Some Distribution-free Two-Sample Tests Applicable to High Dimension, Low Sample Size Data

MUNMUN BISWAS

Indian Statistical Institute, Kolkata
December, 2015

MUNMUN BISWAS

Thesis submitted to the Indian Statistical Institute
in partial fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy.
December, 2015

Thesis Advisor : Dr. Anil K. Ghosh

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

To my mother, friends and teachers

Acknowledgement

Individual effort can only flourish in a fertile land of societal interactions. Knowledge of any kind, in a same sense, is the fruit of a collective play. This is an attempt to acknowledge the collective process in my research work.

First, I would like to express my sincere gratitude to my supervisor Dr. Anil K. Ghosh. Without his constant academic guidance, this thesis would not appear in its present shape. His academic excellence as well as continuous effort helped me to learn and construct my research findings. The constant cooperation and friendly interactions I had with him, provided me with lots of impetus to survive throughout my tenure as a researcher.

Besides my advisor, I would like to thank my co-authors Minerva Mukherjee and Prony Kanti Mondal. It was a pleasant experience to work with them. I am also thankful to all the anonymous reviewers, who carefully read the manuscripts submitted to various journals during the tenure of my research. Their insightful comments and helpful suggestions encouraged me to widen the scope of my research and improve the quality of this work significantly.

I am indebted to all my teachers in Indian Statistical Institute, Kolkata, from whom I learnt a lot throughout the last seven years, during my tenure as an M.Stat student first and then as a Ph.D scholar. I am greatly thankful to Prof. Probal Chaudhuri for his academic influence on me. I did one reading course on analysis with Late Prof. Samesh Chandra Bagchi and was profoundly touched by his embracing personality. Also, I am grateful to Prof. Alok Goswami, Prof. Arup Bose, Prof. Gopal K. Basak of Stat. Math. Unit of Indian Statistical Institute, Kolkata, for teaching me various courses during the first two years of my research tenure and also introducing me to many frontline areas of research. I am thankful to the Indian Statistical Institute for providing me with excellent infrastructure to carry out my research work. I would also like to thank various administrative departments of the Indian Statistical Institute for helping me to resolve various administrative matters. Special thanks to the Office of the Stat. Math. Unit for being extremely helpful in various ways.

I express my deepest respect to my mother, who struggled a lot to raise me as a single mother. She unknowingly influenced me in taking hard challenges in life, by potraying her very own journey. My decade long friend Ratul, with whom I share most of my ideological views, was in constant assistance in blooming up my abilities. Words will be insufficient to worth these two comradeships.

I would like to thank all my Stat. Math. Unit friends and co-researchers, who provided me with a comfortable existance in my fourth floor office. Special thanks to Subhajit da and then Anirvan da for guiding me in several academic and technical issues. I spent a great quality time with Moumita di, Sudipta da, Suman da, Samyadeb, Amit, Some Shankar, Sumana, Srijan in organizing various inter-disciplinary and socially significant discussions in ISI, Kolkata campus, during my tenure as a Ph.D student. All these hours together have added colors to my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advancement of data acquisition technologies and computing resources have greatly facilitated the analysis of massive data sets in various fields of sciences. Researchers from different disciplines rigorously investigate these data sets to extract useful information for new scientific discoveries. Many of these data sets contain large number of features but small number of observations. For instance, in the fields of chemometrics (see e.g., Schoonover et al. (2003)), medical image analysis (see e.g., Yushkevich et al. (2001)) and microarray gene expression data analysis (see e.g., Eisen and Brown (1999), Alter et al. (2000)), we often deal with data of dimensions higher than several thousands but sample sizes of the order of a few hundreds or even less. Such high dimension, low sample size (HDLSS) data present a substantial challenge to the statistics community. Many well known classical multivariate methods cannot be used in such situations. For example, because of the singularity of the estimated pooled dispersion matrix, the classical Hotelling's T^2 statistic (see e.g., Anderson (2003)) cannot be used for two-sample test when the dimension of the data exceeds the combined sample size. Over the last few years, researchers are getting more interested in developing statistical methods that are applicable to HDLSS data. In this thesis, we develop some nonparametric methods that can be used for high dimensional two-sample problems involving two independent samples as well as those involving matched pair data.

In a two-sample testing problem, one usually tests the equality of two d-dimensional probability distributions F and G based on two sets of independent observations

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}$ from F and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$ from G. This problem is well investigated in the literature, and several parametric and nonparametric tests are available for it.

Parametric methods assume a common parametric form for F and G, where we test the equality of the parameter values (which could be scalar or finite dimensional vector valued) in two distributions. For instance, if F and G are assumed to be normal (Gaussian) with a common but unknown dispersion, one uses the Fisher's t statistic (when $d = 1$) or the Hotelling's T^2 statistic (when $d > 1$) to test the equality of their locations (see e.g., Mardia et al. (1979); Anderson (2003)). Though these tests have several optimality properties for data having normal distributions, they are not robust against outliers and can mislead our inference if the underlying distributions are far from being normal. Since the performance of parametric methods largely depends on the validity of underlying model assumptions, nonparametric methods are often preferred because of their flexibility and robustness.

