q(y_1) \xrightarrow{D} \int h(x_1, y_1)d\mathbb{G}_P(x_1)d\mathbb{G}_Q(y_1)$ as $\min\{n, m\}$ goes to infinity.*

Proof. Let $f_0 = 1, f_1, f_2, \ldots$ be an orthonormal basis of $L_2(\mathbb{X}, \mathcal{A}, P)$ and $g_0 = 1, g_1, g_2, \ldots$ be an orthonormal basis of $L_2(\mathbb{X}, \mathcal{A}, Q)$. Define, $\mathcal{F} = \{f_0, f_1, f_2, \ldots\} \cup \{g_0, g_1, g_2, \ldots\}$. Using a similar argument as in Lemma A5.1, we get the result. However, a brief sketch of the proof is given below.

Let us define, $T_n(f) := \sqrt{nm} \int h(x_1, y_1) d\hat{G}_P(x_1) d\hat{G}_Q(y_1)$. For $g, h \in L_2(\mathbb{X}^2, \mathcal{A}^2, P \otimes Q)$, $\mathrm{Cov}(T_n(g), T_n(h)) = \int \tilde{g}(u, v)\tilde{h}(u, v) dP(u) dQ(v)$, and $\mathbb{E}\{T_n(h)\} = \mathbb{E}\{\tilde{h}(X_1, Y_1)\} = 0$, where $\tilde{h}, \tilde{g}$ are as in (5.5). Now write h as a series $\sum_{k_1,k_2} \langle h, f_{k_1} \times g_{k_2} \rangle f_{k_1} \times g_{k_2}$ where $k_1, k_2 \geq 1$ (under the given assumptions). Since, $h_\ell = \sum_{(k_1,k_2) \in [\ell]^2} \langle h, f_{k_1} \times g_{k_2} \rangle f_{k_1} \times g_{k_2}$ converges to h in $L_2(P \otimes Q)$, $T_n(h_\ell)$ converges to $T_n(h)$ in $L_2(P^n \otimes Q^m)$ (since $\mathrm{Var}\big(h - h_\ell\big) \to 0$ as $\ell \to \infty$).

Clearly, the finite-dimensional distributions of the process $\{T_n(f_{k_1} \times g_{k_2}) \mid (k_1, k_2) \in \mathbb{N}^2\}$ converges to those of $\{\int f_{k_1}(u) g_{k_2}(v) d\mathbb{G}_P(u) d\mathbb{G}_Q(v) \mid (k_1, k_2) \in \mathbb{N}^2\}$. Then by continuous mapping theorem, as $\min\{n, m\} \to \infty$, $\sum_{k_1,k_2} \langle h_\ell, f_{k_1} \times g_{k_2} \rangle T_n(f_{k_1} \times g_{k_2}) \xrightarrow{D} \sum_{k_1,k_2} \langle h_\ell, f_{k_1} \times g_{k_2} \rangle \int f_{k_1}(u) g_{k_2}(v) d\mathbb{G}_P(u) d\mathbb{G}_Q(v)$, which can also be written as $\int h_\ell(x_1, x_2) d\mathbb{G}_P(x_1) d\mathbb{G}_Q(x_2)$.

Now note that $\big\{\int f_i(u) d\mathbb{G}_P(u)\big\}$ and $\big\{\int g_i(u) d\mathbb{G}_P(u)\big\}$ are two independent sequence of $\mathcal{N}_1(0, 1)$ random variables. So, for any $\ell_1 < \ell_2$, we have

$$\begin{aligned}
&\mathbb{E}\Big| \int h_{\ell_1}(x_1, x_2) d\mathbb{G}_P(x_1) d\mathbb{G}_Q(x_2) - \int h_{\ell_2}(x_1, x_2) d\mathbb{G}_P(x_1) d\mathbb{G}_Q(x_2) \Big|^2 \\
&= \mathbb{E}\Big| \sum_{(i,j) \in [\ell_2]^2 \setminus [\ell_1]^2} \langle h, f_i \times g_j \rangle \int f_i(u) d\mathbb{G}_P(u) \int g_j(u) d\mathbb{G}_Q(u) \Big|^2 \leq \sum_{(i,j) \in [\ell_2]^2 \setminus [\ell_1]^2} \big| \langle h, f_i \times f_j \rangle \big|^2
\end{aligned}$$

The rest of the proof follows by using similar arguments as in the proof of Lemma A5.1. ■

Remark A5.1. We can also establish distributional convergence of the variable $\int h(x_1, \ldots, x_k) \prod_{i \in S} d\hat{G}_P(x_i) \prod_{i \in S^c} d\hat{G}_Q(x_i)$ for any $S \subset \{1, \ldots, k\}$, where h satisfies conditions similar to Lemma A5.1 and A5.2. However, the proof is similar to the above, so we omit it here.

Proof of Theorem 5.1. Let $X \sim F$, and $Y \sim G$ and define $\varphi : \mathcal{H} \to \mathbb{C}$ as $\varphi(f) = \mathbb{E}\{e^{i\langle X, f \rangle}\} - \mathbb{E}\{e^{i\langle Y, f \rangle}\}$. Notice that the function φ is continuous, and $\varphi(f) = 0 \ \forall f \in \mathcal{H}$ implies $F = G$. One can see that if $F = G$, $T(F^f, G^f) = 0$ for any $f \in \mathcal{H}$, and hence $\eta^{\nu'}(F, G) = 0$ for any probability distribution ν' on $\mathcal{H}$. Now $\eta^{\nu'}(F, G) = 0$ implies that there exists a Borel measurable set $E \in \mathcal{B}(\mathcal{H})$ such that $\nu'(E) = 1$ and $T(F^f, G^f) = 0 \ \forall f \in E$. Hence, by the assumption on T, we have $\varphi(f) = 0 \ \forall f \in E$. Thus when $\mathrm{supp}\{\nu'\}$ contains the surface of the unit sphere, for any $f \in \mathcal{H}$, $\langle X, f \rangle = \|f\| \langle X, f/\|f\| \rangle$ and $\langle Y, f \rangle = \|f\| \langle Y, f/\|f\| \rangle$ have the same distribution. Hence, we have $\varphi(f) = 0 \ \forall f \in \mathcal{H}$. ■

Proof of Theorem 5.2. Let us first consider the following claim.

Claim: Suppose that X is a random variable on a Hilbert space $\mathcal{H}$ with distribution F. If $\mathrm{supp}\{F\} \subset \mathcal{H}_0$, where $\mathcal{H}_0$ is a closed subspace of $\mathcal{H}$, then X and QX have the same distribution, for $Q : \mathcal{H} \to \mathcal{H}_0$ being the projection operator onto $\mathcal{H}_0$.

The claim can be proved by showing that the characteristic functions of X and QX are identical. Now take any arbitrary $\theta \in \mathcal{H}$ and note that

$$\begin{aligned}\mathbb{E}\{e^{i\langle X,\theta\rangle}\} &= \int_{\mathcal{H}} e^{i\langle x,\theta\rangle}\, dF(x) = \int_{\mathcal{H}_0} e^{i\langle x,\theta\rangle}\, dF(x) = \int_{\mathcal{H}_0} e^{i\langle Qx,Q\theta\rangle + i\langle (I-Q)x,(I-Q)\theta\rangle}\, dF(x) \\ &= \int_{\mathcal{H}_0} e^{i\langle Qx,Q\theta\rangle + i\langle 0,(I-Q)\theta\rangle}\, dF(x) = \int_{\mathcal{H}_0} e^{i\langle Qx,\theta\rangle}\, dF(x) = \mathbb{E}\{e^{i\langle QX,\theta\rangle}\}.\end{aligned}$$

Hence, the claim holds. Denote $\mathcal{H}_0 = \overline{span\{\text{supp}\{F\} \cup \text{supp}\{G\}\}}$, a closed subspace of $\mathcal{H}$. Let Q denote the projection operator from $\mathcal{H}$ to $\mathcal{H}_0$. Then

$$\begin{aligned}\varphi(f) &= \mathbb{E}\{e^{i\langle X,f\rangle}\} - \mathbb{E}\{e^{i\langle Y,f\rangle}\} = \int_{\mathcal{H}} e^{i\langle x,f\rangle} dF(x) - \int_{\mathcal{H}} e^{i\langle y,f\rangle} dG(y) \\ &= \int_{\mathcal{H}_0} e^{i\langle x,f\rangle} dF(x) - \int_{\mathcal{H}_0} e^{i\langle y,f\rangle} dG(y) \\ &= \int_{\mathcal{H}_0} e^{i\langle Qx,Qf\rangle + i\langle (I-Q)x,(I-Q)f\rangle} dF(x) - \int_{\mathcal{H}_0} e^{i\langle Qy,Qf\rangle + i\langle (I-Q)y,(I-Q)f\rangle} dG(y) \\ &= \int_{\mathcal{H}_0} e^{i\langle Qx,Qf\rangle} dF(x) - \int_{\mathcal{H}_0} e^{i\langle Qy,Qf\rangle} dG(y) \\ &= \int_{\mathcal{H}_0} e^{i\langle x,Qf\rangle} dF(x) - \int_{\mathcal{H}_0} e^{i\langle y,Qf\rangle} dG(y) \qquad \text{(Using Claim with } \theta = Qf) \\ &= \varphi(Qf). \end{aligned} \tag{5.6}$$

This shows that it is enough to consider the value of $\varphi(\cdot)$ on $\mathcal{H}_0$. Now if $\eta^{(F+G)/2}(F,G) = 0$, we have $\varphi(f) = 0$ for all $f \in \text{supp}\{F\} \cup \text{supp}\{G\}$. So, $\langle X, f\rangle$ and $\langle Y, f\rangle$ are independent and identically distributed for any fixed $f \in \text{supp}\{F\} \cup \text{supp}\{G\}$. Now for any $g \in \mathcal{H}_0$, we can write $g = \sum_{i=1}^{\infty} a_i f_i$ for $\{f_i\} \subset \text{supp}\{F\} \cup \text{supp}\{G\}$ and $\{a_i\} \subset \mathbb{R}$. Clearly, $\langle X, g\rangle = \sum_{i=1}^{\infty} a_i \langle X, f_i\rangle$ and $\langle Y, g\rangle = \sum_{i=1}^{\infty} a_i \langle Y, f_i\rangle$ are also independent and identically distributed. This implies $\varphi(g) = 0\ \forall g \in \mathcal{H}_0$, and using equation (5.6), we have $\varphi(g) = \varphi(Qg) = 0\ \forall g \in \mathcal{H}$. So, if $\eta^{(F+G)/2}(F,G) = 0$, X and Y are identically distributed. ■

Proof of Proposition 5.1. (a) It is easy to see that

$$\begin{aligned}\mathbb{E}\{g(X_1,X_2,X_3;Y_1,Y_2,Y_3)\} &= \frac{1}{2}\int T_\phi(F^f,G^f)dF(f) + \frac{1}{2}\int T_\phi(F^f,G^f)dG(f) \\ &= \frac{1}{2}\int T_\phi(F^f,G^f)d(F+G)(f) = \eta_\phi(F,G).\end{aligned}$$

(b) The proof of this statement follows from Theorem 5.2.

(c) Note that if $U : \mathcal{H} \to \mathcal{G}$ is a unitary operator (i.e., U is a linear bijective map with $\langle Ux, Uy\rangle = \langle x, y\rangle$), then $g(\cdot)$ remains invariant over this operation, i.e., $g(UX_1, UX_2, UX_3; UY_1, UY_2, UY_3) = g(X_1, X_2, X_3; Y_1, Y_2, Y_3)$. Thus.

$$\begin{aligned}\eta_\phi(F,G) = \mathbb{E}\{g(X_1,X_2,X_3;Y_1,Y_2,Y_3)\} &= \mathbb{E}\{g(UX_1,UX_2,UX_3;UY_1,UY_2,UY_3)\} \\ &= \eta_\phi(F\circ U^{-1}, G\circ U^{-1}).\end{aligned}$$

(d) This follows by simply applying the Dominated Convergence Theorem. ■

Proof of Theorem 5.3. First let us note that $\eta_\phi(F,G) = \mathbb{E}\{g^*(X_1,X_2,X_3;Y_1,Y_2,Y_3)\}$ and define the function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ as $\varphi(t,s) = \eta_\phi(F+t\sqrt{n}(\hat{F}_n-F), G+s\sqrt{m}(\hat{G}_m-G))$, where $\hat{F}_n$ and $\hat{G}_m$ are the empirical probability distributions based on the observed data $\mathcal{X}$ and $\mathcal{Y}$, respectively. Clearly, φ is a bivariate polynomial with random coefficients. It is easy to see that the coefficients are tight (by Lemma A5.1 and A5.2). Also note that $\hat{\eta}^\phi_{n,m} = \varphi(1/\sqrt{n}, 1/\sqrt{m})$ and $\varphi(0,0) = \eta_\phi(F,G)$. Hence the limiting distribution of $\hat{\eta}^\phi_{n,m}$ is determined by the leading non-zero coefficient of $\varphi(1/\sqrt{n}, 1/\sqrt{m})$. Now, for any $n, m \in \mathbb{N}$,

$$\varphi(1/\sqrt{n}, 1/\sqrt{m}) = \varphi(0,0) + \frac{1}{\sqrt{n}}\frac{\partial}{\partial t}\varphi(t,s)\Big|_{(0,0)} + \frac{1}{\sqrt{m}}\frac{\partial}{\partial s}\varphi(t,s)\Big|_{(0,0)} + r_{n,m},$$

where by Lemma A5.1 and A5.2, $r_{n,m}$ is of stochastic order $O_P\big(\{1/\sqrt{n}+1/\sqrt{m}\}^2\big)$. Note that

$$\frac{\partial}{\partial t}\varphi(t,s)\Big|_{(0,0)} = \sqrt{n}\int h_1^*(x)d(\hat{F}_n - F)(x), \;\; \frac{\partial}{\partial s}\varphi(t,s)\Big|_{(0,0)} = \sqrt{m}\int h_2^*(y)d(\hat{G}_m - G)(y),$$

where $h_1^*(x) = 3\ \mathbb{E}\{g^*(x,X_2,X_3;Y_1,Y_2,Y_3)\}$ and $h_2^*(y) = 3\ \mathbb{E}\{g^*(X_1,X_2,X_3;y,Y_2,Y_3)\}$. Later, we shall find these functions up to additive constant terms. From the above discussions, we have

$$\begin{aligned}&\hat{\eta}_{n,m} - \eta(F,G)\\ &= \frac{1}{\sqrt{n}}\int h_1^*(x)\sqrt{n}d(\hat{F}_n - F)(x) + \frac{1}{\sqrt{m}}\int h_2^*(y)\sqrt{m}d(\hat{G}_m - G)(y) + O_P\Big(\big\{\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\big\}^2\Big).\end{aligned}$$

Since, $\mathcal{X}$ and $\mathcal{Y}$ are independently generated, the corresponding empirical processes $\sqrt{n}(\hat{F}_n - F)$ and $\sqrt{m}(\hat{G}_m - G)$ are also independent. Now if atleast one of h_1^* and h_2^* is a non-zero function, applying CLT, we have

$$\sqrt{nm/(n+m)}\big\{\hat{\eta}^\phi_{n,m} - \eta_\phi(F,G)\big\} \xrightarrow{D} \mathcal{N}_1\Big(0, \lambda\mathrm{Var}\big(h_1^*(X_1)\big) + (1-\lambda)\mathrm{Var}\big(h_2^*(Y_1)\big)\Big).$$

If h_1^* and h_2^* are both zero functions, the limiting distribution of $\hat{\eta}^\phi_{n,m}$ is determined by the coefficients of higher order terms in $\varphi(t,s)$. Generally, in such situations, one may have

$$\eta^\phi_{n,m} - \eta_\phi(F,G) = \frac{1}{2n}\frac{\partial^2}{\partial t^2}\varphi(t,s)\Big|_{(0,0)} + \frac{1}{2m}\frac{\partial^2}{\partial s^2}\varphi(t,s)\Big|_{(0,0)} + \frac{1}{\sqrt{nm}}\frac{\partial^2}{\partial t\partial s}\varphi(t,s)\Big|_{(0,0)} + r'_{n,m}, \quad (5.7)$$

where

$$\begin{aligned}\frac{\partial^2}{\partial t^2}\varphi(t,s)\Big|_{(0,0)} &= n\int h_1^*(u,v)d(\hat{F}_n - F)(u)d(\hat{F}_n - F)(v),\\ \frac{\partial^2}{\partial t\partial s}\varphi(t,s)\Big|_{(0,0)} &= \sqrt{nm}\int h_2^*(u,v)d(\hat{F}_n - F)(u)d(\hat{G}_m - F)(v),\\ \frac{\partial^2}{\partial s^2}\varphi(t,s)\Big|_{(0,0)} &= m\int h_3^*(u,v)d(\hat{G}_m - F)(u)d(\hat{G}_m - F)(v),\end{aligned}$$

for $h_1^*(u,v), h_2^*(u,v)$ and $h_3^*(u,v)$ being the second order projections of the symmetrized core function g^* and $r'_{n,m} = O_P\big(\{1/\sqrt{n}+1/\sqrt{m}\}^3\big)$ by Lemma A5.1 and A5.2. Here the limiting distribution of $nm/(n+m)\big\{\hat{\eta}^\phi_{n,m} - \eta_\phi(F,G)\big\}$ depends on the limiting distribution of the random vector $(\frac{\partial^2}{\partial t^2}\varphi(t,s)\Big|_{(0,0)}, \frac{\partial^2}{\partial s^2}\varphi(t,s)\Big|_{(0,0)}, \frac{\partial^2}{\partial t\partial s}\varphi(t,s)\Big|_{(0,0)})$. Now, we prove our result below.

First, we find the functions $h_1^*(x)$ and $h_2^*(y)$ up to additive constant terms. Note that adding constant terms to the functions does not change the limiting distribution of the empirical stochastic integrals. So,

$$\begin{aligned} h_1^*(x) &= 3\ \mathbb{E}\big\{g^*(x, X_2, X_3; Y_1, Y_2, Y_3)\big\} \\ &= \Big\{\mathbb{E}\big(g(x, X_2, X_3; Y_1, Y_2, Y_3) + g(X_1, x, X_3; Y_1, Y_2, Y_3) + g(X_1, X_2, x; Y_1, Y_2, Y_3)\big)\Big\}, \end{aligned}$$

where $g(x_1, x_2, x_3; y_1, y_2, y_3)$ is the function defined in Proposition 5.1 (a). Now,

$$\begin{aligned} \mathbb{E}\big(g(x, X_2, X_3; Y_1, Y_2, Y_3)\big) &= \frac{1}{2}\mathbb{E}\Big\{ -\phi\big(|\langle X_2, x\rangle - \langle X_3, x\rangle|^2\big) - \phi\big(|\langle Y_2, x\rangle - \langle Y_3, x\rangle|^2\big) \\ &\qquad + 2\ \phi\big(|\langle X_2, x\rangle - \langle Y_1, x\rangle|^2\big) + 2\ \phi\big(|\langle x, Y_1\rangle - \langle Y_2, Y_1\rangle|^2\big)\Big\} + c_1, \\ \mathbb{E}\big(g(X_1, x, X_3; Y_1, Y_2, Y_3)\big) &= \frac{1}{2}\mathbb{E}\Big\{ -\phi\big(|\langle x, X_1\rangle - \langle X_3, X_1\rangle|^2\big) + 2\ \phi\big(|\langle x, X_1\rangle, \langle Y_1, X_1\rangle|^2\big) \\ &\qquad - \phi\big(|\langle x, Y_1\rangle - \langle X_3, Y_1\rangle|^2\big)\Big\} + c_2, \\ \mathbb{E}\big(g(X_1, X_2, x; Y_1, Y_2, Y_3)\big) &= \frac{1}{2}\mathbb{E}\Big\{ -\phi\big(|\langle X_2, X_1\rangle - \langle x, X_1\rangle|^2\big) - \phi\big(|\langle X_2, Y_1\rangle - \langle x, Y_1\rangle|^2\big)\Big\} + c_3, \end{aligned}$$

where c_1, c_2 and c_3 are constants that depend the distributions F and G. Clearly,

$$\begin{aligned} h_1^*(x)\ =& \Big[\mathbb{E}\Big\{ -\phi\big(|\langle x, X_1\rangle - \langle X_3, X_1\rangle|^2\big) - \phi\big(|\langle X_2, Y_1\rangle - \langle x, Y_1\rangle|^2\big) \\ &+ \phi\big(|\langle x, X_1\rangle - \langle Y_1, X_1\rangle|^2\big) + \phi\big(|\langle x, Y_1\rangle - \langle Y_2, Y_1\rangle|^2\big)\Big\}\Big] + \frac{1}{2}T_\phi(F^x, G^x) + c. \end{aligned}$$

Under H_0, h_1^* is the zero function, and under H_1, it is non-degenerate. We can conclude the same for $h_2(y)$ due to the structural symmetry of g. Hence,

$$\sqrt{nm/(n+m)}\big(\hat{\eta}_{n,m} - \eta(F, G)\big) \xrightarrow{D} \sqrt{1-\lambda}\int h_1^*(x)d\mathbb{G}_F(x) + \sqrt{\lambda}\int h_2^*(y)d\mathbb{G}_G(y)$$

as $\min\{n, m\} \to \infty$ and $n/(n+m) \to \lambda \in [0, 1]$, where $\mathbb{G}_F$ and $\mathbb{G}_G$ are two independent Brownian Bridge processes. The random variable $\sqrt{1-\lambda}\int h_1^*(x)d\mathbb{G}_F(x) + \sqrt{\lambda}\int h_2^*(y)d\mathbb{G}_G(y)$ is a normal random variable with zero mean and variance $(1-\lambda)\mathrm{Var}\big(h_1^*(X)\big) + \lambda\mathrm{Var}\big(h_2^*(Y)\big)$. This completes the proof of part (a).