In the univariate set up, rank based nonparametric tests like the Wilcoxon-Mann-Whitney test, the Kolmogorov-Smirnov maximum deviation test and the Wald-Wolfowitz run test (see e.g., Hollander and Wolfe (1999); Gibbons and Chakraborti (2003)) are often used. These tests are distribution-free, and they outperform the Fisher's t test for a wide variety of non-Gausssian distributions. The Wilcoxon-Mann-Whitney test is used to test the null hypothesis $H_0 : F = G$ when alternative hypothesis H_A suggests a stochastic ordering between F and G. However, the Kolmogorov-Smirnov test and the Wald-Wolfowitz run test are used for the general alternative $H_A : F \neq G$.

Several nonparametric tests are available for the multivariate two-sample problem as well. If we assume a location model for F and G (i.e. $F(\mathbf{x}) = G(\mathbf{x} - \boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \mathbb{R}^d$ and all $\mathbf{x} \in \mathbb{R}^d$), it leads to a two-sample location problem, where we test the equality of the locations of F and G. Perhaps the most simplest among the nonparametric tests for the multivariate two-sample location problem are those based on coordinate-wise signs and ranks (see e.g., Puri and Sen (1971)). Randles and Peters (1990) developed two-sample location tests based on interdirections. Möttönen and Oja (1995) and Choi and Marden (1997) used spatial sign and ranks to develop two-sample location tests for multivariate data. Hettmansperger and Oja (1994) and Hettmansperger et al. (1998) also developed multivariate sign and rank tests, which can be used for two-sample and multisample location problems. Some good reviews of these tests can be found in Marden (1999), Oja and Randles (2004) and Oja (2010). However, most of these above mentioned multivariate tests including the Hotelling's T^2 test perform poorly for high dimensional data, and none of them can be used when the dimension exceeds the combined sample size $n = n_1 + n_2$. In such cases, one can use the Hotelling's T^2 statistic based on the Moore-Penrose generalized inverse of the estimated pooled dispersion matrix, but it usually leads to poor performance in high dimensions (see e.g. Bickel and Levina (2004)). One should also note that unlike univariate nonparametric methods, none of these multivariate tests are distribution-free in finite sample situations. In these cases, one either uses the test based on the large sample distribution of the test statistic or the conditional test based on the permutation principle.

Mardia (1967) was the first to propose a distribution-free test for the bivariate location problem, but no distribution-free generalization of this test is available for $d > 2$. Liu and Singh (1993) used the notion of simplicial depth to develop two separate distribution-free tests for two-sample location and scale problems. Rousson (2002) proposed a distribution-free test based on data depth and principle component direction, which is applicable to two-sample location scale model. But none of these depth based tests can be used when the dimension is larger than the sample size. Recently, several Hotelling's T^2 type two-sample location tests have been proposed in the literature, which can be used in high dimension low sample size situations (see e.g., Bai and Saranadasa (1996); Srivastava and Du (2008); Chen and Qin (2010); Srivastava et al. (2013); Park and Ayyala (2013)). These tests are based on the asymptotic distribution of the test statistics, where the dimension d is assumed to grow with the sample size n. Most of these tests also allow different covariance matrices for the two distributions. So, they can handle high dimensional Behrens-Fisher type problems.

Several nonparametric tests have been proposed for the general two-sample problem as well, where we test the equality of two continuous multivariate distributions F and G without making any further assumptions on them. Friedman and Rafsky (1979) used minimal spanning tree for multivariate generalizations of the Wald-Wolfowitz run test and the Kolmogorov-Smirnov test. Schilling (1986a) and Henze (1988) developed two-sample tests based on nearest neighbor type coincidences. Other nonparametric tests for the general two-sample problem include Hall and Tajvidi (2002), Baringhaus and Franz (2004, 2010), Aslan and Zech (2005), Liu and Modarres (2011) and Gretton et al. (2012). These tests can be used for HDLSS data, but they are not distribution-free in finite sample situations. Bickel (1969) showed that even the most natural multivariate generalization of the Kolmogorov-Smirnov statistic is not distribution-free for $d \geq 2$. Ferger (2000) proposed a distribution-free two-sample test from the perspective of change point detection, but for proper implementation of this test, one needs to find a suitable weight function and an appropriate asymmetric kernel function. Rosenbaum (2005) proposed a simpler distribution-free test for the general two-sample problem based on optimal non-bipartite matching (see e.g., Lu et al. (2011)). This test can be used for HDLSS data if the Euclidean metric is used for distance computation.