To prove part (b), we find the functions $h_1^*(u, v), h_2^*(u, v)$ and $h_3^*(u, v)$ up to additive constant terms. Since, $(g \tilde{+} c)(u, v) = g(u, v) + c - \mathbb{E}h(u, V) - c - \mathbb{E}h(U, v) - c + \mathbb{E}h(U, V) + c = \tilde{g}$ (as defined in Lemma A5.1) for any c, here also the additive constants do not effect the limiting distribution of the empirical stochastic integrals. Note that

$$\begin{aligned} h_1^*(u, v) &= \Big\{\mathbb{E}g\big(u, v, X_3; Y_1, Y_2, Y_3\big) + \mathbb{E}g\big(u, X_2, v; Y_1, Y_2, Y_3\big) \\ &\quad + \mathbb{E}g\big(X_1, u, v; Y_1, Y_2, Y_3\big) + \mathbb{E}g\big(v, u, X_3; Y_1, Y_2, Y_3\big) \\ &\quad + \mathbb{E}g\big(v, X_2, u; Y_1, Y_2, Y_3\big) + \mathbb{E}g\big(X_1, u, v; Y_1, Y_2, Y_3\big)\Big\} \\ &=: \Big\{T_1(u, v) + T_2(u, v)\Big\}, \end{aligned}$$

where $T_1(u, v) = \mathbb{E}g\big(u, v, X_3; Y_1, Y_2, Y_3\big) + \mathbb{E}g\big(u, X_2, v; Y_1, Y_2, Y_3\big) + \mathbb{E}g\big(X_1, u, v; Y_1, Y_2, Y_3\big)$ and $T_2(u, v) = T_1(v, u)$. Hence, it is enough to find $T_1(u, v)$. Now, note that

$$
\begin{aligned}
&\mathbb{E}\big(g(u,v,X_3;Y_1,Y_2,Y_3)\big)\\
&= \frac{1}{2}\mathbb{E}\Big\{ -\phi\big(|\langle v,u\rangle - \langle X_3,u\rangle|^2\big) - \phi\big(|\langle Y_2,u\rangle - \langle Y_3,u\rangle|^2\big) + 2\,\phi\big(|\langle v,u\rangle - \langle Y_1,u\rangle|^2\big)\\
&\qquad - \phi\big(|\langle v,Y_1\rangle - \langle X_3,Y_1\rangle|^2\big) - \phi\big(|\langle Y_2,Y_1\rangle - \langle Y_3,Y_1\rangle|^2\big) + 2\,\phi\big(|\langle u,Y_1\rangle - \langle Y_2,Y_1\rangle|^2\big)\Big\},\\
&\mathbb{E}\big(g(u,X_2,v;Y_1,Y_2,Y_3)\big)\\
&= \frac{1}{2}\mathbb{E}\Big\{ -\phi\big(|\langle X_2,u\rangle - \langle v,u\rangle|^2\big) - \phi\big(|\langle Y_2,u\rangle - \langle Y_3,u\rangle|^2\big) + 2\,\phi\big(|\langle X_2,u\rangle - \langle Y_1,u\rangle|^2\big)\\
&\qquad - \phi\big(|\langle X_2,Y_1\rangle - \langle v,Y_1\rangle|^2\big) - \phi\big(|\langle Y_2,Y_1\rangle - \langle Y_3,Y_1\rangle|^2\big) + 2\,\phi\big(|\langle u,Y_1\rangle - \langle Y_2,Y_1\rangle|^2\big)\Big\},\\
&\mathbb{E}\big(g(X_1,u,v;Y_1,Y_2,Y_3)\big)\\
&= \frac{1}{2}\mathbb{E}\Big\{ -\phi\big(|\langle u,X_1\rangle - \langle v,X_1\rangle|^2\big) - \phi\big(|\langle Y_2,X_1\rangle - \langle Y_3,X_1\rangle|^2\big) + 2\,\phi\big(|\langle u,X_1\rangle - \langle Y_1,X_1\rangle|^2\big)\\
&\qquad - \phi\big(|\langle u,Y_1\rangle - \langle v,Y_1\rangle|^2\big) - \phi\big(|\langle Y_2,Y_1\rangle - \langle Y_3,Y_1\rangle|^2\big) + 2\,\phi\big(|\langle X_1,Y_1\rangle - \langle Y_2,Y_1\rangle|^2\big)\Big\}.
\end{aligned}
$$

Therefore, under H_0,

$$
\begin{aligned}
T_1(u,v) \;=& \mathbb{E}\Big\{ -\phi\big(|\langle v,Y_1\rangle - \langle X_3,Y_1\rangle|^2\big)\Big\} + 3\mathbb{E}\Big\{\phi\big(|\langle u,X_1\rangle - \langle Y_1,X_1\rangle|^2\big)\Big\}\\
&- \mathbb{E}\Big\{\phi\big(|\langle u,Y_1\rangle - \langle v,Y_1\rangle|^2\big)\Big\} - b_1,
\end{aligned}
$$

where $b_1 = \mathbb{E}\Big\{\phi\big(|\langle Y_2,Y_1\rangle - \langle Y_3,Y_1\rangle|^2\big)\Big\}$. Similarly, we get

$$
\begin{aligned}
T_2(u,v) \;=& \mathbb{E}\Big\{ -\phi\big(|\langle u,Y_1\rangle - \langle X_3,Y_1\rangle|^2\big)\Big\} + 3\mathbb{E}\Big\{\phi\big(|\langle v,X_1\rangle - \langle Y_1,X_1\rangle|^2\big)\Big\}\\
&- \mathbb{E}\Big\{\phi\big(|\langle v,Y_1\rangle - \langle u,Y_1\rangle|^2\big)\Big\} - b_1,
\end{aligned}
$$

for the same constant b_1. Hence, under H_0,

$$
\begin{aligned}
h_1^*(u,v) =& -2\mathbb{E}\Big\{\phi\big(|\langle u,Y_1\rangle - \langle v,Y_1\rangle|^2\big)\Big\} + 2\mathbb{E}\Big\{\phi\big(|\langle u,Y_1\rangle - \langle Y_2,Y_1\rangle|^2\big)\Big\}\\
&+ 2\mathbb{E}\Big\{\phi\big(|\langle v,Y_1\rangle - \langle Y_2,Y_1\rangle|^2\big)\Big\} - 2b_1.
\end{aligned}
$$

Clearly, this is symmetric and non-zero. By the structural symmetry of g, we also have $h_3^*(u,v) = h_1^*(u,v)$. Also, note that

$$
\begin{aligned}
h_2^*(u,v) \;=& \Big\{\mathbb{E}g\big(u,X_2,X_3;v,Y_2,Y_3\big) + \mathbb{E}g\big(u,X_2,X_3;Y_1,v,Y_3\big) + \mathbb{E}g\big(u,X_2,X_3;Y_1,Y_2,v\big)\\
&+ \mathbb{E}g\big(X_1,u,X_3;v,Y_2,Y_3\big) + \mathbb{E}g\big(X_1,u,X_3;Y_1,v,Y_3\big) + \mathbb{E}g\big(X_1,u,X_3;Y_1,Y_2,v\big)\\
&+ \mathbb{E}g\big(X_1,X_2,u;v,Y_2,Y_3\big) + \mathbb{E}g\big(X_1,X_2,u;Y_1,v,Y_3\big) + \mathbb{E}g\big(X_1,X_2,u;Y_1,Y_2,v\big)\Big\}.
\end{aligned}
$$

Under H_0, this can be simplified to

$$
\begin{aligned}
h_2^*(u,v) =& -2\mathbb{E}\Big\{\phi\big(|\langle u,Y_1\rangle - \langle Y_2,Y_1\rangle|^2\big)\Big\} - 2\mathbb{E}\Big\{\phi\big(|\langle v,Y_1\rangle - \langle Y_2,Y_1\rangle|^2\big)\Big\}\\
&+ 2\;\mathbb{E}\Big\{\phi\big(|\langle u,Y_1\rangle - \langle v,Y_1\rangle|^2\big)\Big\} + 2b_1.
\end{aligned}
$$

Under H_0, we have $h_1^*(u,v) = h_3^*(u,v) = 2\;h(u,v)$ (say), and hence $h_2^*(u,v) = -2\;h(u,v)$. Clearly, all of them are non-zero functions. Thus, the limiting distribution of $\hat{\eta}_{n,m}^{\phi}$ is determined by

the joint asymptotic distribution of the vector $\big(\int h(u,v)d\hat{G}_F(u)d\hat{G}_F(v), \int h(u,v)d\hat{G}_F(u)d\hat{G}'_F(v), \int h(u,v)d\hat{G}'_F(u)d\hat{G}'_F(v)\big)$ where $\hat{G}_F = \sqrt{n}(\hat{F}_n - F)$ and $\hat{G}'_F = \sqrt{m}(\hat{G}_m - F)$. For joint convergence of this vector, we look at the distributional convergence of

$$t_1 \int h(u,v)d\hat{G}_F(u)d\hat{G}_F(v) + t_2 \int h(u,v)d\hat{G}_F(u)d\hat{G}'_F(v) + t_3 \int h(u,v)d\hat{G}'_F(u)d\hat{G}'_F(v)$$

for some real numbers t_1, t_2 and t_3. We first write h as a series with respect to an orthonormal basis $f_0 = 1, f_1, f_2, \ldots$ of $L_2(F)$ as

$$h(u,v) = \sum_{(k_1,k_2)\in\mathbb{N}^2} \langle h, f_{k_1} \times f_{k_2}\rangle f_{k_1} \times f_{k_2}$$

and truncate the series at ℓ to get

$$h_\ell(u,v) = \sum_{(k_1,k_2)\in[\ell]^2} \langle h, f_{k_1} \times f_{k_2}\rangle f_{k_1} \times f_{k_2}.$$

By continuous mapping theorem, the variable $t_1 \int h_\ell(u,v)d\hat{G}_F(u)d\hat{G}_F(v) + t_2 \int h_\ell(u,v) d\hat{G}_F(u)d\hat{G}'_F(v) + t_3 \int h_\ell(u,v)d\hat{G}'_F(u)d\hat{G}'_F(v)$ converges in distribution to $t_1 \int h_\ell(u,v)d\mathbb{G}_F(u)d\mathbb{G}_F(v) + t_2 \int h_\ell(u,v)d\mathbb{G}_F(u)d\mathbb{G}'_F(v) + t_3 \int h_\ell(u,v)d\mathbb{G}'_F(u)d\mathbb{G}'_F(v)$. Also, take $\ell_2 > \ell_1 > N$ such that $\|h_{\ell_1} - h_{\ell_2}\|_{L_2(F^2)} < \epsilon$. Then, using triangle inequality, we have

$$\begin{aligned}
&\Big\{\mathbb{E}\Big|t_1 \int h_{\ell_1}(u,v)d\mathbb{G}_F(u)d\mathbb{G}_F(v) + t_2 \int h_{\ell_1}(u,v)d\mathbb{G}_F(u)d\mathbb{G}'_F(v) + t_3 \int h_{\ell_1}(u,v)d\mathbb{G}'_F(u)d\mathbb{G}'_F(v) \\
&- t_1 \int h_{\ell_2}(u,v)d\mathbb{G}_F(u)d\mathbb{G}_F(v) - t_2 \int h_{\ell_2}(u,v)d\mathbb{G}_F(u)d\mathbb{G}'_F(v) - t_3 \int h_{\ell_2}(u,v)d\mathbb{G}'_F(u)d\mathbb{G}'_F(v)\Big|^2\Big\}^{1/2} \\
&\le |t_1| \ \Big\{\mathbb{E}\Big|\int h_{\ell_1}(u,v)d\mathbb{G}_F(u)d\mathbb{G}_F(v) - \int h_{\ell_2}(u,v)d\mathbb{G}_F(u)d\mathbb{G}_F(v)\Big|^2\Big\}^{1/2} \\
&\quad + |t_2| \ \Big\{\mathbb{E}\Big|\int h_{\ell_1}(u,v)d\mathbb{G}_F(u)d\mathbb{G}'_F(v) - \int h_{\ell_2}(u,v)d\mathbb{G}_F(u)d\mathbb{G}'_F(v)\Big|^2\Big\}^{1/2} \\
&\quad + |t_3| \ \Big\{\mathbb{E}\Big|\int h_{\ell_1}(u,v)d\mathbb{G}'_F(u)d\mathbb{G}'_F(v) - \int h_{\ell_2}(u,v)d\mathbb{G}'_F(u)d\mathbb{G}'_F(v)\Big|^2\Big\}^{1/2}.
\end{aligned}$$

Now, using Lemma A5.1 and Lemma A5.2, we get the distributional convergence of the above random variable to

$$t_1 \int h(u,v)d\mathbb{G}_F(u)d\mathbb{G}_F(v) + t_2 \int h(u,v)d\mathbb{G}_F(u)d\mathbb{G}'_F(v) + t_3 \int h(u,v)d\mathbb{G}'_F(u)d\mathbb{G}'_F(v).$$

Hence, under H_0, using equation (5.7) and continuous mapping theorem, $nm/(n+m)\eta^{\phi}_{n,m}$ converges in distribution to $(1-\lambda)\int h(u,v)d\mathbb{G}_F(u)d\mathbb{G}_F(v) - 2\sqrt{\lambda(1-\lambda)}\int h(u,v)d\mathbb{G}_F(u)d\mathbb{G}'_F(v) + \lambda\int h(u,v)d\mathbb{G}'_F(u)d\mathbb{G}'_F(v)$ as $\min\{n,m\} \to \infty$ and $n/(n+m) \to \lambda \in [0,1]$, where $\mathbb{G}_F$ and $\mathbb{G}'_F$ are two independent F-Brownian Bridge processes. Since $h(\cdot,\cdot)$ is symmetric, using Fredholm theory of integral equations, we have $h(u,v) = \sum_{i=1}^{\infty}\lambda_i\varphi_i(u)\varphi_i(v)$, where $\{\lambda_i\}$ and $\{\varphi_i\}$ are the sequences of eigenvalues and eigenfunctions of the integral equation $\int h(u,v)\gamma(v)dF(v) = \lambda\gamma(u)$. Here, the equality holds in the L_2 sense. Then the limiting distribution of $nm/(n+m)\hat{\eta}^{\phi}_{n,m}$ is same as that of $\sum_{i=1}^{\infty}\lambda_i\big(\sqrt{1-\lambda}\int\varphi_i(u)d\mathbb{G}_F(u) - \sqrt{\lambda}\int\varphi_i(u)d\mathbb{G}'_F(u)\big)^2$ (since the stochastic integrals are in Wiener

sense). This is identically distributed as $\sum_{i=1}^{\infty} \lambda_i Z_i^2$ where $\{Z_i\}$ is a sequence of i.i.d. $\mathcal{N}_1(0,1)$ random variables. This completes the proof. ■

Proof of Corollary 5.1. The proof follows from Theorem 5.3. ■

Proof of Theorem 5.4. Let $\mathcal{U} = \{U_1 = X_1, \ldots, U_n = X_n, U_{n+1} = Y_1, \ldots, U_N = Y_m\}$ ($N = n + m$) be the pooled data, and π be a random permutation of $\{1, 2, \ldots, N\}$ independent of the observed data. Define the two-sample permutation empirical measures and the pooled measure as

$$\tilde{P}_{n,N} = \frac{1}{n}\sum_{i=1}^{n} \delta_{U_{\pi(i)}}, \qquad \tilde{Q}_{m,N} = \frac{1}{m}\sum_{i=n+1}^{N} \delta_{U_{\pi(i)}} \qquad \text{and} \qquad H_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{U_i}, \quad \text{respectively.}$$

Then it follows from Theorem 3.7.1 in Van der Vaart & Wellner (2013) that over a suitable class of functions $\mathcal{F}$, $\sqrt{n}(\tilde{P}_{n,N} - H_N)$ converges in distribution to $\sqrt{1-\lambda}\mathbb{G}_H$, where $H = \lambda F + (1-\lambda)G$ and $\lim n/N = \lambda$. Since $\sqrt{m}(\tilde{Q}_{m,N} - H_N) = -\sqrt{n/m}\sqrt{n}(\tilde{P}_{n,N} - H_N)$, the distributional convergence of $\sqrt{m}(\tilde{Q}_{m,N} - H_N)$ follows trivially.

Now, using arguments as in Theorem 5.3 (b), for any fixed alternative, we have the distributional convergence of $nm/(n+m)\hat{\eta}^{\phi,\pi}_{n,m}$ to $(1-\lambda)^2 \int h(u,v) d\mathbb{G}_H(u) d\mathbb{G}_H(v) + 2\lambda(1-\lambda) \int h(u,v) d\mathbb{G}_H(u) d\mathbb{G}_H(v) + \lambda^2 \int h(u,v) d\mathbb{G}_H(u) d\mathbb{G}_H(v) = \int h(u,v) d\mathbb{G}_H(u) d\mathbb{G}_H(v)$ as $\min\{n,m\} \to \infty$ and $n/N \to \lambda$. Writing $h(u,v) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(u)\varphi_i(v)$, where the λ_i's and the φ_i's are solutions of the integral equation $\int h(u,v) f(v) dH(v) = \lambda f(u)$, we can show that this limiting distribution is same as the distribution of $\sum_{k=1}^{\infty} \lambda_k Z_k^2$ where $\{Z_k\}$ is a sequence of i.i.d. $\mathcal{N}_1(0,1)$ random variables. This completes the proof. ■

Proof of Proposition 5.2. To prove this, let us first define,

$$M(t) = \frac{1}{N!}\Big\{ \sum_{\pi \in \mathcal{S}_N} \mathrm{I}[\hat{\eta}^{\phi,\pi}_{n,m} \le t] \Big\}, \qquad M_B(t) = \frac{1}{B}\Big\{ \sum_{i=1}^{B} \mathrm{I}[\hat{\eta}^{\phi,\pi_i}_{n,m} \le t] \Big\}.$$

Here M and M_B are distribution functions conditioned on the observed pooled data $\mathcal{U}$. Then

$$\begin{aligned}
|p_{n,m} - p_{n,m,B}| &= \Big| \frac{1}{N!}\Big\{ \sum_{\pi \in \mathcal{S}_N} \mathrm{I}[\hat{\eta}^{\phi,\pi}_{n,m} \ge \hat{\eta}_{n,m}] \Big\} - \frac{1}{B+1}\Big\{ \sum_{i=1}^{B} \mathrm{I}[\hat{\eta}^{\phi,\pi_i}_{n,m} \ge \hat{\eta}_{n,m}] + 1 \Big\} \Big| \\
&= \Big| \frac{1}{N!}\Big\{ \sum_{\pi \in \mathcal{S}_N} \mathrm{I}[\hat{\eta}^{\phi,\pi}_{n,m} < \hat{\eta}_{n,m}] \Big\} - \frac{1}{B+1}\Big\{ \sum_{i=1}^{B} \mathrm{I}[\hat{\eta}^{\phi,\pi_i}_{n,m} < \hat{\eta}_{n,m}] \Big\} \Big| \\
&= |M(\hat{\eta}_{n,m}-) - \frac{B}{B+1} M_B(\hat{\eta}_{n,m}-)| \\
&\le |M(\hat{\eta}_{n,m}-) - M_B(\hat{\eta}_{n,m}-)| + |\frac{M_B(\hat{\eta}_{n,m}-)}{B+1}| \le \sup_{t \in \mathbb{R}} |M(t) - M_B(t)| + \frac{1}{B+1}.
\end{aligned}$$

Conditioned on the pooled data $\mathcal{U}$, the Dvoretzky-Keifer-Wolfwitz inequality (Massart (1990)) gives $\mathbb{P}\{\sup_{t \in \mathbb{R}} |M(t) - M_B(t)| > \epsilon\} \le 2e^{-2B\epsilon^2}$. Hence, conditioned on $\mathcal{U}$, as B grows to infinity, the randomized p-value $p_{n,m,B}$ converges almost surely to $p_{n,m}$. ■

Proof of Theorem 5.5. Under Assumption (A5.1), it is easy to see that $F^{(N)} = (1-\delta/\sqrt{N})F + \delta/\sqrt{N}L$ has the Radon-Nikodym derivative as $\Big(1+\frac{\delta}{\sqrt{N}}\big(\ell(z)-1\big)\Big)$ w.r.t. F. Hence, if $Z_1,\ldots,Z_N \overset{iid}{\sim} F$, then the log-likelihood ratio is given by

$$L_N = \log\Big\{\prod_{i=1}^{N}\frac{dF^{(N)}}{dF}(Z_i)\Big\} = \sum_{i=1}^{N}\log\Big\{\frac{dF^{(N)}}{dF}(Z_i)\Big\} = \sum_{i=1}^{N}\log\Big(1+\frac{\delta}{\sqrt{N}}\big(\ell(Z_i)-1\big)\Big).$$

Using the fact $\log(1+y) = y - \frac{y^2}{2} + \frac{1}{2}y^2\beta(y)$, where $\lim_{y\to 0}\beta(y) = 0$ and $\beta(y)$ is continuous, we get

$$L_N = \sum_{i=1}^{N}\frac{\delta}{\sqrt{N}}\big(\ell(Z_i)-1\big) - \sum_{i=1}^{N}\frac{\delta^2}{2N}\big(\ell(Z_i)-1\big)^2 + \sum_{i=1}^{N}\frac{\delta^2}{2N}\big(\ell(Z_i)-1\big)^2\beta\Big(\frac{\delta}{\sqrt{N}}\big(\ell(Z_i)-1\big)\Big).$$

Under Assumption (A5.1), as N grows to infinity,

$$\sum_{i=1}^{N}\frac{\delta^2}{N}\big(\ell(Z_i)-1\big)^2 \overset{a.s.}{\to} \delta^2\mathbb{E}\Big(\big(\ell(Z_1)-1\big)^2\Big).$$

Hence, we only need to show that

$$\sum_{i=1}^{N}\frac{\delta^2}{N}\big(\ell(Z_i)-1\big)^2\beta\Big(\frac{\delta}{\sqrt{N}}\big(\ell(Z_i)-1\big)\Big)$$

converges to zero in probability. Notice that

$$\sum_{i=1}^{N}\frac{\delta^2}{N}\big(\ell(Z_i)-1\big)^2\beta\Big(\frac{\delta}{\sqrt{N}}\big(\ell(Z_i)-1\big)\Big) \le \max_{1\le i\le N}|\beta\Big(\frac{\delta}{\sqrt{N}}\big(\ell(Z_i)-1\big)\Big)|\sum_{i=1}^{N}\frac{\delta^2}{N}\big(\ell(Z_i)-1\big)^2.$$

Due to Assumption (A5.1), it suffices to show that $\max_{1\le i\le N}|\beta\Big(\frac{\delta}{\sqrt{N}}\big(\ell(Z_i)-1\big)\Big)| \overset{P}{\to} 0$, which follows if $\max_{1\le i\le N}|\frac{\delta}{\sqrt{N}}\big(\ell(Z_i)-1\big)| \overset{P}{\to} 0$ (as $\lim_{y\to 0}\beta(y) = 0$ and it is continuous). Note that

$$\begin{aligned}
\mathbb{P}\Big\{\max_{1\le i\le N}|\frac{1}{\sqrt{N}}\big(\ell(Z_i)-1\big)| > \epsilon\Big\} &\le \sum_{i=1}^{N}\mathbb{P}\Big\{|\frac{1}{\sqrt{N}}\big(\ell(Z_i)-1\big)| > \epsilon\Big\} \\
&= N\mathbb{P}\Big\{|\frac{1}{\sqrt{N}}\big(\ell(Z_1)-1\big)| > \epsilon\Big\} \\
&\le \frac{1}{\epsilon^2}\mathbb{E}\Big\{\big(\ell(Z_1)-1\big)^2 I\big(|\frac{1}{\sqrt{N}}\big(\ell(Z_1)-1\big)| > \epsilon\big)\Big\}.
\end{aligned}$$

By the Dominated Convergence Theorem, the right-hand side converges to zero. Hence, we have

$$\left|\log\Big\{\prod_{i=1}^{N}\frac{dF^{(N)}}{dF}(Z_i)\Big\} - \frac{\delta}{\sqrt{N}}\sum_{i=1}^{N}\Big(\ell(Z_i)-1\Big) + \frac{\delta^2}{2}\mathbb{E}\Big\{\ell(Z_1)-1\Big\}^2\right| \to 0$$

in probability under F as N goes to infinity. ■

Lemma A5.3. *Under Assumption (A5.1) and $\delta_N = \delta/\sqrt{N}$, the process $\{\sqrt{n}\int f(u)\big(\hat{F}_n - F\big)(u) \mid f\in\mathcal{F}\}$ converges in distribution to the process $\{\int \tilde{f}d\big(\mathbb{G}_F + \delta(L-F)\big) \mid \tilde{f}(u) = f(u) - \mathbb{E}_F f(Z), f\in\mathcal{F}\}$ in $\ell^\infty(\mathcal{F})$, where $\mathcal{F}$ is a Donsker class of measurable functions with $\sup_{f\in\mathcal{F}}|Pf| < \infty$, and $\mathbb{G}_F$ is the F-Brownian Bridge process on $\ell^\infty(\mathcal{F})$.*

Proof. Here, we only show the finite-dimensional distribution convergence of the empirical process. The tightness of the process under $F^{(N)}$ can be obtained using Theorem 30.12 from Van der Vaart & Wellner (2013). Let $Z_1, Z_2, \ldots, Z_N$ be a sequence of i.i.d. random variables and $F^{(N)} = (1 - \delta/\sqrt{N})F + \delta/\sqrt{N}L$, where L satisfies Assumption (A5.1). Let $\hat{F}_N$ be the corresponding empirical probability distribution. Take any $f \in \mathcal{F}$ with $\mathbb{E}_F f^2(Z) < \infty$ and note that the joint limiting distribution of $\int f(u)\sqrt{N}d(\hat{F}_N - F)(u)$ and $\log\big\{\prod_{i=1}^N dF^{(N)}/dF(Z_i)\big\}$ is same as that of $\int f(u)\sqrt{N}(\hat{F}_N - F)(u)$ and $\frac{\delta}{\sqrt{N}}\sum_{i=1}^N \big(\ell(Z_i) - 1\big) - \frac{\delta^2}{2}\mathbb{E}\big\{\ell(Z_1) - 1\big\}^2$ (by Theorem 5.5 and Slutsky's theorem), which is $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ -\frac{\delta^2}{2}\mathbb{E}\big\{\ell(Z_1) - 1\big\}^2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \int \tilde{f}^2(u)dF(u) & \tau \\ \tau & \delta^2\mathbb{E}\big\{\ell(Z_1) - 1\big\}^2 \end{pmatrix},$$

$\tau = \delta\big(\int \tilde{f}(u)dL(u)\big)$ and $\tilde{f}(u) = f(u) - \mathbb{E}\ f(Z)$. According to Le Cam's third lemma, this implies that under $F^{(N)}$, $\int f(u)\sqrt{N}d(\hat{F}_N - F)(u) \xrightarrow{D} \mathcal{N}_1\Big(\delta\big(\int \tilde{f}(u)dL(u)\big), \int \tilde{f}^2(u)dF(u)\Big)$. Using Cramer-Wold device, one can further show that under $F^{(N)}$, the finite dimensional distributions of $\big\{\int f(u)\sqrt{N}(\hat{F}_N - F)(u) \mid f \in \mathcal{F}\big\}$ converges to that of $\big\{\int f(u)d\big(\mathbb{G}_F + \delta(L - F)\big)(u) \mid f \in \mathcal{F}\big\}$. Hence, under the contiguous alternative $F^{(N)}$, the empirical process converges in distribution to the process $\big\{\int f(u)d\big(\mathbb{G}_F + \delta(L - F)\big)(u) \mid f \in \mathcal{F}\big\}$, where $\mathcal{F}$ is a F- Donsker class of functions. ■

Proof of Theorem 5.6. Using Lemma A5.3, one can show that under $(F^{(n)}, G^{(m)})$, the empirical processes $\sqrt{n}(\hat{F}_n - F)$ and $\sqrt{m}(\hat{G}_m - F)$ converge in distribution to the processes $\mathbb{G}_F$ and $\mathbb{B}'_F = \mathbb{G}'_F + \delta(L - F)$ respectively, where $\mathbb{G}_F$ and $\mathbb{G}'_F$ are independent F-Brownian Bridge processes. Now, if $f_0 = 1, f_1, f_2, \ldots$ are orthonormal basis of $L_2(F)$, then it is easy to see $\int f_i(u)d\mathbb{G}_F(u) + \delta\int f_i(u)dL(u)$ has a normal distribution with mean $\delta\int f_i(u)dL(u)$ and variance one. Also, we have

$$\Big|\int f_i(u)dL(u)\Big| = \Big|\int f_i(u)\ell(u)dF(u)\Big| \le \Big(\int \ell^2(u)dF(u)\Big)^{1/2},$$

and $\max_i \mathrm{Var}\big(\int f_i(u)d\mathbb{B}'_F(u)\big) = 1$. Now, as in Theorem 5.3 (b), using equation (5.7), the decomposition $h(u, v) = \sum_{i=1}^\infty \lambda_i\varphi_i(u)\varphi_i(v)$ of the second order Hoeffding projection of the core function g^* defined in equation (5.3) and the contiguity of $(F^{(n)}, G^{(m)})$, $nm/(n+m)\hat{\eta}^{\phi}_{n,m}$ converges in distribution to $\sum_{i=1}^\infty \lambda_i\big(\sqrt{1-\lambda}Z_i - \sqrt{\lambda}(Z'_i + \delta\int \varphi_i(u)dL(u))\big)^2$ (as $\int \varphi_k(u)dF(u) = 0,\ \forall\ k$) under $(F^{(n)}, G^{(m)})$. Here $\lambda = \lim n/(n+m)$ and $\{Z_i\}$, $\{Z'_i\}$ are two sequences of independent $\mathcal{N}_1(0, 1)$ random variables. It is easy to show that this limiting distribution is the same as the distribution of $\sum_{i=1}^\infty \lambda_i\big(Z_i - \sqrt{\lambda}\delta\int \varphi_i(u)dL(u)\big)^2$. This completes the proof. ■

Proof of Theorem 5.7. Under the contiguous alternative $(F^{(n)}, G^{(m)})$ (as in Theorem 5.6), for any f with finite $\int f^2(u)dF(u)$, we need to find the joint limiting distribution of $(\frac{1}{n}\sum_{i=1}^n f(U_{\pi(i)}) - \frac{1}{N}\sum_{i=1}^N f(U_i))$ and $\frac{\delta}{N}\sum_{i=1}^N \big(\ell(U_i) - 1\big) - \frac{\delta^2}{2}\mathbb{E}\big\{\ell(U_1) - 1\big\}^2$ assuming $U_i \overset{iid}{\sim} F$.

Note that

$$\begin{pmatrix} \left(\frac{1}{n}\sum_{i=1}^{n} f(U_{\pi(i)}) - \frac{1}{N}\sum_{i=1}^{N} f(U_i)\right) \\ \frac{\delta}{N}\sum_{i=1}^{N}\left(\ell(U_i)-1\right) \end{pmatrix} \overset{D}{=} \begin{pmatrix} \left(\frac{1}{n}\sum_{i=1}^{n} f(U_i) - \frac{1}{N}\sum_{i=1}^{N} f(U_i)\right) \\ \frac{\delta}{N}\sum_{i=1}^{N}\left(\ell(U_i)-1\right) \end{pmatrix}$$

$$= \mathbf{M}_1\mathbf{M}_2 \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}\left(f(U_i)-\mathbb{E}f(U_1)\right) \\ \frac{1}{m}\sum_{i=n+1}^{N}\left(f(U_i)-\mathbb{E}f(U_1)\right) \\ \frac{\delta}{N}\sum_{i=1}^{n}\left(\ell(U_i)-1\right) \\ \frac{\delta}{N}\sum_{i=n+1}^{N}\left(\ell(U_i)-1\right) \end{pmatrix}, \text{ where } \mathbf{M}_1 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{M}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{n}{N} & \frac{m}{N} & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\overset{D}{=} \mathbf{M}_1\mathbf{M}_2 \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}\left(f(U_i)-\mathbb{E}f(U_1)\right) \\ \frac{1}{m}\sum_{i=1}^{m}\left(f(U_i')-\mathbb{E}f(U_1)\right) \\ \frac{\delta}{N}\sum_{i=1}^{n}\left(\ell(U_i)-1\right) \\ \frac{\delta}{N}\sum_{i=1}^{m}\left(\ell(U_i')-1\right) \end{pmatrix} =: \mathbf{M}_1\mathbf{M}_2 \begin{pmatrix} W_{n1} \\ W_{n2} \\ W_{n3} \\ W_{n4} \end{pmatrix},$$

where $U_1', U_1', \ldots, U_m' \overset{iid}{\sim} F$ independent of $U_1, U_2, \ldots, U_n$. Applying multivariate CLT, we get $\sqrt{n}(W_{n1}, W_{n2}, W_{n3}, W_{n4}) \xrightarrow{D} \mathcal{N}_4(\mathbf{0}_4, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \begin{pmatrix} \int \tilde{f}^2(u)dF(u) & 0 & \lambda\delta\int \tilde{f}(u)dL(u) & 0 \\ 0 & \frac{\lambda}{(1-\lambda)}\int \tilde{f}^2(u)dF(u) & 0 & \lambda\delta\int \tilde{f}(u)dL(u) \\ \lambda\delta\int \tilde{f}(u)dL(u) & 0 & \delta^2\lambda^2\int(\ell(u)-1)^2dF(u) & 0 \\ 0 & \lambda\delta\int \tilde{f}(u)dL(u) & 0 & \delta^2\lambda(1-\lambda)\int(\ell(u)-1)^2dF(u) \end{pmatrix},$$

and $\tilde{f}(u) = f(u) - \mathbb{E}_F f(U)$. Then by continuous mapping theorem, we get $\sqrt{n}\Big(\left(\frac{1}{n}\sum_{i=1}^{n} f(U_{\pi(i)}) - \frac{1}{N}\sum_{i=1}^{N} f(U_i)\right), \frac{\delta}{N}\sum_{i=1}^{N}\left(\ell(U_i)-1\right)\Big) \xrightarrow{D} \mathcal{N}_2(\mathbf{0}_2, \mathbf{\Sigma}_0)$, where

$$\mathbf{\Sigma}_0 = \begin{pmatrix} (1-\lambda)\int \tilde{f}^2(u)dF(u) & 0 \\ 0 & \lambda\delta^2\int\left(\ell(u)-1\right)^2dF(u) \end{pmatrix}.$$

Now, using Le Cam's third lemma and Cramer-Wold device, we can say that the finite-dimensional distributions of the process $\sqrt{n}(\tilde{P}_{n,N} - H_N)$ converge to those of $\sqrt{1-\lambda}\mathbb{G}_F$, where $\tilde{P}_{n,N}$ is the permutation empirical measure (as defined in the proof of Theorem 5.4) when $U_1, \ldots, U_N \overset{iid}{\sim} F$. The tightness of the process follows from Theorem 30.12 of Van der Vaart & Wellner (2013). Now applying arguments as in Theorem 5.3 (b), under $(F^{(n)}, G^{(m)})$, we get $nm/(n+m)\hat{\eta}_{n,m}^{\phi,\pi} \xrightarrow{D} \sum_{k=1}^{\infty}\lambda_k Z_k^2$ for some square integrable sequence $\{\lambda_i\}$ (same as in Theorem 5.3 (b)) and a sequence of independent $\mathcal{N}_1(0,1)$ variables $\{Z_i\}$. This completes the proof. ■

Chapter 6

Test of Independence for Functional Data

Like Chapter 5, here also, random functions are modeled as elements of infinite-dimensional function spaces. Let $X^{(i)}$ $(i = 1, 2, \ldots, d)$ be a random function on a separable Hilbert space $\mathcal{H}_i$ with a probability distribution P_i and $\mathbb{P}$ be the joint distribution of $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})$ defined on $\mathcal{H} = \bigoplus_{i=1}^{d} \mathcal{H}_i$, (the space $\mathcal{H}_1 \times \ldots \times \mathcal{H}_d$ endowed with the inner-product $\langle ., . \rangle = \langle ., . \rangle_1 + \cdots + \langle ., . \rangle_d$, for $\langle ., . \rangle_i$ being the inner-product associated with $\mathcal{H}_i$). In this chapter, we develop a test for mutual independence of these $X^{(i)}$'s, i.e., we test the null hypothesis as $H_0 : \mathbb{P} = \mathrm{P}_1 \otimes \cdots \otimes \mathrm{P}_d$ against the alternative hypothesis $H_1 : \mathbb{P} \neq \mathrm{P}_1 \otimes \cdots \otimes \mathrm{P}_d$.

First, we propose a dependency measure that is motivated by the Cramer-Wold device. We know that $X^{(1)}, \ldots, X^{(d)}$ are independent if and only if $\langle X^{(1)}, f_1 \rangle, \ldots, \langle X^{(d)}, f_d \rangle$ are independent for all $\boldsymbol{f} = (f_1, \ldots, f_d) \in \bigoplus_{i=1}^{d} \mathcal{H}_i$. So, one can consider a measure of dependence among $\langle X^{(1)}, f_1 \rangle, \langle X^{(2)}, f_2 \rangle, \ldots, \langle X^{(d)}, f_d \rangle$ and aggregate it over all possible choices of $\boldsymbol{f}$. For any metric space $\mathbb{X}$, let $\mathcal{B}(\mathbb{X})$ be the Borel σ-algebra endowed on $\mathbb{X}$ and $\mathcal{M}(\mathbb{X})$ be the space of all probability measures on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. Now, consider a statistical functional $T : \mathcal{M}(\mathbb{R}^d) \to \mathbb{R}$ such that for any $G \in \mathcal{M}(\mathbb{R}^d)$, $T(G)$ measures the dependence among the components of the corresponding random vector. Therefore, for any fixed $\boldsymbol{f} \in \mathcal{H}$, if $\mathbb{P}^{\boldsymbol{f}}$ denotes the joint probability distribution of $\langle X^{(1)}, f_1 \rangle, \langle X^{(2)}, f_2 \rangle, \ldots, \langle X^{(d)}, f_d \rangle$, $T(\mathbb{P}^{\boldsymbol{f}})$ measures the dependence between $\langle X^{(1)}, f_1 \rangle, \langle X^{(2)}, f_2 \rangle, \ldots, \langle X^{(d)}, f_d \rangle$. We can integrate it with respect to a suitable probability measure $\nu' \in \mathcal{M}(\bigoplus_{i=1}^{d} \mathcal{H}_i)$ to define

$$\xi^{\nu'}(\mathbb{P}) = \int_{\bigoplus_{i=1}^{d} \mathcal{H}_i} T(\mathbb{P}^f)\, d\nu'(f).$$

There are several choices of T for measuring the dependence among d Euclidean random variables (see, e.g., Schweizer & Wolff, 1981; Gaißer, Ruppert & Schmid, 2010; Póczos, Ghahramani & Schneider, 2012; Roy et al., 2022). Here we consider a measure T having the following properties.

1. $T(G) \geq 0$ for all $G \in \mathcal{M}(\mathbb{R}^d)$,
2. $T(G) = 0$ if and only if the components of the corresponding random vector are independent.

If $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$ are independent, (i.e., $\mathbb{P} = \otimes_{i=1}^{d} \mathrm{P}_i$), then for any functional T with these properties, we have $\xi^{\nu'}(\mathbb{P}) = 0$, but for any $\mathbb{P} \in \mathcal{M}(\bigoplus_{i=1}^{d} \mathcal{H}_i)$, $\xi^{\nu'}(\mathbb{P}) = 0$ only implies $T(\mathbb{P}^{\boldsymbol{f}}) = 0$ almost surely w.r.t. the measure ν', not necessarily the independence among the $X^{(i)}$'s.

If the $\mathcal{H}_i$'s are Euclidean spaces, using a result in Rawat & Sitaram (2000), one can show that $\xi^{\nu'}(\mathbb{P}) = 0$ implies the independence of $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$ if ν' is not singular with respect to the Lebesgue measure (see Roy et al., 2020). In the absence of Lebesgue measure in infinite dimensional Hilbert spaces, we need an analogous result. One such result is stated below.

Theorem 6.1. *If* $\text{supp}\{\nu'\} = \bigoplus_{j=1}^{d} \mathcal{H}_j$, *then* $\xi^{\nu'}(\mathbb{P}) = 0$ *implies* $\mathbb{P} = \otimes_{j=1}^{d} \mathrm{P}_j$.

Theorem 6.1 gives a characterization of independence among $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$. One can generate several $\boldsymbol{f}$'s from the probability distribution ν' to approximate $\xi^{\nu'}(\mathbb{P})$. However, if the generated values are nearly orthogonal to the support of $\mathbb{P}$, $\xi^{\nu'}(\mathbb{P})$ takes very small values even when the $X^{(j)}$'s are highly dependent. Therefore, in practice, it is advantageous to work with $\nu' = \mathbb{P}$. For this choice of ν', the characterization property of $\xi^{\mathbb{P}}(\mathbb{P})$ holds if $\overline{span\{\text{supp}\{\mathbb{P}\}\}} = \bigoplus_{j=1}^{d} \mathcal{H}_j^0$, for $\mathcal{H}_j^0$ being a closed subspace of $\mathcal{H}_j$ $(j = 1, 2, \ldots, d)$. This is asserted by the following theorem.

Theorem 6.2. *If* $\overline{span\{\text{supp}\{\mathbb{P}\}\}} = \bigoplus_{j=1}^{d} \mathcal{H}_j^0$, *then* $\xi^{\mathbb{P}}(\mathbb{P}) = 0$ *implies* $\mathbb{P} = \otimes_{j=1}^{d} \mathrm{P}_j$.

The above condition is quite natural in functional data analysis. In practice, observed functions are often approximated by a linear combination of finitely many basis functions of a suitable function space, and any finite-dimensional subspace of a Hilbert space is closed.

6.1 Projected Hilbert-Schmidt Independence Criterion

There are several choices of T satisfying the two properties (1 and 2) mentioned before. In this chapter, we consider T to be the d-variate Hilbert-Schmidt Independence Criterion (dHSIC) (Pfister et al., 2018) based on the mean embedding of multivariate distributions into a reproducing kernel Hilbert space $\mathbb{H} = H^1 \otimes \cdots \otimes H^d$. The mean embedding $\mathcal{E} : \mathcal{M}(\mathbb{R}^d) \to \mathbb{H}$ of a d-variate distribution $\mathbb{Q}$ is defined as $\mathcal{E}(\mathbb{Q}) = \int K(x, \cdot) d\mathbb{Q}(x)$, where $K = K^1 \otimes \cdots \otimes K^d$ is the reproducing kernel associated with $\mathbb{H}$. The kernel $K(\cdot, \cdot)$ is called a characteristic kernel if the map $\mathcal{E}$ is one-to-one. If $\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_d$ are one-dimensional marginals of $\mathbb{Q}$, dHSIC is defined as $\text{dHSIC}(\mathbb{Q}) = \|\mathcal{E}(\mathbb{Q}) - \mathcal{E}(\mathcal{Q}_1 \otimes \cdots \otimes \mathcal{Q}_d)\|_{\mathbb{H}}^2$, where $\|\cdot\|_{\mathbb{H}}$ is the norm in $\mathbb{H}$. If $K(\cdot, \cdot)$ is a characteristic kernel, $\text{dHSIC}(\mathbb{Q}) = 0$ if and only if $\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_d$ are independent. For any fixed $\boldsymbol{f}$, here we use $T(\mathbb{P}^{\boldsymbol{f}}) = \text{dHSIC}(\mathbb{P}^{\boldsymbol{f}})$. Since the measure $\xi(\mathbb{P}) = \xi^{\mathbb{P}}(\mathbb{P}) = \int_{\bigoplus_{i=1}^{d} \mathcal{H}_i} T(\mathbb{P}^{\boldsymbol{f}}) d\mathbb{P}(\boldsymbol{f})$ is obtained by aggregating dHSIC computed along different projection directions, we call it projected HSIC (pHSIC). This can be expressed as

$$
\begin{aligned}
\xi(\mathbb{P}) =& \mathbb{E}_{\mathbf{X_1}}\Big[\mathbb{E}_{\mathbf{X_2},\mathbf{X_3}}\Big(\prod_{j=1}^{d} K^j(\langle X_2^{(j)}, X_1^{(j)}\rangle, \langle X_3^{(j)}, X_1^{(j)}\rangle)\Big)\Big] \\
&+ \mathbb{E}_{\mathbf{X_1}}\Big[\prod_{j=1}^{d} \mathbb{E}_{X_2^{(j)}, X_3^{(j)}}\Big(K^j(\langle X_2^{(j)}, X_1^{(j)}\rangle, \langle X_3^{(j)}, X_1^{(j)}\rangle)\Big)\Big] \\
&- 2\mathbb{E}_{\mathbf{X_1}}\mathbb{E}_{\mathbf{X_2}}\Big[\prod_{j=1}^{d} \mathbb{E}_{X_3^{(j)}}\Big(K^j(\langle X_2^{(j)}, X_1^{(j)}\rangle, \langle X_3^{(j)}, X_1^{(j)}\rangle)\Big)\Big],
\end{aligned}
$$

where $\mathbf{X}_i = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(d)}) \overset{iid}{\sim} \mathbb{P}\,(i = 1, 2, 3)$ and $K^j : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ $(i = 1, 2, \ldots, d)$ is a symmetric, bounded, positive semi-definite characteristic kernel (note that K is a characteristic kernel if and only if all K^js are characteristic kernels, see Gretton (2015)). Following Pfister et al. (2018), we can also write $\xi(\mathbb{P}) = \mathbb{E}\{g(\mathbf{X}_1, \ldots, \mathbf{X}_{2d+1})\}$, where

$$\begin{aligned} g(\mathbf{X}_1, \ldots, \mathbf{X}_{2d+1}) = & \frac{1}{(2d+1)!} \sum_{\pi \in \mathcal{S}_{2d+1}} \Bigg\{ \prod_{j=1}^{d} K^j(\langle X^{(j)}_{\pi(1)}, X^{(j)}_{\pi(2d+1)} \rangle, \langle X^{(j)}_{\pi(2)}, X^{(j)}_{\pi(2d+1)} \rangle) \\ & + \prod_{j=1}^{d} K^j(\langle X^{(j)}_{\pi(2j-1)}, X^{(j)}_{\pi(2d+1)} \rangle, \langle X^{(j)}_{\pi(2j)}, X^{(j)}_{\pi(2d+1)} \rangle) \\ & - 2 \prod_{j=1}^{d} K^j(\langle X^{(j)}_{\pi(1)}, X^{(j)}_{\pi(2d+1)} \rangle, \langle X^{(j)}_{\pi(j+1)}, X^{(j)}_{\pi(2d+1)} \rangle) \Bigg\}. \end{aligned} \tag{6.1}$$

Some intriguing properties of pHSIC are stated below as Proposition 6.1.

Proposition 6.1. *The dependency measure ξ has the following properties.*

(a) If the K^js $(j = 1, 2, \ldots, d)$ are characteristic kernels, under the assumption $\overline{span\{\text{supp}\{\mathbb{P}\}\}} = \bigoplus_{j=1}^{d} \mathcal{H}_j^0$, we have $\xi(\mathbb{P}) = 0$ if and only if $\mathbb{P} = \otimes_{j=1}^{d} \mathrm{P}_j$.

(b) If the K^js $(j = 1, 2, \ldots .d)$ are equal, $\xi(\mathbb{P})$ is invariant under permutation of $(X^{(1)}, \ldots, X^{(d)})$.

(c) $\xi(\mathbb{P})$ is invariant under unitary operations on $(X^{(1)}, \ldots, X^{(d)})$, i.e., if $U_j : \mathcal{H}_j \to \mathcal{X}_j$ is an unitary operator for all $j = 1, 2, \ldots, d$, then $\xi(\mathbb{P} \circ U^{-1}) = \xi(\mathbb{P})$, where $U(x^{(1)}, x^{(2)}, \ldots, x^{(d)}) = (U_1(x^{(1)}), U_2(x^{(2)}), \ldots, U_d(x^{(d)}))$.

(d) Let $\{\mathbf{X}_n : n \geq 1\}$ be a sequence of d-component Hilbertian random vectors such that $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$. If the K^js $(j = 1, 2, \ldots, d)$ are bounded and continuous, then $\lim_{n \to \infty} \xi(\mathcal{L}(\mathbf{X}_n)) = \xi(\mathcal{L}(\mathbf{X}))$, where $\mathcal{L}(\mathbf{Z})$ is the distribution of $\boldsymbol{Z}$.

Remark 6.1. *Proposition 6.1(a) essentially says that ξ characterizes independence, and it is irreducible in the sense that it is not a function of lower order marginals $\{\xi(\mathbb{P} \circ \pi^{-1}_{i_1, \ldots, i_k}) : \{i_1, \ldots, i_k\} \subset \{1, 2, \ldots, d\}\}$, where $\pi_{i_1, \ldots, i_k}(x^{(1)}, x^{(2)}, \ldots, x^{(d)}) = (x^{(i_1)}, \ldots, x^{(i_k)})$.*

Remark 6.2. *Proposition 6.1(c) implies that ξ only depends on the structure of the inner product defined on the Hilbert space, but not on the space used for modeling. For example, modeling the same data as random variables in $L_2[0, 1]$ and in $L_2[0, 10]$ leads to the same value of $\xi(\mathbb{P})$.*

6.1.1 Estimation of pHSIC

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample on $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)}) \sim \mathbb{P}$. Replacing $\mathbb{P}$ in $\xi(\mathbb{P})$ by its empirical version $\hat{\mathbb{P}}_n$, which puts mass $1/n$ on each of the data point, we get

$$\begin{aligned} \hat{\xi}_n = & \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \prod_{\ell=1}^{d} K^\ell(\langle X_j^{(\ell)}, X_i^{(\ell)} \rangle, \langle X_k^{(\ell)}, X_i^{(\ell)} \rangle) + \frac{1}{n^{2d+1}} \sum_{i=1}^{n} \prod_{\ell=1}^{d} \sum_{1 \leq j,k \leq n} K^\ell(\langle X_j^{(\ell)}, X_i^{(\ell)} \rangle, \langle X_k^{(\ell)}, X_i^{(\ell)} \rangle) \\ & - \frac{2}{n^{d+2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \prod_{\ell=1}^{d} \sum_{k=1}^{n} K^\ell(\langle X_j^{(\ell)}, X_i^{(\ell)} \rangle, \langle X_k^{(\ell)}, X_i^{(\ell)} \rangle) \end{aligned}$$

as an estimator of $\xi(\mathbb{P})$ for $n \geq 2d+1$. This estimator has properties similar to $\xi(\mathbb{P})$. It can be viewed as a V-statistic with the core function g (see (6.1)) and is consistent under fairly general assumptions. This is stated as a theorem below.

Theorem 6.3. *Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} \mathbb{P}$, *and* $\sigma_c^*(g) = Var\big(\mathbb{E}\{g(\mathbf{X}_1, \ldots, \mathbf{X}_{2d+1})|\mathbf{X}_1, \ldots, \mathbf{X}_c\}\big)$ *for* $c = 1, 2$. *If the* K^j*'s are bounded, we have the following results.*

(a) Under H_0, *if* $\sigma_2^*(g) > 0$, *there exists a real sequence* $\{\lambda_i\}$ *such that* $n\hat{\xi}_n \overset{D}{\to} \sum_{i=1}^{\infty} \lambda_i Z_i^2$, *where* $\{Z_i\}$ *is a sequence of i.i.d* $\mathcal{N}_1(0,1)$ *random variables. If* $\sigma_2^*(g) = 0$, $n\hat{\xi}_n$ *becomes a degenerate random variable.*

(b) Under H_1, *if* $\sigma_1^*(g) > 0$, $\sqrt{n}(\hat{\xi}_n - \xi(\mathbb{P})) \overset{D}{\to} (2d+1)\sqrt{\sigma_1^*(g)}Z_1$, *where* $Z_1 \sim N(0,1)$. *If* $\sigma_1^*(g) = 0$, $n\mathbb{E}\big(\hat{\xi}_n - \xi(\mathbb{P})\big)^2 \to 0$ *as* $n \to \infty$.

As a corollary of Theorem 6.3, we get the consistency of $\hat{\xi}_n$, which is stated below.

Corollary 6.1. *If* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} \mathbb{P}$, $\hat{\xi}_n$ *converges in probability to* $\xi(\mathbb{P})$ *as* $n \to \infty$.

6.1.2 Test of Independence based on pHSIC

We have seen that $\xi(\mathbb{P})$ serves as a useful measure of dependency. It is non-negative, and under fairly general assumptions, it takes value 0 if and only of $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$ are independent. We have also observed that it can be consistently estimated by $\hat{\xi}_n$. So, if $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$ are dependent, $\hat{\xi}_n$ is supposed to take higher values. Based on this idea, we can construct a test using $\hat{\xi}_n$ as the test statistic and reject H_0 for large values of $\hat{\xi}_n$. For a level α test ($0 < \alpha < 1$), this threshold r_α is chosen based on the permutation method.

Our test needs the kernel functions $K^j(\cdot,\cdot)$'s to be specified. Throughout this article, for our numerical work, we use the Gaussian kernel $K^j(x,y) = K_{\sigma_j}(x,y) = \exp\{-(x-y)^2/2\sigma_j^2\}$ ($j = 1, 2, \ldots, d$), which involves a bandwidth parameter $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_d)$. Henceforth, we use the notations $\xi_{\boldsymbol{\sigma}}(\mathbb{P})$ and $\hat{\xi}_{\boldsymbol{\sigma},n}$ to denote the dependency measure and the corresponding estimator based on Gaussian kernel (with bandwidth $\boldsymbol{\sigma}$), respectively. Since, it is a characterizing kernel (see, .g., Gretton et al., 2007), we have $\xi_{\boldsymbol{\sigma}}(\mathbb{P}) = 0$ if and only if $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$ are independent (see Proposition 2.1(a)). Again, for a fixed choice of $\boldsymbol{\sigma}$, $\hat{\xi}_{\boldsymbol{\sigma},n} \overset{P}{\to} \xi_{\boldsymbol{\sigma}}(\mathbb{P})$ (see Theorem 6.3). So, for any fixed $\boldsymbol{\sigma}$, the consistency of the resulting test follows from that. However, the finite sample power of the test may depend on the choice of $\boldsymbol{\sigma}$. A popular way to choose the value of $\boldsymbol{\sigma}$ is based on "median heuristic" (see, e.g., Gretton et al., 2007). Following that idea, we can choose $2\hat{\sigma}_j^2$ ($j = 1, 2, \ldots, d$) equal to the median of $\{\langle x_i^{(j)}, x_k^{(j)} - x_\ell^{(j)}\rangle^2 : i = 1, 2, \ldots, n, 1 \leq k < \ell \leq n\}$. Note that as n increases, $2\hat{\sigma}_j^2 \overset{P}{\to} 2\sigma_{0j}^2 = Median\{\langle X_1^{(j)}, X_2^{(j)} - X_3^{(j)}\rangle^2\}$, where $X_1^{(j)}, X_2^{(j)}, X_3^{(j)}$ are i.i.d. copies of $X^{(j)} \sim \mathrm{P}_j$ ($j = 1, 2, \ldots, d$). If $\boldsymbol{\sigma}_0 = (\sigma_{01}, \sigma_{02}, \ldots, \sigma_{0d})$ is unique, our test remains consistent for such a data-driven choice of the bandwidth. This result is stated below.

Theorem 6.4. *If* $\boldsymbol{\sigma}_{(n)} = (\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_d)$ *is the bandwidth chosen based on median heuristic, for any fixed alternative, the power of the test based on* $\hat{\xi}_{\boldsymbol{\sigma}_{(n)},n}$ *converges to* 1 *as* n *increases to infinity.*

Recently, Miao, Zhang & Wong (2023) considered smoothing the observed functions using wavelets and used HSIC on the recovered functions to construct a test of independence. However, they proposed their method for two functions only, whereas our test based on $\hat{\xi}_{\boldsymbol{\sigma}_{(n)},n}$ can be conveniently used even for more than two functions.

6.2 Results for independence between two random functions

We analyze some simulated data sets to compare the finite sample performance of our proposed test with dCov (Lyons, 2013), bCov (Pan et al., 2020) and aCov (Lai et al., 2021) tests. In the case of aCov test, the raw data need to be transformed using a system of basis functions. We used two different choices of basis functions, the Fourier basis and the spline basis. The corresponding tests are referred to as aCov_1 and aCov_2, respectively. Throughout this section, all tests are considered to have a 5% nominal level. The cut-offs of all tests are computed using the permutation principle based on 500 random permutations.

Here, all simulated functions are considered as elements of $L_2[0,1]$. We consider the model $X^{(1)}(t) = \sum_{i=1}^{9} \frac{1}{i^2}\xi_i\phi_i(t)$ and $X^{(2)}(t) = \sum_{i=1}^{9} \frac{1}{i^3}\eta_i\phi_i(t)$ for $t \in [0,1]$, where $\{\phi_k(t)\}_{k\geq 0}$ is the trigonometric basis defined as $\{\phi_0(t) = 1, \phi_{2k}(t) = \sqrt{2}\sin(2\pi kt),\ \phi_{2k+1}(t) = \sqrt{2}\cos(2\pi kt)\}_{k\geq 0}$, and the (ξ_i, η_i)'s are i.i.d. bivariate random vectors. It is easy to see that $X^{(1)}$ and $X^{(2)}$ are independent if and only if ξ_i and η_i are independent for all $i = 1, 2, \ldots, 9$. Using this model, we generated observations on $X^{(1)}$ and $X^{(2)}$, and they were observed on 101 equally spaced grid points on $[0,1]$. Each experiment was repeated 1000 times to estimate the power of a test by the proportion of times it rejected H_0. We begin with the following examples.

Example 6.1. *Generate (ξ_i, η_i)'s from $\mathcal{N}_2(\mathbf{0}_2, \mathbf{I}_2)$.*

Clearly, in Example 6.1, $X^{(1)}$ and $X^{(2)}$ are independent. We used it to check the level properties of different tests. We carried out our experiment with samples of different sizes, and in all cases, all tests had powers close to the nominal level of 0.05 (see Figure 6.1(a)).

Example 6.2. *Generate the (ξ_i, η_i)'s from the standard bivariate Cauchy distribution.*

In this example, $X^{(1)}$ and $X^{(2)}$ are dependent, and this dependency was well detected by the bCov test and our proposed test (see Figure 6.1(b)). The dCov test also had competitive performance, but both versions of the aCov test had much lower power.

Example 6.3. *Generate the (ξ_i, η_i)'s from $\mathcal{N}(0,0;1,1;\rho)$ distribution.*

Here we considered different values of ρ ranging between -1 and 1 and computed the powers of different tests keeping the sample size $n = 40$. Figure 6.2(a) shows that all tests had satisfactory performance, but among them, dCov and aCov tests had an edge over the other tests.

Example 6.4. *Generate the (ξ_i, η_i)'s from a mixture distribution $(1-\gamma)F + \gamma G$, where F denotes the $\mathcal{N}(0,0;1,1;0.5)$ and G denotes a distribution with two independent standard Cauchy variates.*

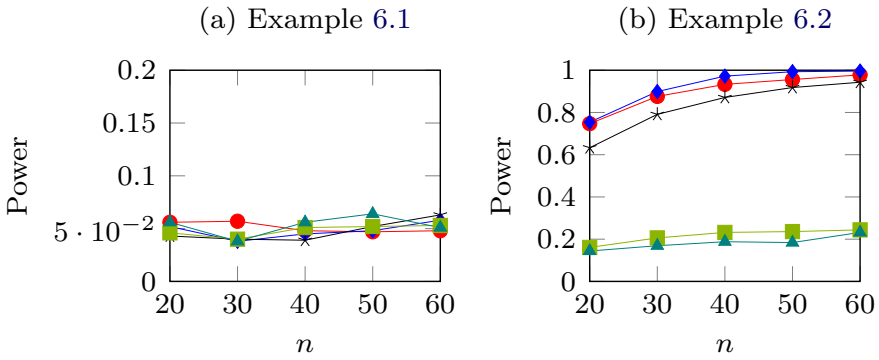


Fig. 6.1 *Powers of pHSIC (●), bCov (◆), dCov (⋆), aCov$_1$ (■) and aCov$_2$ (▲) tests in Examples 6.1-6.2.*

Example 6.4 deals with a contaminated version of Example 6.3 with $\rho = 0.5$. Here, we considered different contamination proportions $\gamma \in [0, 0.5]$, and the results are reported in Figure 6.2(b). Note that as γ increases, the strength of dependence between $X^{(1)}$ and $X^{(2)}$ decreases. So, one would expect the powers of the tests to decrease. Figure 6.2(b) shows that the impact of contamination was more on the dCov test. Its power dropped sharply as γ increased. The aCov tests exhibited a similar performance but showed little resistance against contamination. Our test was most robust against this contamination, and for higher values of γ, it had the highest power among the competitors considered here. The bCov test also had a competitive performance.

Next, we consider some examples, where we generate the (ξ_i, η_i)'s from the six unusual bivariate distributions considered in Newton (2009).

Example 6.5. *Generate the (ξ_i, η_i)'s from the distributions (referred to as (a) Circle, (b) W, (c) Diamond, (d) Parabola, (e) Two parabolas, and (f) Four clouds) given in Figure 6.3.*

Scatter plots of the coefficients generated from these distributions are given in Figure 6.3. In these examples, ξ_i and η_i are uncorrelated, but they are dependent in all cases, barring Example 6.5 (f) (Four clouds). Results reported in Figure 6.4 clearly show that in Examples 6.5 (a)-(e), our test

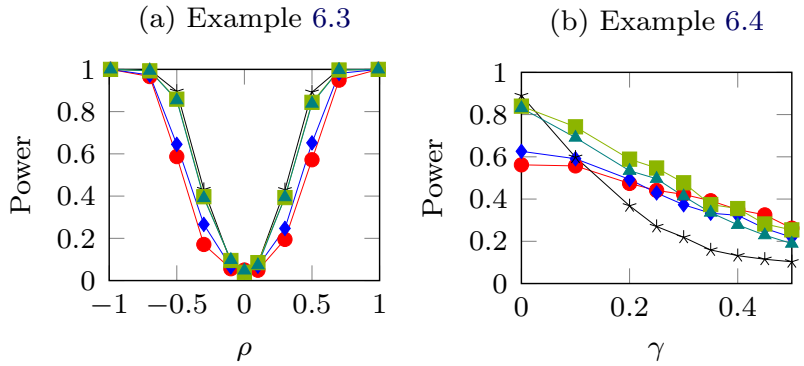


Fig. 6.2 *Powers of pHSIC (●), bCov (◆), dCov (⋆), aCov$_1$ (■) and aCov$_2$ (▲) tests in Examples 6.3-6.4.*

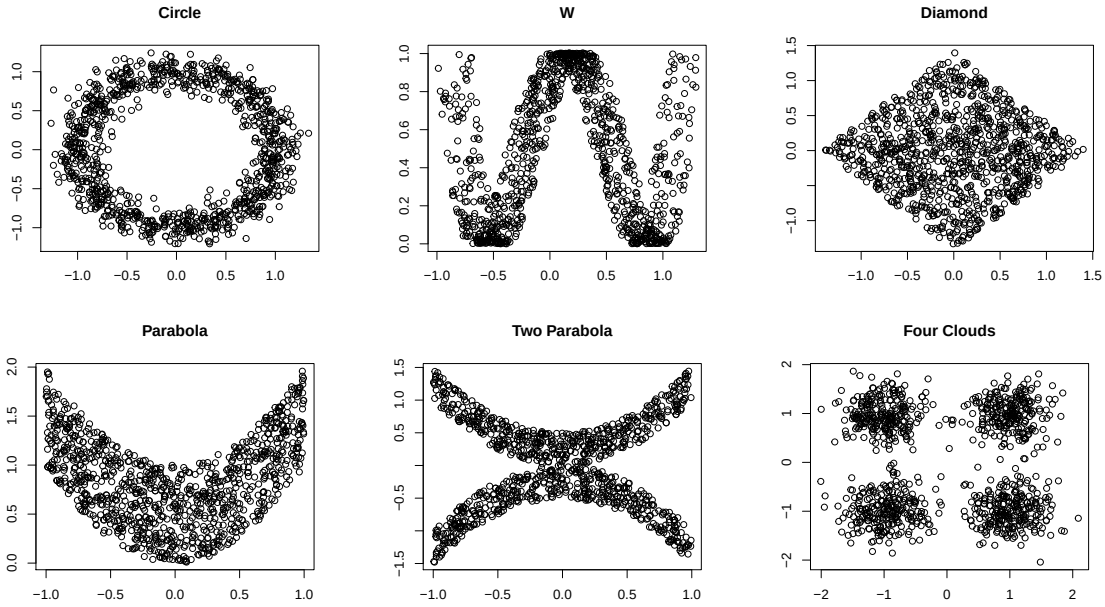


Fig. 6.3 *Scatter plots of observations from the six unusual bivariate distributions in Newton (2009).*

had excellent performance. The bCov test also had competitive performance in these examples. In Examples 6.5 (a) (Circle) and 6.5 (e) (Two parabolas), powers of these two tests were quite similar,

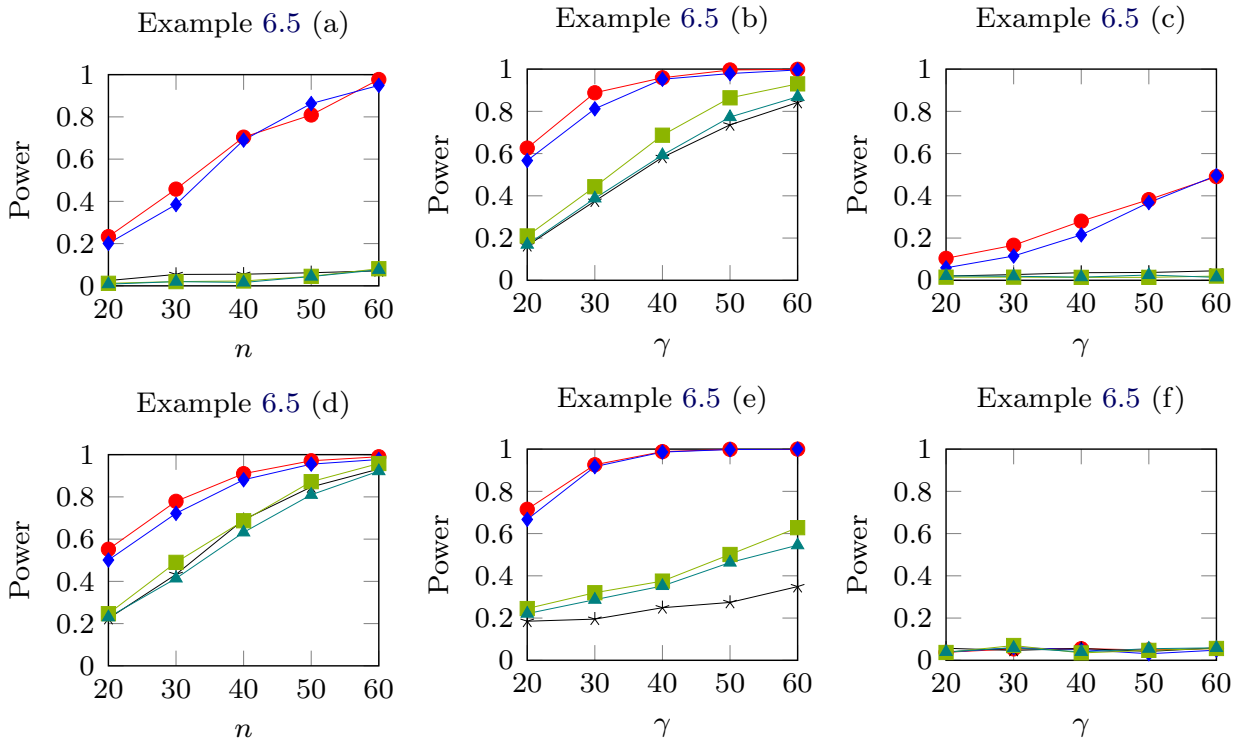


Fig. 6.4 *Powers of pHSIC (●), bCov (◆), dCov (⋆), $aCov_1$ (■) and $aCov_2$ (▲) tests in Examples 6.5(a)-(f).*

but in Examples 6.5 (b)-(d) (W, Diamond and Parabola), our proposed test had an edge, especially when the sample size was small. In Examples 6.5 (a) (Circle) and 6.5 (c) (Diamond), dCov and aCov tests had powers close to the nominal level $\alpha = 0.05$. In Example 6.5 (e) (Two parabolas) also, they had much lower powers. They performed slightly better in Examples 6.5 (b) (W) and 6.5 (d) (Parabola), but even in those two cases, they were outperformed by the bCov test and our proposed test. In Example 6.5 (f) (Four clouds), where $X^{(1)}$ and $X^{(2)}$ are independent, all tests had powers close to the nominal level $\alpha = 0.05$.

6.3 Results for independence among multiple random functions

We also consider some testing problems involving more than two random functions. Note that aCov and dCov tests cannot be used in such situations. So, we compare our performance only with the bCov test. We consider four examples (Examples 6.6-6.9) in this section. In the first two examples, we use the model

$$X^{(1)}(t) = \sum_{i=1}^{9} \frac{1}{i^2}\xi_i\phi_i(t), \qquad X^{(2)}(t) = \sum_{i=1}^{9} \frac{1}{i^3}\eta_i\phi_i(t), \qquad X^{(3)}(t) = \sum_{i=1}^{9} \frac{1}{i^4}\mu_i\phi_i(t)$$

for $t \in [0,1]$, where $\{\phi_k(t)\}$ is the trigonometric basis as defined before.

Example 6.6. *Generate the (ξ_i, η_i, μ_i)'s ($i = 1, 2, \ldots, 9$) independently from the uniform distribution over the sphere $\mathcal{S} = \{(x, y, z) : \ x^2 + y^2 + z^2 \le 1\}$.*

Example 6.7. *Generate the (ξ_i, η_i, μ_i)'s independently from a hyperbolic paraboloid distribution, i.e., $\xi_i, \eta_i \overset{iid}{\sim} \text{Unif}(-1, 1)$ and $\mu_i = \eta_i^2 - \xi_i^2 + \frac{1}{4}Z_i$, for $Z_i \sim \mathcal{N}_1(0, 1)$ independent of ξ_i and η_i.*

We considered samples of different sizes, and in each case, the powers were computed based on 1000 repetitions of the experiment. Figure 6.5 clearly shows that in these two examples, our proposed test outperformed the bCov test. In Example 6.7, the bCov test had a somewhat competitive performance, but in Example 6.6, the power of our proposed test was much higher.

Next, we consider some block diagonal examples, which can be viewed as extensions of some special cases of Example 6.5.

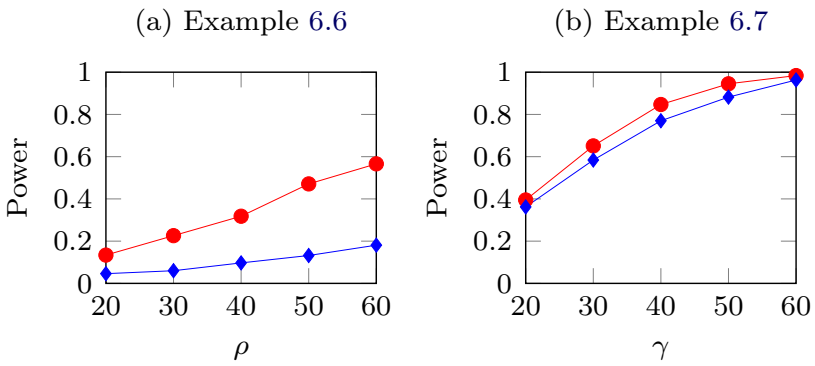


Fig. 6.5 *Powers of pHSIC (●) and bCov (◆) tests in Examples 6.6 and 6.7.*

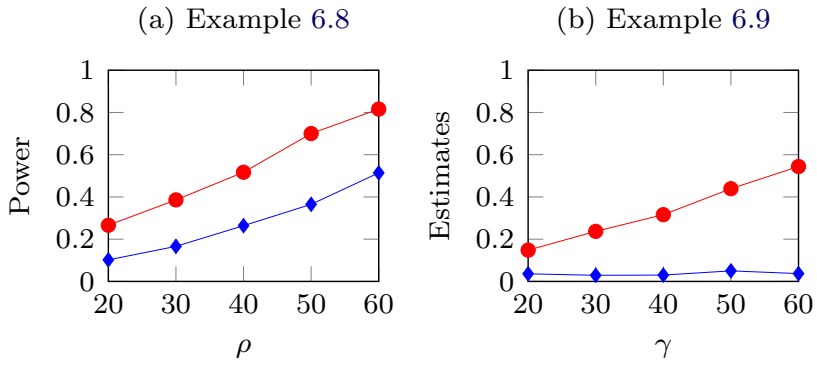


Fig. 6.6 *Powers of pHSIC (●), and bCov (◆) tests in Examples 6.8 and 6.9.*

Example 6.8. *Generate observations on* $(X^{(1)}, X^{(2)})$ *as in Example 6.5 (a) (Circle) and repeat the same method for generating observations on* $(X^{(3)}, X^{(4)})$ *independently.*

Example 6.9. *Generate observations on* $(X^{(1)}, X^{(2)})$ *and* $(X^{(3)}, X^{(4)})$ *independently, following the model used in Example 6.5 (c) (Diamond).*

Here also we considered different sample sizes, and in each case, the powers of the tests were computed based on 1000 replications. Figure 6.6 clearly shows that the proposed test performed much better than the bCov test in these two examples, and the differences between their powers were more prominent than what was observed in Examples 6.5 (a) and 6.5 (c). We also considered similar extensions of the other cases in Example 6.5, but in those cases, the performance of our test was either similar or marginally better than the bCov test. Therefore, to avoid repetition, we decided not to report those results here.

6.4 Proofs and mathematical details

Proof of Theorem 6.1. Let P_j $(j = 1, 2, \ldots, d)$ be the marginal distribution of $X^{(j)}$ and $\mathbb{P}$ be the joint distribution of $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$. Define $\varphi : \bigoplus_{j=1}^{d} \mathcal{H}_j \to \mathbb{C}$ as

$$\varphi(f_1, f_2, \ldots, f_d) = \mathbb{E}\{e^{i\sum_{j=1}^{d}\langle X^{(j)}, f_j\rangle}\} - \prod_{j=1}^{d} \mathbb{E}\{e^{i\langle X^{(j)}, f_j\rangle}\}.$$

Notice that the function φ is continuous. So, if φ is zero on a dense subset of $\bigoplus_{j=1}^{d} \mathcal{H}_j$, it will imply $\varphi(f_1, f_2, \ldots, f_d) = 0$ $\forall (f_1, \ldots, f_d) \in \bigoplus_{j=1}^{d} \mathcal{H}_j$, or

$$\mathbb{E}\{e^{i\sum_{j=1}^{d}\langle X^{(j)}, f_j\rangle}\} = \prod_{j=1}^{d} \mathbb{E}\{e^{i\langle X^{(j)}, f_j\rangle}\} \ \forall (f_1, \ldots, f_d) \in \bigoplus_{j=1}^{d} \mathcal{H}_j,$$

i.e., the mutual independence of $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$. Now $\xi^{\nu'}(\mathbb{P}) = 0$ implies there exists a Borel measurable set $E \in \mathcal{B}(\bigoplus_{j=1}^{d} \mathcal{H}_j)$ such that $\nu'(E) = 1$ and $T(\mathbb{P}^{\boldsymbol{f}}) = 0$ $\forall \boldsymbol{f} \in E$. Hence, by the assumption on T (see page 127), we have $\varphi(f_1, f_2, \ldots, f_d) = 0$ $\forall (f_1, f_2, \ldots, f_d) \in E$. So, when

$\text{supp}\{\nu'\} = \bigoplus_{j=1}^{d} \mathcal{H}_j$, we must have $\overline{E} = \bigoplus_{j=1}^{d} \mathcal{H}_j$, otherwise $\overline{E} \subset \bigoplus_{j=1}^{d} \mathcal{H}_j$ will contain the support of ν', which leads to a contradiction. This completes the proof. ■

Proof of Theorem 6.2. Before proving the main result, let us consider the following claim.

Claim: Suppose that X is a random function taking values in a Hilbert space $\mathcal{H}$, and it has the probability distribution F. Also assume that $\text{supp}\{F\} \subset \mathcal{H}_0$, where $\mathcal{H}_0$ is a closed subspace of $\mathcal{H}$. If $Q : \mathcal{H} \to \mathcal{H}_0$ is the projection operator onto $\mathcal{H}_0$, then X and QX have the same distribution. This can be shown using the characteristic function of F. Note that for any arbitrary $\theta \in \mathcal{H}$,

$$\begin{aligned}\mathbb{E}\{e^{i\langle X,\theta\rangle}\} &= \int_{\mathcal{H}} e^{i\langle x,\theta\rangle}\, dF(x) = \int_{\mathcal{H}_0} e^{i\langle x,\theta\rangle}\, dF(x) = \int_{\mathcal{H}_0} e^{i\langle Qx,Q\theta\rangle + i\langle (I-Q)x,(I-Q)\theta\rangle}\, dF(x) \\ &= \int_{\mathcal{H}_0} e^{i\langle Qx,Q\theta\rangle + i\langle 0,(I-Q)\theta\rangle}\, dF(x) = \int_{\mathcal{H}_0} e^{i\langle Qx,\theta\rangle}\, dF(x) = \mathbb{E}\{e^{i\langle QX,\theta\rangle}\}.\end{aligned}$$

Now, we shall show that $\varphi(f_1, f_2, \ldots, f_d) = 0$, $\forall (f_1, f_2, \ldots, f_d) \in \bigoplus \mathcal{H}_j$. Recall the condition $\overline{span\{\text{supp}\{\mathbb{P}\}\}} = \bigoplus \mathcal{H}_j^0$ and define Q_j as the projection operator from $\mathcal{H}_j$ to $\mathcal{H}_j^0$. Then

$$\begin{aligned}\varphi(f_1, \ldots, f_d) &= \mathbb{E}\{e^{i\sum_{j=1}^{d}\langle X^{(j)}, f_j\rangle}\} - \prod_{j=1}^{d} \mathbb{E}\{e^{i\langle X^{(j)}, f_j\rangle}\} \\ &= \int_{\bigoplus_{i=1}^{d}\mathcal{H}_i} e^{i\sum_{j=1}^{d}\langle x^j, f_j\rangle} \mathrm{d}\mathbb{P}(x^1, \ldots, x^d) - \prod_{j=1}^{d} \int_{\mathcal{H}_j} e^{i\langle x, f_j\rangle} \mathrm{dP}_j(x) \\ &= \int_{\bigoplus_{i=1}^{d}\mathcal{H}_i^0} e^{i\sum_{j=1}^{d}\langle x^j, f_j\rangle} \mathrm{d}\mathbb{P}(x^1, \ldots, x^d) - \prod_{j=1}^{d} \int_{\mathcal{H}_j^0} e^{i\langle x, f_j\rangle} \mathrm{dP}_j(x) \\ &= \int_{\bigoplus_{i=1}^{d}\mathcal{H}_i^0} e^{i\sum_{j=1}^{d}\langle Q_j x^j, Q_j f_j\rangle} \mathrm{d}\mathbb{P}(x^1, \ldots, x^d) - \prod_{j=1}^{d} \int_{\mathcal{H}_j^0} e^{i\langle Q_j x, Q_j f_j\rangle} \mathrm{dP}_j(x) \\ &= \int_{\bigoplus_{i=1}^{d}\mathcal{H}_i^0} e^{i\sum_{j=1}^{d}\langle x^j, Q_j f_j\rangle} \mathrm{d}\mathbb{P}(x^1, \ldots, x^d) - \prod_{j=1}^{d} \int_{\mathcal{H}_j^0} e^{i\langle x, Q_j f_j\rangle} \mathrm{dP}_j(x) \quad (\text{as } X^{(j)} \overset{D}{=} Q_j X^{(j)}) \\ &= \varphi(Q_1 f_1, \ldots, Q_d f_d).\end{aligned}$$

This shows that under the assumption $\overline{span\{\text{supp}\{\mathbb{P}\}} = \bigoplus_{j=1}^{d} \mathcal{H}_j^0$, we only need to consider the value of $\varphi(\cdot)$ on $\bigoplus_{j=1}^{d} \mathcal{H}_j^0$. Now, if $\xi(\mathbb{P}) = 0$, we have $\varphi(f_1, \ldots, f_d) = 0$ for all $(f_1, f_2, \ldots, f_d) \in \text{supp}\{\mathbb{P}\}$. So, $\langle X^{(1)}, f_1\rangle, \langle X^{(2)}, f_2\rangle, \ldots, \langle X^{(d)}, f_d\rangle$ are independent for all $(f_1, f_2, \ldots, f_d) \in \text{supp}\{\mathbb{P}\}$. Now, for any $(g_1, g_2, \ldots, g_d) \in \overline{span\{\text{supp}\{\mathbb{P}\}\}}$, we can write $g_j = \sum_{i=1}^{\infty} a_{ji} f_{ji}$ for $\{f_{ji}\} \subset \mathcal{H}_j$ and $\{a_{ji}\} \subset \mathbb{R}$. This implies $\langle X^{(j)}, g_j\rangle = \sum_{i=1}^{\infty} a_{ji}\langle X^{(j)}, f_{ji}\rangle$ are independent for all $j = 1, 2, \ldots, d$. Therefore, $\langle X^{(1)}, g_1\rangle, \ldots, \langle X^{(d)}, g_d\rangle$ are independent for any fixed $(g_1, g_2, \ldots, g_d) \in \overline{span\{\text{supp}\{\mathbb{P}\}\}}$, i.e., $\varphi(g_1, g_2, \ldots, g_d) = 0$ $\forall (g_1, \ldots, g_d) \in \overline{span\{\text{supp}\{\mathbb{P}\}\}}$. Now, by the assumption $\overline{span\{\text{supp}\{\mathbb{P}\}\}} = \bigoplus \mathcal{H}_j^0$, we have $\varphi(g_1, g_2, \ldots, g_d) = 0$ $\forall (g_1, \ldots, g_d) \in \bigoplus \mathcal{H}_j$, i.e., $\mathbb{E}\{e^{i\sum_{j=1}^{d}\langle X^{(j)}, f_j\rangle}\} = \prod_{j=1}^{d} \mathbb{E}\{e^{i\langle X^{(j)}, f_j\rangle}\}$ for all $(f_1, \ldots, f_d) \in \bigoplus_{j=1}^{d} \mathcal{H}_j$. ■

Proof of Proposition 6.1. (a) The proof follows from Theorem 6.2.

(b) Notice that the main term in the expression of $\xi(\mathbb{P})$ is $\prod_{j=1}^{d} K^j(\langle X_2^{(j)}, X_1^{(j)}\rangle, \langle X_3^{(j)}, X_1^{(j)}\rangle)$. If K^j's are all equal say K_0, then for any permutation $\pi : \{1, \ldots, d\} \to \{1, \ldots, d\}$, we have

$$\prod_{j=1}^{d} K_0(\langle X_2^{(j)}, X_1^{(j)}\rangle, \langle X_3^{(j)}, X_1^{(j)}\rangle) = \prod_{j=1}^{d} K_0(\langle X_2^{(\pi(j))}, X_1^{(\pi(j))}\rangle, \langle X_3^{(\pi(j))}, X_1^{(\pi(j))}\rangle).$$

The permutation invariance of $\xi(\mathbb{P})$ follows from this expression.

(c) Since $\langle x, y\rangle = \langle Ux, Uy\rangle$ holds under any unitary operation U, we have

$$\prod_{j=1}^{d} K^j(\langle X_2^{(j)}, X_1^{(j)}\rangle, \langle X_3^{(j)}, X_1^{(j)}\rangle) = \prod_{j=1}^{d} K^j(\langle U_j X_2^{(j)}, U_j X_1^{(j)}\rangle, \langle U_j X_3^{(j)}, U_j X_1^{(j)}\rangle),$$

for unitary operations U_j on $\mathcal{H}_j$ $(j = 1, 2, \ldots, d)$. This ensures the unitary operation invariance property for pHSIC.

(d) Let $\mathbf{X}_n$ be a sequence of d-component Hilbertian random variables such that $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$. Since K^j's are bounded and continuous, it follows that g is a bounded continuous function on $(\bigoplus_{j=1}^{d} \mathcal{H}_j)^{2d+1}$. Now, using the Dominated Convergence Theorem, it follows that

$$\lim_{n\to\infty} \xi(\mathcal{L}(\mathbf{X}_n)) = \lim_{n\to\infty} \mathbb{E}(g(\mathbf{X}_{n,1}, \ldots, \mathbf{X}_{n,2d+1})) = \mathbb{E}(g(\mathbf{X}_1, \ldots, \mathbf{X}_{2d+1})) = \xi(\mathcal{L}(\mathbf{X})). \qquad \blacksquare$$

Proof of Theorem 6.3. Before going into the details of the proof, first notice that our estimator $\hat{\xi}_n$ can also be written as

$$\hat{\xi}_n = \frac{1}{n^{2d+1}} \sum_{i_1=1}^{n} \cdots \sum_{i_{2d+1}=1}^{n} g(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{2d+1}}),$$

$$\begin{aligned}\text{where } \; g(\mathbf{X}_1, \ldots, \mathbf{X}_{2d+1}) = &\frac{1}{(2d+1)!} \sum_{\pi \in \mathcal{S}_{2d+1}} \{\prod_{j=1}^{d} K^j(\langle X_{\pi(1)}^{(j)}, X_{\pi(2d+1)}^{(j)}\rangle, \langle X_{\pi(2)}^{(j)}, X_{\pi(2d+1)}^{(j)}\rangle) \\ &+ \prod_{j=1}^{d} K^j(\langle X_{\pi(2j-1)}^{(j)}, X_{\pi(2d+1)}^{(j)}\rangle, \langle X_{\pi(2j)}^{(j)}, X_{\pi(2d+1)}^{(j)}\rangle) \\ &- 2\prod_{j=1}^{d} K^j(\langle X_{\pi(1)}^{(j)}, X_{\pi(2d+1)}^{(j)}\rangle, \langle X_{\pi(j+1)}^{(j)}, X_{\pi(2d+1)}^{(j)}\rangle)\},\end{aligned}$$

for $\mathcal{S}_{2d+1}$ being the set of all permutations of $\{1, 2, \ldots, 2d+1\}$. By the assumption of the theorem, we have $|K^j(x, y)| < c^{(j)}$ for all $j = 1, 2, \ldots, d$. This implies $|g| \leq 4\prod_{j=1}^{d} c^{(j)}$ and ensures that $\mathbb{E}\{|g(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{2d+1}})|^r\} < \infty$ for any $r \in \mathbb{N}$ and all $1 \leq i_1 < i_2 < \ldots < i_{2d+1} \leq n$. Therefore, from proposition 3.15 of Shao (2003), we get

$$\text{Bias}(\hat{\xi}_n) = O(n^{-1}) \;\text{ and }\; \text{Var}(\hat{\xi}_n) = \text{Var}(\tilde{\xi}_n) + O(n^{-2}), \tag{6.2}$$

where $\tilde{\xi}_n = \frac{1}{(n)_{2d+1}} \sum_{1\leq i_1<i_2<\cdots<i_{2d+1}\leq n} g(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{2d+1}})$ is the corresponding U-statistic. So, the consistency of $\hat{\xi}_n$ holds, but depending on the degeneracy of g the convergence rate varies.

Here, we find the first-order projection of the core function g and use results from the supplementary materials from Pfister et al. (2018) to prove the theorem. First, we find a closed form expression of $g_1(\boldsymbol{Z}_1) = \mathbb{E}\{g(\mathbf{X}_1, \ldots, \mathbf{X}_{2d+1}) \mid \mathbf{X}_1 = \boldsymbol{Z}_1\}$. For that, we deal with individual terms in g and capture those $\pi(i)$'s that are equal to 1 for each $\pi \in \mathcal{S}_{2d+1}$.

Look at the term $e_1(\pi) := \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X^{(j)}_{\pi(1)}, X^{(j)}_{\pi(2d+1)}\rangle, \langle X^{(j)}_{\pi(2)}, X^{(j)}_{\pi(2d+1)}\rangle) \mid \mathbf{X}_1 = \boldsymbol{Z}_1\}$.

Case $\pi(1) = 1$ or $\pi(2) = 1$ and $\pi(2d+1) \neq 1$:
We have $e_1(\pi) = \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle Z^{(j)}_1, X^{(j)}_1\rangle, \langle X^{(j)}_2, X^{(j)}_1\rangle)\}$. There are $2 * (2d)!$ such terms in $\mathcal{S}_{2d+1}$.

Case $\pi(2d+1) = 1$:
We have $e_1(\pi) = \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X^{(j)}_1, Z^{(j)}_1\rangle, \langle X^{(j)}_2, Z^{(j)}_1\rangle)\}$. There are $(2d)!$ such terms in $\mathcal{S}_{2d+1}$.

Case $\pi(1) \neq 1$ or $\pi(2) \neq 1$ and $\pi(2d+1) \neq 1$:
We have $e_1(\pi) = \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X^{(j)}_1, X^{(j)}_3\rangle, \langle X^{(j)}_2, X^{(j)}_3\rangle)\}$. There are $(2d+1)! - 3 * (2d)! = (2d-2) * (2d)!$ such terms are there in $\mathcal{S}_{2d+1}$.

Next consider the term $e_2(\pi) = \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X^{(j)}_{\pi(2j-1)}, X^{(j)}_{\pi(2d+1)}\rangle, \langle X^{(j)}_{\pi(2j)}, X^{(j)}_{\pi(2d+1)}\rangle) \mid \mathbf{X}_1 = \boldsymbol{Z}_1\}$.

Case $\exists\ r \in \{1, \ldots, 2d\}$ such that $\pi(2r-1) = 1$ or $\pi(2r) = 1$ and $\pi(2d+1) \neq 1$:
We have $e_2(\pi) = \mathbb{E}\{\prod_{j=1, j\neq r}^{d} K^j(\langle X^{(j)}_1, X^{(j)}_2\rangle, \langle X^{(j)}_3, X^{(j)}_2\rangle) * k^r(\langle Z^{(r)}_1, X^{(r)}_2\rangle, \langle X^r_3, X^r_2\rangle)\}$. For each r, there are $2 * (2d)!$ such terms in $\mathcal{S}_{2d+1}$.

Case $\pi(2d+1) = 1$:
We have $e_2(\pi) = \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X^{(j)}_1, Z^{(j)}_1\rangle, \langle X^{(j)}_2, Z^{(j)}_1\rangle)\}$. There are $(2d)!$ such terms in $\mathcal{S}_{2d+1}$.

Now consider the final term $e_3(\pi) = \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X^{(j)}_{\pi(1)}, X^{(j)}_{\pi(2d+1)}\rangle, \langle X^{(j)}_{\pi(j+1)}, X^{(j)}_{\pi(2d+1)}\rangle) \mid \mathbf{X}_1 = \boldsymbol{Z}_1\}$.

Case $\pi(2d+1) = 1$:
We have $e_3(\pi) = \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X^{(j)}_1, Z^{(j)}_1\rangle, \langle X^{(j)}_{j+1}, Z^{(j)}_1\rangle)\}$. There are $(2d)!$ such terms in $\mathcal{S}_{2d+1}$.

Case $\exists\ r \in \{1, \ldots, d\}$ such that $\pi(r+1) = 1$:
We have $e_3(\pi) = \mathbb{E}\{\prod_{j=1, j\neq r}^{d} K^j(\langle X^{(j)}_1, X^{(j)}_{2d+1}\rangle, \langle X^{(j)}_{j+1}, X^{(j)}_{2d+1}\rangle) * k^r(\langle X^{(r)}_1, X^{(r)}_{2d+1}\rangle, \langle Z^{(r)}_1, X^{(r)}_{2d+1}\rangle)\}$. For each r, there are $(2d)!$ such terms in $\mathcal{S}_{2d+1}$.

Case $\pi(1) = 1$:
We have $e_3(\pi) = \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle Z^{(j)}_1, X^{(j)}_{2d+1}\rangle, \langle X^{(j)}_{j+1}, X^{(j)}_{2d+1}\rangle)\}$. There are $(2d)!$ such terms in $\mathcal{S}_{2d+1}$.

Case $\pi(1) \neq 1$, $\pi(r+1) \neq 1, \forall r \in \{1, 2, \ldots, d\}, \pi(2d+1) \neq 1$:
We have $e_3(\pi) = \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X^{(j)}_1, X^{(j)}_{2d+1}\rangle, \langle X^{(j)}_{j+1}, X^{(j)}_{2d+1}\rangle)\}$. There are $(2d+1)! - (d+2) * (2d)! = (d-1) * (2d)!$ such terms in $\mathcal{S}_{2d+1}$.

Thus, combining all the cases, we get

$$
\begin{aligned}
g_1(\mathbf{Z}_1) = {} & \frac{2}{2d+1}\mathbb{E}\{\prod_{j=1}^{d} K^j(\langle Z_1^{(j)}, X_1^{(j)}\rangle, \langle X_2^{(j)}, X_1^{(j)}\rangle)\} + \frac{1}{2d+1}\mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X_1^{(j)}, Z_1^{(j)}\rangle, \langle X_2^{(j)}, Z_1^{(j)}\rangle)\} \\
& + \frac{2d-2}{2d+1}\mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X_1^{(j)}, X_2^{(j)}\rangle, \langle X_3^{(j)}, X_2^{(j)}\rangle)\} + \frac{1}{2d+1}\mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X_{2j-1}^{(j)}, Z_1^{(j)}\rangle, \langle X_{2j}^{(j)}, Z_1^{(j)}\rangle)\} \\
& + \frac{2}{2d+1}\sum_{r=1}^{d}\mathbb{E}\{\prod_{j=1, j\neq r}^{d} K^j(\langle X_{2j-1}^{(j)}, X_{2d+1}^{(j)}\rangle, \langle X_{2j}^{(j)}, X_{2d+1}^{(j)}\rangle) k^r(\langle X_{2r-1}^{(r)}, X_{2d+1}^{(r)}\rangle, \langle Z_1^{(r)}, X_{2d+1}^{(r)}\rangle)\} \\
& - \frac{2}{2d+1}\mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X_1^{(j)}, Z_1^{(j)}\rangle, \langle X_{j+1}^{(j)}, Z_1^{(j)}\rangle)\} - \frac{2}{2d+1}\mathbb{E}\{\prod_{j=1}^{d} K^j(\langle Z_1^{(j)}, X_{2d+1}^{(j)}\rangle, \langle X_{j+1}^{(j)}, X_{2d+1}^{(j)}\rangle)\} \\
& - \frac{2}{2d+1}\sum_{r=1}^{d}\mathbb{E}\{\prod_{j=1, j\neq r}^{d} K^j(\langle X_1^{(j)}, X_{2d+1}^{(j)}\rangle, \langle X_{j+1}^{(j)}, X_{2d+1}^{(j)}\rangle) k^r(\langle X_1^{(r)}, X_{2d+1}^{(r)}\rangle, \langle Z_1^{(r)}, X_{2d+1}^{(r)}\rangle)\} \\
& - \frac{2(d-1)}{2d+1}\mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X_1^{(j)}, X_{2d+1}^{(j)}\rangle, \langle X_{j+1}^{(j)}, X_{2d+1}^{(j)}\rangle)\}.
\end{aligned}
$$

Now, under H_0, $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$ are independent. In particular, we can write

$$
\begin{aligned}
\mathbb{E}\{\prod_{j=1}^{d} K^j(\langle Z_1^{(j)}, X_{2d+1}^{(j)}\rangle, \langle X_{j+1}^{(j)}, X_{2d+1}^{(j)}\rangle)\} &= \mathbb{E}_{\mathbf{X}_{2d+1}}\{\prod_{j=1}^{d}\mathbb{E}\{K^j(\langle Z_1^{(j)}, X_{2d+1}^{(j)}\rangle, \langle X_2^{(j)}, X_{2d+1}^{(j)}\rangle)\}\}, \\
\mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X_1^{(j)}, Z_1^{(j)}\rangle, \langle X_{j+1}^{(j)}, Z_1^{(j)}\rangle)\} &= \prod_{j=1}^{d}\mathbb{E}\{K^j(\langle X_1^{(j)}, Z_1^{(j)}\rangle, \langle X_2^{(j)}, Z_1^{(j)}\rangle)\} \\
&= \prod_{j=1}^{d}\mathbb{E}\{K^j(\langle X_{2j-1}^{(j)}, Z_1^{(j)}\rangle, \langle X_{2j}^{(j)}, Z_1^{(j)}\rangle)\}, \\
\mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X_1^{(j)}, X_2^{(j)}\rangle, \langle X_3^{(j)}, X_2^{(j)}\rangle)\} &= \mathbb{E}\{\prod_{j=1}^{d} K^j(\langle X_1^{(j)}, X_{2d+1}^{(j)}\rangle, \langle X_{j+1}^{(j)}, X_{2d+1}^{(j)}\rangle)\} \text{ and}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\{\prod_{j=1, j\neq r}^{d} K^j(\langle X_{2j-1}^{(j)}, X_{2d+1}^{(j)}\rangle, \langle X_{2j}^{(j)}, X_{2d+1}^{(j)}\rangle) * k^r(\langle X_{2r-1}^{(r)}, X_{2d+1}^{(r)}\rangle, \langle Z_1^{(r)}, X_{2d+1}^{(r)}\rangle)\} \\
&= \mathbb{E}_{\mathbf{X}_{2d+1}}\{\prod_{j=1, j\neq r}\mathbb{E}\{K^j(\langle X_1^{(j)}, X_{2d+1}^{(j)}\rangle, \langle X_{j+1}^{(j)}, X_{2d+1}^{(j)}\rangle) * \mathbb{E}\{k^r(\langle X_1^{(r)}, X_{2d+1}^{(r)}\rangle, \langle Z_1^{(r)}, X_{2d+1}^{(r)}\rangle)\}\}.
\end{aligned}
$$

Thus under H_0, we have $g_1(\boldsymbol{Z}_1) = 0$. So, under H_0, $\hat{\xi}_n$ is a degenerate V-statistic with the order of degeneracy being at least one. It is difficult to show that higher-order degeneracy will not exist in this situation. So, following Theorems C.5, C.6, C.7, and C.9 from the supplementary material of Pfister et al. (2018), we summarize our result as follows.

- Under H_0, $\sigma_1^*(g) = \mathrm{Var}\big(\mathbb{E}\{g(\mathbf{X}_1, \ldots, \mathbf{X}_{2d+1})|\mathbf{X}_1\}\big) = 0$. Therefore, if $\sigma_2^*(g) > 0$, we have $n\hat{\xi}_n \xrightarrow{D} \sum_{i=1}^{\infty}\lambda_i Z_i^2$, where $\sum_{i=1}^{\infty}\lambda_i^2 = \sigma_2^*(g)$ and $\{Z_i\}$ is a sequence of i.i.d. $\mathcal{N}_1(0,1)$ random

variables. If $\sigma_2^*(g) = 0$, then $\mathrm{Var}(n\hat{\xi}_n) = O(n^{-\frac{1}{2}})$ and $\mathbb{E}\{n\hat{\xi}_n\} = \binom{2d+1}{2}\mathbb{E}\{g_2(\mathbf{X}_1, \mathbf{X}_1)\} + O(n^{-1})$, where $g_2(\mathbf{x}, \mathbf{y}) = \mathbb{E}\{g(\mathbf{x}, \mathbf{y}, \mathbf{X}_3, \ldots, \mathbf{X}_{2d+1})\}$. Thus in this scenario, $n\hat{\xi}_n$ becomes a degenerate random variable.

- Under H_1, it is difficult to check whether g is a non-degenerate core function. If $\sigma_1^*(g) > 0$, we have $\sqrt{n}(\hat{\xi}_n - \xi(\mathbb{P})) \xrightarrow{D} \mathcal{N}_1(0, (2d+1)^2\sigma_1^*(g))$. If $\sigma_1^*(g) = 0$, using (6.2), we get $\mathbb{E}\{n(\hat{\xi} - \xi(\mathbb{P}))^2\} = n\mathrm{Var}(\hat{\xi}_n) + n\{\mathrm{Bias}(\hat{\xi}_n)\}^2 = O(\frac{1}{n})$. Hence, in this scenario, $n\mathbb{E}\big(\hat{\xi}_n - \xi(\mathbb{P})\big)^2 \to 0$, as $n \to \infty$.

This completes our proof. ■

Proof pf Corollary 6.1. This result follows from Theorem 6.3. ■

Proof of Theorem 6.4. Notice that if $K^j(x, y) = K_{\sigma_j}(x, y) = \exp\{-|x - y|^2/2\sigma_j^2\}$ for all $j = 1, \ldots, d$, then for any $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \in (0, \infty)^d$, we have

$$\begin{aligned}
&\Big| \prod_{j=1}^{d} K_{\sigma_{1j}}(x^{(j)}, y^{(j)}) - \prod_{j=1}^{d} K_{\sigma_{2j}}(x^{(j)}, y^{(j)}) \Big| \\
&= \Big| \prod_{j=1}^{d} \exp(-|x^{(j)} - y^{(j)}|_2^2/2\sigma_{1j}^2) - \prod_{j=1}^{d} \exp(-|x^{(j)} - y^{(j)}|_2^2/2\sigma_{2j}^2) \Big| \\
&\leq \sum_{j=1}^{d} \sup_{x \in [0,\infty)} \Big| \exp(-x/2\sigma_{1j}^2) - \exp(-x/2\sigma_{2j}^2) \Big|.
\end{aligned}$$

So, $|\hat{\xi}_{\sigma_1,n} - \hat{\xi}_{\sigma_2,n}| \leq 4\sum_{j=1}^{d} \sup_{x \in [0,\infty)} \Big| \exp(-x/2\sigma_{1j}^2) - \exp(-x/2\sigma_{2j}^2) \Big| = 4\sum_{j=1}^{d} \sup_{x \in [0,\infty)} [h_j(x)]^{1/2}$, where $h_j(x) = \Big| \exp(-x/2\sigma_{1j}^2) - \exp(-x/2\sigma_{2j}^2) \Big|^2$. To find the supremum of h_j over the domain $[0, \infty)$, first notice that h_j is differentiable on $(0, \infty)$, and

$$h_j'(x) = -\frac{1}{\sigma_{1j}^2}\exp(-x/\sigma_{1j}^2) - \frac{1}{\sigma_{2j}^2}\exp(-x/\sigma_{2j}^2) + \Big(\frac{1}{\sigma_{1j}^2} + \frac{1}{\sigma_{2j}^2}\Big)\exp\left(-x\Big(\frac{1}{2\sigma_{1j}^2} + \frac{1}{2\sigma_{2j}^2}\Big)\right) = 0$$

has two solutions

$$2\log\{\max\{\sigma_{1j}^2/\sigma_{2j}^2, 1\}\}\sigma_{1j}^2\sigma_{2j}^2/(\sigma_{1j}^2 - \sigma_{2j}^2) \text{ and } 2\log\{\min\{\sigma_{1j}^2/\sigma_{2j}^2, 1\}\}\sigma_{1j}^2\sigma_{2j}^2/(\sigma_{1j}^2 - \sigma_{2j}^2).$$

Since h_j is zero at the origin, we have the maximizer $2\log\{\sigma_{1j}^2/\sigma_{2j}^2\}\sigma_{1j}^2\sigma_{2j}^2/(\sigma_{1j}^2 - \sigma_{2j}^2)$. Thus,

$$\begin{aligned}
&\sup_{x \in [0,\infty)} \Big| \exp(-x/2\sigma_{1j}^2) - \exp(-x/2\sigma_{2j}^2) \Big| \\
&= \Big| \exp\Big\{ -\log\Big(\frac{\sigma_{1j}^2}{\sigma_{2j}^2}\Big)\frac{\sigma_{2j}^2}{\sigma_{1j}^2 - \sigma_{2j}^2} \Big\} - \exp\Big\{ -\log\Big(\frac{\sigma_{1j}^2}{\sigma_{2j}^2}\Big)\frac{\sigma_{1j}^2}{\sigma_{1j}^2 - \sigma_{2j}^2} \Big\} \Big|.
\end{aligned}$$

One can check that as $\sigma_{1j} \to \sigma_{2j}$, this quantity converges to 0. This shows that if $\boldsymbol{\sigma}_{(n)}$ is a sequence of bandwidths converging to $\boldsymbol{\sigma}_0$ in probability, we have $\hat{\xi}_{\boldsymbol{\sigma}_{(n)},n} \xrightarrow{P} \xi_{\boldsymbol{\sigma}_0}(\mathbb{P})$. Since the elements of $\boldsymbol{\sigma}_0$ are positive, $\xi_{\boldsymbol{\sigma}_0}(\mathbb{P})$ is non-negative and takes the value zero if and only if $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$ are independent. This proves the consistency of the test against fixed alternatives. ■

Chapter 7

Concluding Remarks

In this thesis, we have developed some nonparametric methods for high-dimensional and functional data and investigated their theoretical properties under appropriate regularity conditions. By analyzing several simulated and real datasets, we have also amply demonstrated the usefulness of these proposed methods. However, our methods are not above all limitations. During our investigation, we noticed some shortcomings of the proposed methods and also identified some interesting related issues, which can be investigated as future research problems. We briefly discuss them in this final chapter of the thesis. A short summary of our contributions is also given in this chapter.

In Chapter 2, we investigated the high-dimensional behavior of our two-sample tests based on ball divergence, and under appropriate assumptions, proved their consistency both in HDLSS and HDHSS regimes. We also proved their minimax rate optimality and established their consistency even for shrinking alternatives. In this chapter, we considered tests based on different distance functions. In the HDLSS asymptotic regime, while the test based on the ℓ_2 distance can discriminate between two distributions differing in their location or scales, those based on the generalized distance function $\varphi_{h,\psi}$ (see page 13) can also differentiate two distributions differing outside the first two moments if the functions h and ψ are chosen appropriately. If the underlying distributions have light exponential tails, the tests based on ℓ_2 and ℓ_1 distances usually perform better, but if they have heavy polynomial tails, it is better to use tests based on a suitably chosen bounded ψ function. Finding the optimal choice of ψ for a given data set is a challenging problem. It would be helpful if one could develop a method for a data-driven choice of this function.

Analyzing several simulated and real data sets, we have shown that the proposed tests can outperform the start-of-the-art tests in a wide variety of high-dimensional two-sample problems. However, one major problem with these tests is the computation issue. Their computing costs grow up linearly with the dimension d but at a faster rate $O(n^3)$ with the sample size n. So, to make them applicable to large data sets, one needs to come up with scalable versions.

The proposed tests can also be generalized to k-sample problems, where we test for the equality of k multivariate distributions $F_1, F_2, \ldots, F_k$. For a ball with a specified center and radius, the variance of the probability measures of that ball corresponding to these distributions gives

us some idea about the difference among the F_i's around that specified center. The average (or weighted average) of these variances computed for different balls with random centers and radii can be used as a generalized measure of ball divergence. Clearly, it is non-negative, and it takes the value 0 if and only if all the F_i's are identical. A suitable estimate of this measure can be used to construct a k-sample test.

Recently, Pan et al. (2020) used the notion of ball divergence to develop ball covariance, a measure of dependence among several Banach-valued random variables, and proposed a test of independence based on it. This has been discussed in Chapter 6 in the context of functional data. However, the theory presented in Chapter 2 can be used to study the high dimensional behavior of that test, particularly when the sample size increases with the dimension. However, it is not clear whether the test is minimax rate optimal. This can be investigated in future work.

In Chapter 3, we proposed a new measure $\zeta(\mathrm{P})$ for spherical asymmetry of a probability distribution P and constructed a consistent estimator $\hat{\zeta}_n$ of this measure based on data augmentation. We also developed a test of spherical symmetry based on this estimator and studied its large sample behavior when the dimension may or may not grow with the sample size. Recall that the proposed MMD-type measure $\zeta(\mathrm{P})$ was constructed by aggregating the squared differences between $\varphi_1(\mathbf{t})$ and $\varphi_2(\mathbf{t})$, the characteristic functions of the random vector $\mathbf{X}$ and that of its spherically symmetric variant $\mathbf{X}'$, while the probability measure corresponding to $\mathcal{N}_d(\mathbf{0}_d, \frac{1}{d}\mathbf{I}_d)$ was used as the weight function $W(\cdot)$ for aggregation over different choices of $\mathbf{t}$. (see page 40). Though our extensive simulation studies amply demonstrated the superiority of the proposed test over some state-of-the-art methods, its performance may vary depending on the choice of W. One can introduce an additional scale parameter in W or use other choices of W (e.g., Laplace or Cauchy) as well. A suitable data-driven choice of W may lead to further improvement in the performance of the proposed test. We leave these as possible future extensions of our work.

However, the test proposed in Chapter 3 often fails to have satisfactory performance in the HDLSS setup. Moreover, the resampling algorithm needed for calibration of the test increases the computing cost. To take care of these issues, in Chapter 4, we proposed some distribution-free methods for testing the spherical symmetry of a multivariate distribution. We mainly considered sign and runs tests and their modified versions in this thesis. These tests are also based on the idea of data augmentation. Under appropriate regularity conditions, we proved the consistency of these tests in the HDLSS and HDHSS asymptotic regimes and demonstrated their utility using several simulated and real data sets. They outperformed the state-of-the-art methods in a wide variety of high-dimensional examples. In high dimensions, the modified sign and runs tests usually outperform the test proposed in Chapter 3, especially when the dimension is larger than the sample size. But when the sample size is much larger than the dimension, sometimes it is better to use the test based on the MMD-type measure discussed in Chapter 3. Among the modified sign and

runs tests, there is no clear winner, but the former one has been observed to yield higher power in most of our examples. For constructing our modified sign and runs tests, here we used Bonferroni's method for size correction. Instead, one can also use the methods available for controlling the false discovery rate based on p-values (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001) or e-values (Wang & Ramdas, 2022). However, these methods did not make any visible difference in the performance of our proposed tests. Instead of restricting to sign test, one can also consider a more general class of linear rank tests, but finding the optimal score function in a given problem is a matter of concern. Similarly, instead of runs test based on number of runs, one may construct a test based on the length of the longest run as well. However, it had relatively inferior performance in our examples, and that is why we decided not to include it in this thesis.

Recall that our sign and runs tests need the construction of the shortest covering path on the augmented dataset. As we have mentioned before, this is an NP-complete problem, and here, we have used a heuristic method based on Prim's algorithm for this purpose. Instead of the shortest covering path, one can consider other graph-based tests as well. For instance, one can construct the spanning tree with $n-1$ edges having the minimum cost that covers either a data point or its spherically symmetric counterpart. Though there are algorithms (see, e.g., Prim, 1957; Kruskal, 1956) for constructing the minimum spanning tree, finding such a tree that covers n out of $2n$ vertices (each representing one observation in the augmented data set) again turns out to be an NP-complete problem. A heuristic method based on Prim's algorithm (Prim, 1957) can be used there as well, and after constructing the tree, sign statistic can be defined in the same way. This sign statistic will also have the distribution-free property, and its null distribution will match with that of the univariate sign statistic. Currently, we do not have any quantification of this approximation. It will be an interesting problem to find some theoretical results on the approximation factor for this algorithm. This can be explored as a potential research problem. The runs statistic can also be computed using the idea of Friedman & Rafsky (1979), but unfortunately, the resulting test won't be distribution-free in two or higher dimensions. One can also construct a test based on nearest neighbor type coincidences (see, e.g., Henze, 1988; Schilling, 1986; Hall & Tajvidi, 2002; Mondal, Biswas & Ghosh, 2015). However, the resulting test won't be distribution-free, and one needs to use a resampling algorithm for its calibration, which will increase the computing cost.

Since the main focus of Chapter 4 was on a high-dimensional test for spherical symmetry, we did not pay much attention to the usual large sample behavior of the proposed sign test (or linear rank test) and runs test in the classical asymptotic regime. However, from our discussion, it is clear that the asymptotic null distributions of the linear rank statistic and the runs statistic are the same as given by Theorems 4.4 and 4.6. The large sample consistency of the resulting tests can be proved as well. Theorem A4.1 also establishes the Pitman efficiency of the linear rank test, but such a result for the runs test is yet to be derived.

Following our idea based on data augmentation, tests for other types of symmetry can also be constructed. For instance, one can construct a test for general symmetry (i.e., $\boldsymbol{X} \stackrel{D}{=} -\boldsymbol{X}$), angular symmetry (i.e., $\boldsymbol{X}/\|\boldsymbol{X}\| \stackrel{D}{=} -\boldsymbol{X}/\|\boldsymbol{X}\|$), coordinate wise symmetry (i.e., $(X_1, \ldots, X_d) \stackrel{D}{=} (\epsilon_1 X_1, \ldots, \epsilon_d X_d)$ for all $(\epsilon_1, \ldots, \epsilon_d) \in \{-1, 1\}^n$) or any $\mathcal{G}$-symmetry (i.e., $\boldsymbol{X} \stackrel{D}{=} G\boldsymbol{X}$ for all $G \in \mathcal{G}$). In such cases, we can use an MMD-type measure to construct a test statistic, but one needs to properly exploit the exchangeability structure of the observed data and their symmetric variants under the null hypothesis of symmetry to develop an appropriate resampling algorithm for calibration. Graph-based ideas can also be used for constructing distribution-free sign and runs tests, but instead of inner-products, one may need to consider costs based on other suitable functions to properly discriminate between the underlying distribution and its symmetric variant. The theoretical properties of the resulting tests need to be properly investigated as well.

In Chapter 3 and Chapter 4 of this thesis, we mainly considered the case where the null hypothesis specified the center of symmetry (which was taken as the origin $\mathbf{0}_d$). If the null hypothesis does not specify the center, it calls for a test of spherical symmetry about an unknown center. We briefly discussed this issue in Chapter 4, where we described a plug-in method (that estimates the location from the data) and a method based on sample splitting (that uses the differences between pairs of observations). The former one works well when the sample size is large compared to the dimension of the data, but when the dimension is comparable to or larger than the sample size, it often leads to inflated type I errors by the proposed tests. The sample splitting idea helps to take care of this problem, but the loss of observations due to sample splitting reduces the finite sample powers of the resulting tests. We are yet to achieve a satisfactory solution in this regard. Another interesting problem would be the construction of a test for elliptic symmetry of a high-dimensional probability distribution. If the sample size is large compared to the dimension of the data, we can use a plug-in method that estimates the location and the scatter of the underlying distribution for standardization and apply the tests of spherical symmetry on the standardized data, but this idea does not work when the dimension of the data is large compared to the sample size. One needs to come up with an appropriate alternative method to take care of this issue.

In Chapter 5, we proposed a two-sample test for functional data, where the observations $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ on the random functions $X \sim F$ and $Y \sim G$ were modeled as elements of a separable Hilbert space $\mathcal{H}$. We considered linear projections $\langle X_1, f\rangle, \ldots, \langle X_n, f\rangle$ and $\langle Y_1, f\rangle, \ldots, \langle Y_m, f\rangle$ of the observations along a particular direction $f \in \mathcal{H}$ and used a measure based on the Baringhaus-Franz (BF) statistic (see Baringhaus *&* Franz, 2010) to estimate the difference between the two distributions along that direction. These estimated differences were aggregated judiciously over several choices $f \in \mathcal{H}$ to construct the test statistic. This idea of aggregation over several linear projections has been used in the literature for multivariate data as well. For instance, the multivariate Cramer test (Baringhaus *&* Franz, 2004) and the test based on projection averaging

(Kim, Balakrishnan & Wasserman, 2020) use the same idea, where appropriate probability measures on $\mathcal{S}^{d-1}$ are used for aggregation. Using that idea one can think of aggregating the measures by using a probability distribution on the unit ball in $\mathcal{H}$, but such a probability distribution sometimes unnecessarily puts weights on many directions that are orthogonal to the observed data, and as a result, the power of resulting the test goes down. That is why instead of using such a probability distribution, we opted for aggregation based on the empirical distribution function $\frac{1}{2}(\hat{F}_n + \hat{G}_m)$. We derived the limiting distribution of our proposed test statistic and proved the large sample consistency of the permutation test. We also derived a new local asymptotic normality result for functional data and proved that our test is statistically efficient in the Pitman sense. Analyzing several simulated and real data sets, we amply demonstrated the superior performance of our test over several state-of-the-art tests.

Note that the BF statistic comes with an associated function ϕ. In this chapter, we considered several choices of ϕ, which were motivated by the suggestion of Baringhaus & Franz (2010). Based on our empirical experience, we recommend using $\phi(z) = \sqrt{z}$ when the observations on X and those on Y are nearly orthogonal. Otherwise, the use of $\phi(z) = 1 - \exp(-z/2)$ or $\phi(z) = \log(1 + z)$ usually yields better performance, especially in the presence of outliers and extreme observations. However, one may opt for other appropriate choices of ϕ or use a different statistic for computing the difference between two distributions along various projection directions. Similarly, instead of relying on the empirical distribution function $\frac{1}{2}(\hat{F}_n + \hat{G}_m)$, one can use other suitable methods for aggregation.

The proposed test can be generalized to k-sample problems as well, where one needs to use a suitable k-sample criterion for measuring the differences among the one-dimensional linear projections corresponding to different distributions $F_1, F_2, \ldots, F_k$. One can construct a consistent estimate of this measure, repeat it for several linear projections, and aggregate them judiciously to come up with a test statistic for the k-sample test. The large sample behavior of the resulting test can also be investigated using the theory presented in this chapter. However, the empirical performance of the resulting k-sample test may depend on the univariate measure of difference and the aggregation method.

In Chapter 6, we used the projection-based idea to present a general recipe for measuring dependence among multiple random functions and proposed a test for their mutual independence based on that measure. In particular, we considered different linear projections of the components of the observations and used the d-variate Hilbert-Schmidt Independence Criterion (dHSIC) to compute the dependence among the linear projections along those directions. The measures corresponding to different projections were aggregated to come up with the test statistic. This idea of projection averaging was also explored in Lai et al. (2021) for two random functions, where they used a Gaussian weight function for aggregation. Their idea can be extended to multiple

random functions as well, but its computing cost grows faster as the number of variables increases. Our test does not have such an undesirable property and can be conveniently used for multiple random functions. It is also invariant under unitary operations, and because of that, it depends only on the underlying geometry of the Hilbert space and not the space itself. The aCov test does not have this property. Recall that our test statistic involves a kernel function and the associated bandwidth parameter. In this thesis, we used the Gaussian kernel, where the bandwidth was chosen based on median heuristic (Gretton et al., 2012). We proved the large sample consistency of our test even for this data-driven choice of bandwidth. Analyzing several simulated data sets, we have amply demonstrated the usefulness of our test against the state-of-the-art methods.

However, the proper choice of the kernel function and the associated bandwidth still remains an open issue. Our empirical experience suggests that the finite sample performance of the proposed test may depend heavily on the value of the bandwidth parameter. The median heuristic method worked well in most of the examples considered in this thesis, but this may not be the optimal choice in general. So, if one can come up with a suitable data-driven method for selecting the optimal bandwidth, the performance of the proposed test can be improved further. Instead of selecting a single bandwidth, one can also use a multi-scale approach (Sarkar & Ghosh, 2018a), where we implement the test for several choices of the bandwidth parameter and aggregate them judiciously to arrive at the final decision. However, the multi-scale approach increases the computing time substantially. In our examples, there was no visible difference in the performance of the resulting test. That is why we decided not to include it in this thesis.

In the modern machine learning era, we often deal with huge data sets where both sample size and dimension are of the order of billions. Traditional resampling algorithms like permutation or bootstrap (as considered in Chapters 2, 5, and 6) become computationally demanding for such large data sets. To take care of this problem, several approximate algorithms (see, e.g. Politis, Romano & Wolf, 1999; Bickel, Götze & van Zwet, 1997; Kleiner et al., 2014; Sengupta, Volgushev & Shao, 2016) have been proposed in the literature as alternatives to the usual bootstrap. However, these ideas are not directly applicable for the calibration of our tests. Some non-trivial modifications are needed, but even then, the theoretical guarantees of the resulting algorithms are not very clear at this moment. Note that our calibration is based on the permutation principle. So, it would be great if we could come up with computationally efficient alternatives to these permutation methods. Similarly, one may look for a computationally efficient version of the resampling algorithm proposed in Chapter 3. Note that most of the test statistics considered in this thesis are either complete U-statistics or complete V-statistics. Instead, one may consider using incomplete U and V statistics (see Chen & Kato (2019) and Shekhar, Kim & Ramdas (2022)) , which are easily computable and scalable to large data sets. However, these ideas may lead to a loss in statistical efficiency. Therefore, one needs to come up with more suitable algorithms that are capable of handling such

datasets without sacrificing statistical efficiency. One option is to approximate the tail behaviour of the resampling test statistics using saddle point approximation Davison & Hinkley (1988) or large deviation principle Dembo & Zeitouni (2010). These ideas deserve further exploration as future research problems.

Appendix A

Brief Descriptions of Competing Tests

In different chapters of this thesis, we have compared our proposed tests with different state-of-the-art tests available in the literature. Brief descriptions of those tests are given below. Recall that in Chapters 2-4, we considered some hypothesis testing problems for high-dimensional data, whereas in Chapters 5 and 6, some inference problems related to functional data were considered. Here, we arrange the competing tests accordingly.

Chapters 2-4: Tests for high dimensional data

In Chapter 2, we considered the problem of testing the equality of two high-dimensional probability distributions, while in Chapters 3 and 4, we considered the problem of testing the spherical symmetry of a high-dimensional distribution.

Chapter 2: Two-sample test

Suppose that $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ and $\mathcal{Y} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m\}$ are two independent samples from two d-dimensional probability distributions F and G, respectively. Here, we test the null hypothesis $H_0 : F = G$ against the alternative hypothesis $H_1 : F \neq G$.

• **FR test** (Friedman & Rafsky, 1979): Consider an edge weighted complete graph on the vertex set $\mathcal{X} \cup \mathcal{Y}$ with edge weigths being the pairwise Euclidean distances. The FR test rejects H_0 for small values of

$$T_{FR} = 1 + \sum_{i=1}^{n+m-1} \mathbb{M}_i,$$

where $\mathbb{M}_i$ takes the value 1 if the i-th edge of the minimum spanning tree in the complete graph connects two observations from different samples.

• **SHP test** (Biswas, Mukhopadhyay & Ghosh, 2014): Consider the same setup as in the FR test and reject H_0 for small values of

$$T_{SHP} = 1 + \sum_{i=1}^{n+m-1} \mathbb{S}_i$$

where $\mathbb{S}_i$ takes the value 1 if the i-th edge of the shortest Hamiltonian path in the complete graph connects two observations from different samples. This test has the exact distribution-free property.

• **BF test** (Baringhaus & Franz, 2004): It rejects H_0 for large values of

$$T_{BF} = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\boldsymbol{X}_i - \boldsymbol{Y}_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{X}_i - \boldsymbol{X}_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\boldsymbol{Y}_i - \boldsymbol{Y}_j\|.$$

• **BG test** (Biswas & Ghosh, 2014): It rejects H_0 for large values of

$$T_{BG} = \left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\boldsymbol{X}_i - \boldsymbol{Y}_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|\right)^2 + \left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\boldsymbol{X}_i - \boldsymbol{Y}_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\boldsymbol{Y}_i - \boldsymbol{Y}_j\|\right)^2.$$

• **NN test** (Schilling, 1986; Henze, 1988): For a fixed $k < n$, it rejects H_0 for large values

$$T_{NN} = \frac{1}{nk}\left\{\sum_{i=1}^{n}\sum_{r=1}^{k}\mathbb{I}_r(\boldsymbol{X}_i) + \sum_{j=1}^{m}\sum_{r=1}^{k}\mathbb{I}_r(\boldsymbol{Y}_i)\right\},$$

where $\mathbb{I}_r(\boldsymbol{Z})$ takes the value one if $\boldsymbol{Z}$ and its r-th nearest neighbor (in terms of the Euclidean distance) come from the same sample.

• **MMD test** (Gretton et al., 2012): It rejects H_0 for large values of

$$T_{MMD} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(\boldsymbol{X}_i, \boldsymbol{X}_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}K(\boldsymbol{Y}_i, \boldsymbol{Y}_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}K(\boldsymbol{X}_i, \boldsymbol{Y}_j),$$

where $K(.,.)$ is a reproducing kernel of an RKHS, which is also a characteristic kernel. In practice, we take $K(\boldsymbol{X}, \boldsymbol{Y}) = \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{X} - \boldsymbol{Y}\|^2\right\}$ and σ is chosen based on median heuristic.

Chapters 3 and 4: Test of spherical symmetry

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent observations from a d-dimensional probability distribution P. Here, we want to test the null hypothesis (H_0) that P is a spherically symmetric distribution.

• **OT test** (Huang & Sen, 2023): Let $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ be i.i.d. observations from $\mathcal{N}_d(\boldsymbol{0}_d, \mathbf{I}_d)$. If $\boldsymbol{X}_i$ ($i = 1, 2 = \ldots, n$) has the j-th largest Euclidean norm among $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and $\|\boldsymbol{Q}_{(1)}\| < \cdots < \|\boldsymbol{Q}_{(n)}\|$, then the rank vector of $\boldsymbol{X}_i$ is $\boldsymbol{H}_{(j)}$, i.e., $R_n(\boldsymbol{X}_i) = \boldsymbol{Q}_{(j)}$. The signed-rank vector is defined as $\|R_n(\boldsymbol{X}_i)\|\frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|}$ and it rejects H_0 for large values of

$$T_{OT} = \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\|R_n(\boldsymbol{X}_i)\|\frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|}\right\|^2.$$

This test is distribution-free, but the large sample distribution of T_{OT} is used to calibrate this test.

• **DT test** (Diks & Tong, 1999): It rejects H_0 for large values of

$$T_{DT} = \frac{2}{n(n-1)}\sum_{1\le i<j\le n}\exp\left\{-\frac{1}{4\tau^2}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2\right\},$$

where $\tau = 0.25$ is taken in practice, following the suggestion of the authors. However, this choice of τ did not work well for high-dimensional data. So, after discussing it with the authors, we made a small change in the scale and chose $4\tau^2 = (0.25)^2 * \frac{1}{n(n-1)} \sum_{1\le i<j\le n} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2$.

• **PP test** (Fang, Zhu & Bentler, 1993): Let $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d$ be random orthogonal directions in $\mathbb{R}^d$. Then, it rejects H_0 for small values of

$$T_{PP} = \min_{1\le i<j\le d} \left[\frac{1}{n(n-1)} \sum_{1\le l<k\le n} \mathrm{I}[\boldsymbol{\alpha}_i^\top \boldsymbol{X}_l < \boldsymbol{\alpha}_j^\top \boldsymbol{X}_k] \right].$$

However, this test becomes computationally prohibitive in high dimensions. So, we could not use it for the high-dimensional problems considered in Chapter 4.

Tests for functional data

For functional data, observations were modeled as elements of Hilbert spaces. In Chapter 5, we considered the problem of testing the equality of two probability distributions, whereas, in Chapter 6, we considered testing mutual independence among several functional random variables.

Chapter 5: Two-sample test

Let $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots .Y_m\}$ be two sets of independent observations on two functional random variables $X \sim F$ and $Y \sim G$, respectively, which take values in a separable Hilbert space $\mathcal{H}$. Here, we want to test $H_0 : F = G$ against $H_1 : F \neq G$.

• **WD test** (Wynne & Duncan, 2020): It rejects H_0 for large values of

$$T_{WD} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K(X_i, X_j) + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} K(Y_i, Y_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} K(X_i, Y_j),$$

where $K(.,.)$ is the Gaussian kernel as in the MMD test.

• **BD test** (Pan et al., 2018): It rejects H_0 for large values of

$$\begin{aligned} T_{BD} =& \frac{1}{n^2} \sum_{1\le i,j\le n} \left\{ \frac{1}{n} \sum_{k=1}^{n} \delta(X_k, X_j, X_i) - \frac{1}{m} \sum_{k=1}^{m} \delta(Y_k, X_j, X_i) \right\}^2 \\ &+ \frac{1}{m^2} \sum_{1\le i,j\le m} \left\{ \frac{1}{n} \sum_{k=1}^{n} \delta(X_k, Y_j, Y_i) - \frac{1}{m} \sum_{k=1}^{m} \delta(Y_k, Y_j, Y_i) \right\}^2, \end{aligned}$$

where $\delta(s, u, v) = \mathrm{I}\Big[\int \big(s(t) - v(t)\big)^2 \mathrm{d}t \le \int \big(u(t) - v(t)\big)^2 \mathrm{d}t \Big]$.

• **FAD test** (Pomann, Staicu & Ghosh, 2016): It applies the functional principal component analysis on the pooled data and uses the Anderson-Darling's test on the linear transformation of the data projected along major principal component directions to get the p-values. These p-values are combined using the Bonferroni's method to construct the test.

Chapter 6: Test of independence

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent observations on $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(d)})$, where $X^{(i)}$ $(i = 1, 2, \ldots d)$ takes values in a separable Hilbert space $\mathcal{H}_i$. We denote the metric on $\mathcal{H}_i$ by $d_i(\cdot, \cdot)$. We want to test the null hypothesis (H_0) that $X^{(1)}, \ldots, X^{(d)}$ are mutually independent.

• **dCov test** (Lyons, 2013): This test is applicable only when $d = 2$. First define, $a_{ij} = d_1(X_i^{(1)}, X_j^{(1)})$ and $b_{ij} = d_2(X_i^{(2)}, X_j^{(2)})$ for $i, j = 1, \ldots, n$. Let $a_{i\cdot} = \frac{1}{n}\sum_{j=1}^n a_{ij}, b_{i\cdot} = \frac{1}{n}\sum_{j=1}^n b_{ij}$ be the i-th row means, $a_{\cdot j} = \frac{1}{n}\sum_{i=1}^n a_{ij}, b_{\cdot j} = \frac{1}{n}\sum_{i=1}^n b_{ij}$ be the j-th column means and $a_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^n a_{i\cdot}, b_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^n b_{i\cdot}$ be the grand means of the distance matrices $((a_{ij}))$ and $((b_{ij}))$, respectively. Then, it rejects H_0 for large values of

$$T_{\text{dCov}} = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n A_{ij}B_{ij},$$

where $A_{ij} = a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}$ and $B_{ij} = b_{ij} - b_{i\cdot} - b_{\cdot j} + b_{\cdot\cdot}$.

• **aCov test** (Lai et al., 2021): This test is also applicable only when $d = 2$. Given a covariance operator Q, define, $\theta_Q(x, x', x'') = \arccos\left(\frac{\langle Q^{1/2}(x-x''), Q^{1/2}(x'-x'')\rangle}{\|Q^{1/2}(x-x'')\|\cdot\|Q^{1/2}(x'-x'')\|}\right)$. Now let $a_{ijk} = \theta_Q(X_i^{(1)}, X_j^{(1)}, X_k^{(1)})$ and $b_{ijk} = \theta_Q(X_i^{(2)}, X_j^{(2)}, X_k^{(2)})$ for $i, j, k = 1, \ldots, n$. It rejects H_0 for large values of

$$T_{\text{aCov}} = \frac{1}{n^3}\sum_{1\le i,j,k\le n} A_{ijk}B_{ijk},$$

where $A_{ijk} = a_{ijk} - \frac{1}{n}\sum_{j=1}^n a_{ijk} - \frac{1}{n}\sum_{i=1}^n a_{ijk} + \frac{1}{n^2}\sum_{i,j=1}^n a_{ijk}$ and $B_{ijk} = b_{ijk} - \frac{1}{n}\sum_{j=1}^n b_{ijk} - \frac{1}{n}\sum_{i=1}^n b_{ijk} + \frac{1}{n^2}\sum_{i,j=1}^n b_{ijk}$.

Here, the raw data need to be transformed using a system of basis functions. We used two different choices of basis functions, the Fourier basis and the spline basis. The corresponding tests are referred to as aCov_1 and aCov_2, respectively.

• **bCov test** (Pan et al., 2020): This test is applicable even when $d \ge 2$. First define, $\delta_{ij,k}^{(\ell)} = \text{I}[d_l(X_i^{(\ell)}, X_k^{(\ell)}) \le d_l(X_i^{(\ell)}, X_j^{(\ell)})]$ for each $i, j, k = 1, \ldots, n$ and $\ell = 1, \ldots, d$. It rejects H_0 for large values of

$$T_{\text{bCov}} = \frac{1}{n^2}\sum_{1\le i,j\le n}\left(\Delta_{ij} - \prod_{\ell=1}^d \Delta_{ij}^{(\ell)}\right)^2,$$

where $\Delta_{ij} = \frac{1}{n}\sum_{k=1}^n\prod_{\ell=1}^d \delta_{ij,k}^{(\ell)}$ and $\Delta_{ij}^{(\ell)} = \frac{1}{n}\sum_{k=1}^n \delta_{ij,k}^{(\ell)}$ for each $i, j = 1, \ldots, n$ and $l = 1, \ldots, d$.

Bibliography

Ahn, J., Marron, J., Muller, K. M. and Chi, Y.-Y. (2007). 'The high-dimension, low-sample-size geometric representation holds under mild conditions.' *Biometrika*, **94**, 760–766.

Albisetti, I., Balabdaoui, F. and Holzmann, H. (2020). 'Testing for spherical and elliptical symmetry.' *J. Multivariate Anal.*, **180**, 104667.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.' *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 6745–6750.

Aoshima, M. and Yata, K. (2018). 'Two-sample tests for high-dimension, strongly spiked eigenvalue models.' *Statistica Sinica*, **28**, 43–62.

Aslan, B. and Zech, G. (2005). 'New test for the multivariate two-sample problem based on the concept of minimum energy.' *J. Stat. Comput. Simul.*, **75**, 109–119.

Bai, Z. and Saranadasa, H. (1996). 'Effect of high dimension: by an example of a two sample problem.' *Statist. Sinica*, **6**, 311–329.

Banerjee, B. (2024). 'Testing distributional equality for functional random variables.' *J. Multivariate Anal.*, **203**, 105318.

Banerjee, B. and Ghosh, A. K. (2022). 'Test of independence for Hilbertian random variables.' *Stat*, **11**, e474, 12.

Banerjee, B. and Ghosh, A. K. (2024a). 'A consistent test of spherical symmetry for multivariate and high-dimensional data via data augmentation.' *arXiv preprint arXiv:2403.12491*.

Banerjee, B. and Ghosh, A. K. (2024b). 'Exact distribution-free tests of spherical symmetry applicable to high dimensional data.' *arXiv preprint arXiv:2412.05608*.

Banerjee, B. and Ghosh, A. K. (2025). 'On high dimensional behaviour of some two-sample tests based on ball divergence.' *Statist. Sinica*, **35**, (To appear).

Baraud, Y. (2002). 'Non-asymptotic minimax rates of testing in signal detection.' *Bernoulli*, **8**, 577–606.

Baringhaus, L. (1991). 'Testing for spherical symmetry of a multivariate distribution.' *Ann. Statist.*, **19**, 899–917.

Baringhaus, L. and Franz, C. (2004). 'On a new multivariate two-sample test.' *J. Multivariate Anal.*, **88**, 190–206.

Baringhaus, L. and Franz, C. (2010). 'Rigid motion invariant two-sample tests.' *Statist. Sinica*, **20**, 1333–1361.

Benjamini, Y. and Hochberg, Y. (1995). 'Controlling the false discovery rate: a practical and powerful approach to multiple testing.' *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **57**, 289–300.

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

Beran, R., Bilodeau, M. and de Micheaux, P. L. (2007). 'Nonparametric tests of independence between random vectors.' *J. Multivariate Anal.*, **98**, 1805–1824.

Bhattacharya, B. B. (2019). 'A general asymptotic framework for distribution-free graph-based two-sample tests.' *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **81**, 575–602.

Bickel, P. J., Götze, F. and van Zwet, W. R. (1997). 'Resampling fewer than n observations: gains, losses, and remedies for losses.' *Statist. Sinica*, **7**, 1–31.

Biswas, M. and Ghosh, A. K. (2014). 'A nonparametric two-sample test applicable to high dimensional data.' *J. Multivariate Anal.*, **123**, 160–171.

Biswas, M., Mukhopadhyay, M. and Ghosh, A. K. (2014). 'A distribution-free two-sample run test applicable to high-dimensional data.' *Biometrika*, **101**, 913–926.

Biswas, M., Mukhopadhyay, M. and Ghosh, A. K. (2015). 'On some exact distribution-free one-sample tests for high dimension low sample size data.' *Statist. Sinica*, **25**, 1421–1435.

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

Blomqvist, N. (1950). 'On a measure of dependence between two random variables.' *Ann. Math. Statist.*, **21**, 593–600.

Brown, B. M. (1971). 'Martingale central limit theorems.' *Ann. Math. Statist.*, **42**, 59–66.

Cai, T., Liu, W. and Xia, Y. (2013). 'Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings.' *J. Amer. Statist. Assoc.*, **108**, 265–277.

Chakraborty, S. and Zhang, X. (2019). 'Distance metrics for measuring joint dependence with application to causal inference.' *J. Amer. Statist. Assoc.*, **114**, 1638–1650.

Chaudhuri, P. (1996). 'On a geometric notion of quantiles for multivariate data.' *J. Amer. Statist. Assoc.*, **91**, 862–872.

Chaudhuri, P. and Sengupta, D. (1993). 'Sign tests in multidimension: inference based on the geometry of the data cloud.' *J. Amer. Statist. Assoc.*, **88**, 1363–1370.

Chen, S. X. and Qin, Y.-L. (2010). 'A two-sample test for high-dimensional data with applications to gene-set testing.' *Ann. Statist.*, **38**, 808–835.

Chen, X. and Kato, K. (2019). 'Randomized incomplete U-statistics in high dimensions.' *Ann. Statist.*, **47**, 3127–3156.

Chen, Y., Hao, Y., Rakthanmanon, T., Zakaria, J., Hu, B. and Keogh, E. (2015). 'A general framework for never-ending learning from time series streams.' *Data Min. Knowl. Discov.*, **29**, 1622–1664.

Chmielewski, M. A. (1981). 'Elliptically symmetric distributions: a review and bibliography.' *Internat. Statist. Rev.*, **49**, 67–74.

Christiansen, B. (2021). 'The blessing of dimensionality for the analysis of climate data.' *Nonlinear Processes in Geophysics*, **28**, 409–422.

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

Davison, A. C. and Hinkley, D. V. (1988). 'Saddlepoint approximations in resampling methods.' *Biometrika*, **75**, 417–431.

Dembo, A. and Zeitouni, O. (2010). *Large deviations techniques and applications*, vol. 38. Springer-Verlag, Berlin.

Diks, C. and Tong, H. (1999). 'A test for symmetries of multivariate probability distributions.' *Biometrika*, **86**, 605–614.

Ding, X. (2020). 'Some sphericity tests for high dimensional data based on ratio of the traces of sample covariance matrices.' *Statist. Probab. Lett.*, **156**, 108613.

Dutta, S., Ghosh, A. K. and Chaudhuri, P. (2011). 'Some intriguing properties of Tukey's half-space depth.' *Bernoulli*, **17**, 1420–1434.

Dutta, S., Sarkar, S. and Ghosh, A. K. (2016). 'Multi-scale classification using localized spatial depth.' *J. Mach. Learn. Res.*, **17**, 218, 30.

Fan, Y., de Micheaux, P. L., Penev, S. and Salopek, D. (2017). 'Multivariate nonparametric test of independence.' *J. Multivariate Anal.*, **153**, 189–210.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Monograph on Statistics and Applied Probability, Chapman and Hall/CRC Press, London.

Fang, K. T., Zhu, L. X. and Bentler, P. M. (1993). 'A necessary test of goodness of fit for sphericity.' *J. Multivariate Anal.*, **45**, 34–55.

Feng, L. and Liu, B. (2017). 'High-dimensional rank tests for sphericity.' *J. Multivariate Anal.*, **155**, 217–233.

Ferraty, F. and Romain, Y. (2011). *The Oxford Handbook of Functional Data Analaysis*. Oxford University Press, Oxford.

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

Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (2018). *Shrinkage Estimation.* Springer Series in Statistics, Springer, Cham.

Friedman, J. H. and Rafsky, L. C. (1979). 'Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests.' *Ann. Statist.*, **7**, 697–717.

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

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

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability.* A Series of Books in the Mathematical Sciences, W. H. Freeman and Co., San Francisco, California.

Ghosh, A. K. and Biswas, M. (2016). 'Distribution-free high-dimensional two-sample tests based on discriminating hyperplanes.' *TEST*, **25**, 525–547.

Ghosh, A. K. and Chaudhuri, P. (2005). 'On maximum depth and related classifiers.' *Scand. J. Statist.*, **32**, 327–350.

Gibbons, J. D. and Chakraborti, S. (2011). *Nonparametric Statistical Inference.* CRC Press, Boca Raton, Florida.

Gieser, P. W. and Randles, R. H. (1997). 'A nonparametric test of independence between two vectors.' *J. Amer. Statist. Assoc.*, **92**, 561–567.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2011). 'Penalized functional regression.' *J. Comput. Graph. Statist.*, **20**, 830–851.

Goldsmith, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2012). 'Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements.' *J. R. Stat. Soc. Ser. C. Appl. Stat.*, **61**, 453–469.

Gretton, A. (2015). 'A simpler condition for consistency of a kernel independence test.' *arXiv preprint arXiv:1501.06103.*

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012). 'A kernel two-sample test.' *J. Mach. Learn. Res.*, **13**, 723–773.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B. and Smola, A. (2007). 'A kernel statistical test of independence.' In *Adv. Neural Inf. Process. Syst.*, vol. 20. Curran Associates, Inc., Newry, Northern Ireland, UK.

Gretton, A. and Györfi, L. (2010). 'Consistent nonparametric tests of independence.' *J. Mach. Learn. Res.*, **11**, 1391–1423.

Gupta, A. and Song, D. (1997). 'Lp-norm spherical distribution.' *J. Stat. Plan. Inference*, **60**, 241–260.

Hájek, J., Sidák, Z. e. and Sen, P. K. (1999). *Theory of Rank Tests*. Probability and Mathematical Statistics, Academic Press, Inc., San Diego, California.

Hall, P., Marron, J. S. and Neeman, A. (2005). 'Geometric representation of high dimension, low sample size data.' *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 427–444.

Hall, P. and Tajvidi, N. (2002). 'Permutation tests for equality of distributions in high-dimensional settings.' *Biometrika*, **89**, 359–374.

Hall, P. and Van Keilegom, I. (2007). 'Two-sample tests in functional data analysis starting from discrete data.' *Statist. Sinica*, 1511–1531.

Heck, D., Knapp, J., Capdevielle, J. N., Schatz, G. and Thouw, T. (1998). 'CORSIKA: A Monte Carlo code to simulate extensive air showers.' Tech. rep. 51.02.03; LK 01; Wissenschaftliche Berichte, FZKA-6019.

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

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

Henze, N. (1988). 'A multivariate two-sample test based on the number of nearest neighbor type coincidences.' *Ann. Statist.*, **16**, 772–783.

Henze, N., Hlávka, Z. and Meintanis, S. G. (2014). 'Testing for spherical symmetry via the empirical characteristic function.' *Statistics*, **48**, 1282–1296.

Hoeffding, W. (1948). 'A non-parametric test of independence.' *Ann. Math. Statist.*, **19**, 546–557.

Hollander, M., Wolfe, D. A. and Chicken, E. (2014). *Nonparametric Statistical Methods*. John Wiley & Sons, Inc., Hoboken, NJ.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, Ltd., Chichester.

Huang, Z. and Sen, B. (2023). 'Multivariate symmetry: distribution-free testing via optimal transport.' *arXiv preprint arXiv:2305.01839*.

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

John, S. (1972). 'The distribution of a statistic used for testing sphericity of normal distributions.' *Biometrika*, **59**, 169–173.

Johnstone, I. M. (2001). 'On the distribution of the largest eigenvalue in principal components analysis.' *Ann. Statist.*, **29**, 295–327.

Jörnsten, R. (2004). 'Clustering and classification based on the l1 data depth.' *J. Multivariate Anal.*, **90**, 67–89.

Jung, S. and Marron, J. S. (2009). 'PCA consistency in high dimension, low sample size context.' *Ann. Statist.*, **37**, 4104–4130.

Kim, I. (2021). 'Comparing a large number of multivariate distributions.' *Bernoulli*, **27**, 419–441.

Kim, I., Balakrishnan, S. and Wasserman, L. (2020). 'Robust multivariate nonparametric tests via projection averaging.' *Ann. Statist.*, **48**, 3417–3441.

Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M. I. (2014). 'A scalable bootstrap for massive data.' *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **76**, 795–816.

Koltchinskii, V. I. and Li, L. (1998). 'Testing for spherical symmetry of a multivariate distribution.' *J. Multivariate Anal.*, **65**, 228–244.

Kruskal, J. B. (1956). 'On the shortest spanning subtree of a graph and the traveling salesman problem.' *Proc. Amer. Math. Soc.*, **7**, 48–50.

Kussul, E. and Baidyk, T. (2004). 'Improved method of handwritten digit recognition tested on mnist database.' *Image and Vision Computing*, **22**, 971–981.

Lai, T., Zhang, Z., Wang, Y. and Kong, L. (2021). 'Testing independence of functional variables by angle covariance.' *J. Multivariate Anal.*, **182**, 104711.

Lee, A. J. (1990). *U-Statistics: Theory and Practice*. Statistics: Textbooks and Monographs, Marcel Dekker, Inc., New York.

Lehmann, E. L. (2012). 'On the history and use of some standard statistical models.' *Selected Works of EL Lehmann (Edited by J. Rozo)*, 1019–1031.

Lehmann, E. L. and Romano, J. P. (2021). *Testing Statistical Hypotheses*. Springer, Cham.

Li, J. and Chen, S. X. (2012). 'Two sample tests for high-dimensional covariance matrices.' *Ann. Statist.*, **40**, 908–940.

Li, J., Cuesta-Albertos, J. A. and Liu, R. Y. (2012). 'DD-classifier: Nonparametric classification procedure based on DD-plot.' *J. Amer. Statist. Assoc.*, **107**, 737–753.

Liang, J., Fang, K.-T. and Hickernell, F. J. (2008). 'Some necessary uniform tests for spherical symmetry.' *Ann. Inst. Statist. Math.*, **60**, 679–696.

Liu, Z. and Modarres, R. (2011). 'A triangle test for equality of distribution functions in high dimensions.' *J. Nonparametr. Stat.*, **23**, 605–615.

Lyons, R. (2013). 'Distance covariance in metric spaces.' *Ann. Probab.*, **41**, 3284–3305.

Maa, J.-F., Pearl, D. K. and Bartoszyński, R. (1996). 'Reducing multidimensional two-sample data to one-dimensional interpoint comparisons.' *Ann. Statist.*, **24**, 1069–1074.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate analysis.* Academic Press, New York.

Massart, P. (1990). 'The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality.' *Ann. Probab.*, **18**, 1269–1283.

Miao, R., Zhang, X. and Wong, R. K. W. (2023). 'A wavelet-based independence test for functional data with an application to MEG functional connectivity.' *J. Amer. Statist. Assoc.*, **118**, 1876–1889.

Mondal, P. K., Biswas, M. and Ghosh, A. K. (2015). 'On high dimensional two-sample tests based on nearest neighbors.' *J. Multivariate Anal.*, **141**, 168–178.

Nelsen, R. B. (1996). 'Nonparametric measures of multivariate association.' *Distributions with Fixed Marginals and Related Topics, Lecture Notes-Monograph Series*, **28**, 223–232.

Newton, M. A. (2009). 'Introducing the discussion paper by Székely and Rizzo.' *Ann. Appl. Statist.*, **3**, 1233–1235.

Pan, W., Tian, Y., Wang, X. and Zhang, H. (2018). 'Ball divergence: nonparametric two sample test.' *Ann. Statist.*, **46**, 1109–1137.

Pan, W., Wang, X., Zhang, H., Zhu, H. and Zhu, J. (2020). 'Ball covariance: a generic measure of dependence in Banach space.' *J. Amer. Statist. Assoc.*, **115**, 307–317.

Pfister, N., Bühlmann, P., Schölkopf, B. and Peters, J. (2018). 'Kernel-based tests for joint independence.' *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **80**, 5–31.

Póczos, B., Ghahramani, Z. and Schneider, J. (2012). 'Copula-based kernel dependency measures.' In *Proceedings of the 29th International Conference on Machine Learning*, 1635–1642. Omnipress, Madison, Wisconsin, USA.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling.* Springer Series in Statistics, Springer-Verlag, New York.

Pomann, G.-M., Staicu, A.-M. and Ghosh, S. (2016). 'A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis.' *J. Roy. Statist. Soc. Ser. C*, **65**, 395–414.

Prim, R. C. (1957). 'Shortest connection networks and some generalizations.' *Bell System Tech. J.*, **36**, 1389–1401.

Qiu, Z., Chen, J. and Zhang, J.-T. (2021). 'Two-sample tests for multivariate functional data with applications.' *Comput. Stat. Data Anal.*, **157**, 107160.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer-Verlag, New York.

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

Randles, R. H. (1989). 'A distribution-free multivariate sign test based on interdirections.' *J. Amer. Statist. Assoc.*, **84**, 1045–1050.

Ratcliffe, B. L., Ness, M. K., Johnston, K. V. and Sen, B. (2020). 'Tracing the assembly of the Milky Way's disk through abundance clustering.' *The Astrophysical Journal*, **900**, 165.

Rawat, R. and Sitaram, A. (2000). 'Injectivity sets for spherical means on $\mathbb{R}^n$ and on symmetric spaces and on symmetric spaces.' *J. Fourier Anal. Appl.*, **6**, 343–348.

Rosenbaum, P. R. (2005). 'An exact distribution-free test comparing two multivariate distributions based on adjacency.' *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 515–530.

Rousseeuw, P. J. and Driessen, K. V. (1999). 'A fast algorithm for the minimum covariance determinant estimator.' *Technometrics*, **41**, 212–223.

Roy, A., Das, K., Sarkar, S. and Ghosh, A. K. (2021). 'Tests of mutual independence among several random vectors using univariate and multivariate ranks of nearest neighbours.' *J. Stat. Comput. Simul.*, **91**, 1890–1906.

Roy, A. and Ghosh, A. K. (2020). 'Some tests of independence based on maximum mean discrepancy and ranks of nearest neighbors.' *Statistics & Probability Letters*, **164**, 108793.

Roy, A., Ghosh, A. K., Goswami, A. and Murthy, C. A. (2022). 'Some new copula based distribution-free tests of independence among several random variables.' *Sankhya Ser. A*, **84**, 556–596.

Roy, A., Sarkar, S., Ghosh, A. K. and Goswami, A. (2020). 'On some consistent tests of mutual independence among several random vectors of arbitrary dimensions.' *Stat. Comput.*, **30**, 1707–1723.

Sarkar, S., Biswas, R. and Ghosh, A. K. (2020). 'On some graph-based two-sample tests for high dimension, low sample size data.' *Mach. Learn.*, **109**, 279–306.

Sarkar, S. and Ghosh, A. K. (2018a). 'On some high-dimensional two-sample tests based on averages of inter-point distances.' *Stat*, **7**, e187, 16.

Sarkar, S. and Ghosh, A. K. (2018b). 'Some multivariate tests of independence based on ranks of nearest neighbors.' *Technometrics*, **60**, 101–111.

Sarkar, S. and Ghosh, A. K. (2020). 'On perfect clustering of high dimension, low sample size data.' *IEEE Trans. Pattern Anal. Mach. Intell.*, **42**, 2257–2272.

Schilling, M. F. (1986). 'Multivariate two-sample tests based on nearest neighbors.' *J. Amer. Statist. Assoc.*, **81**, 799–806.

Schoonover, J. R., Marx, R. and Zhang, S. L. (2003). 'Multivariate curve resolution in the analysis of vibrational spectroscopy data files.' *Appl. Spectrosc.*, **57**, 154A–170A.

Schweizer, B. and Wolff, E. F. (1981). 'On nonparametric measures of dependence for random variables.' *Ann. Statist.*, **9**, 879–885.

Sengupta, S., Volgushev, S. and Shao, X. (2016). 'A subsampled double bootstrap for massive data.' *J. Amer. Statist. Assoc.*, **111**, 1222–1232.

Shao, J. (2003). *Mathematical Statistics*. Springer-Verlag, New York.

Shekhar, S., Kim, I. and Ramdas, A. (2022). 'A permutation-free kernel two-sample test.' In *Advances in Neural Information Processing Systems*, vol. 35, 18168–18180. Curran Associates, Inc.

Smith, P. J. (1977). 'A nonparametric test for bivariate circular symmetry based on the empirical CDF.' *Comm. Statist. Theory Methods*, **6**, 209–220.

Srivastava, R., Li, P. and Ruppert, D. (2016). 'RAPTT: an exact two-sample test in high dimensions using random projections.' *J. Comput. Graph. Statist.*, **25**, 954–970.

Székely, G. J. and Rizzo, M. L. (2004). 'Testing for equal distributions in high dimension.' *InterStat*, **5**, 1249–1272.

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

Taskinen, S., Kankainen, A. and Oja, H. (2003). 'Sign test of independence between two random vectors.' *Statist. Probab. Lett.*, **62**, 9–21.

Taskinen, S., Oja, H. and Randles, R. H. (2005). 'Multivariate nonparametric tests of independence.' *J. Amer. Statist. Assoc.*, **100**, 916–925.

Tsukada, S.-I. (2019). 'High dimensional two-sample test based on the inter-point distance.' *Comput. Statist.*, **34**, 599–615.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. (Revised and extended from the 2004 French original, Translated by Vladimir Zaiats).

Úbeda-Flores, M. (2005). 'Multivariate versions of blomqvist's beta and spearman's footrule.' *Ann. Inst. Statist. Math.*, **57**, 781–788.

Van Aelst, S. and Rousseeuw, P. (2009). 'Minimum volume ellipsoid.' *WIREs Computational Statistics*, **1**, 71–82.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Van der Vaart, A. W. and Wellner, J. (2013). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

Wainwright, M. J. (2019). *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, Cambridge.

Wang, R. and Ramdas, A. (2022). 'False discovery rate control with e-values.' *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **84**, 822–852.

Wei, S., Lee, C., Wichers, L. and Marron, J. S. (2016). 'Direction-projection-permutation for high-dimensional hypothesis tests.' *J. Comput. Graph. Statist.*, **25**, 549–569.

Wheeden, R. L. and Zygmund, A. (1977). *Measure and Integral: An Introduction to Real Analysis*. Marcel Dekker Inc., New York.

Wynne, G. and Duncan, A. B. (2020). 'A kernel two-sample test for functional data.' *arXiv preprint arXiv:2008.11095*.

Yata, K. and Aoshima, M. (2012). 'Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations.' *J. Multivariate Anal.*, **105**, 193–215.

Yata, K. and Aoshima, M. (2020). 'Geometric consistency of principal component scores for high-dimensional mixture models and its application.' *Scand. J. of Statist.*, **47**, 899–921.

Yushkevich, P., Pizer, S. M., Joshi, S. and Marron, J. S. (2001). 'Intuitive, localized analysis of shape variability.' In *Information Processing in Medical Imaging*, 402–408. Springer, Heidelberg.

Zhang, C., Peng, H. and Zhang, J.-T. (2010). 'Two samples tests for functional data.' *Commun. Stat. Theory Methods*, **39**, 559–578.

Zhu, L., Xu, K., Li, R. and Zhong, W. (2017). 'Projection correlation between two random vectors.' *Biometrika*, **104**, 829–843.

Zou, C., Peng, L., Feng, L. and Wang, Z. (2014). 'Multivariate sign-based high-dimensional tests for sphericity.' *Biometrika*, **101**, 229–236.