Instead of having two independent sets of observations from F and G, one can have n matched paired observations $\binom{\mathbf{x}_1}{\mathbf{y}_1}, \binom{\mathbf{x}_2}{\mathbf{y}_2}, \ldots, \binom{\mathbf{x}_n}{\mathbf{y}_n}$ from a $2d$-variate distribution with d-dimensional marginals F and G for $\mathbf{X}$ and $\mathbf{Y}$, respectively. Note that if F and G satisfy a location model (i.e., $F(\mathbf{x}) = G(\mathbf{x} - \boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \mathbb{R}^d$ and all $\mathbf{x} \in \mathbb{R}^d$), the distribution of $\mathbf{X} - \mathbf{Y}$ is symmetric about $\boldsymbol{\theta}$, and testing the equality of the locations of F and G is equivalent to test $H_0 : \boldsymbol{\theta} = 0$. So, in such cases, it is a common practice to consider it as a one-sample problem, where $\{\boldsymbol{\xi}_i = \mathbf{x}_i - \mathbf{y}_i;\ i = 1, 2, \ldots, n\}$ are used as sample observations to test $H_0 : \boldsymbol{\theta} = 0$ against $H_A : \boldsymbol{\theta} \neq 0$.

This one-sample problem is also well studied in the literature. If the distribution of $\mathbf{X} - \mathbf{Y}$ is assumed to be Gaussian, one uses the Student's t-statistic (when $d = 1$) or the one-sample Hotelling's T^2 statistic (when $d > 1$) to perform the test (see e.g. Mardia et al. (1979); Anderson (2003)). In the univariate case, one can also use distribution-free nonparametric tests (e.g., the sign test or the signed rank test) based on linear rank statistics (see e.g., Hájek et al. (1999); Gibbons and Chakraborti (2003)). Several attempts have also been made to generalize these rank-based tests to multivariate set up. Hodges (1955) and Blumen (1958) proposed distribution-free sign tests for bivariate data. Puri and Sen (1971) proposed tests based coordinate-wise signs and ranks. Randles (1989, 2000) developed one-sample location tests based on interdirections. Chaudhuri and Sengupta (1993) generalized Hodges' bivariate sign test to higher dimension. Other nonparametric tests for the multivariate one-sample problem include Bickel (1965); Hettmansperger et al. (1994); Möttönen et al. (1997); Hettmansperger et al. (1997); Chakraborty et al. (1998) and Hallin and Paindaveine (2002). For a brief overview of these tests, see Marden (1999); Oja and Randles (2004) and Oja (2010). Some of these multivariate nonparametric tests are distribution-free for some specific types of symmetric distributions, but none of them are distribution-free under general symmetry of the distribution of $\mathbf{X} - \mathbf{Y}$. So, one either uses the large sample test or the conditional test in such cases. However, these tests perform poorly for high dimensional data, and they cannot be used when the dimension exceeds the sample size. Recently, several one-sample tests have been proposed in the literature, which are applicable to HDLSS data (see e.g., Bai and Saranadasa (1996); Srivastava and Du (2008); Srivastava (2009); Chen and Qin (2010); Park and Ayyala (2013)). However, these Hotelling's T^2 type tests are mainly concerned with the mean vector of a high-dimensional distribution, and they are not robust. These tests are based on the asymptotic distribution of the test statistic, where the dimension increases with the sample size.

In the next two chapters of this thesis, we propose two nonparametric methods that can be used as general recipes for distribution-free multivariate generalizations of several univariate rank based two-sample tests. In both of these cases, the resulting tests are applicable to HDLSS data, and they retain the distribution-free property of their univariate analogs. Similar methods are used to develop distribution-free rank based tests for matched pair data as well.

In Chapter 2, we develop some two-sample tests using the idea of linear classification. Here we project the multivariate observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$ using a one-dimensional linear projection along a direction $\boldsymbol{\beta}$ and then use univariate distribution-free tests on the projected observations $\boldsymbol{\beta}^T\mathbf{x}_1, \ldots, \boldsymbol{\beta}^T\mathbf{x}_{n_1}, \boldsymbol{\beta}^T\mathbf{y}_1, \ldots, \boldsymbol{\beta}^T\mathbf{y}_{n_2}$. The projection direction $\boldsymbol{\beta}$ is estimated using a linear classifier that aims at separating the data clouds from the two-distributions. Two popular linear classification methods, support vector machines (SVM) (see e.g., Vapnik (1998)) and distance weighted discrimination (DWD) (Marron et al. (2007)) are used for this purpose. In order to develop a distribution-free test, we randomly split each of the two samples on $\mathbf{X}$ and $\mathbf{Y}$ into two disjoint subsamples. We use SVM or DWD to estimate $\boldsymbol{\beta}$ based on one subsample containing some of the $\mathbf{x}$'s and one subsample containing some of the $\mathbf{y}$'s. Then we project the observations in the remaining two subsamples using that estimated $\boldsymbol{\beta}$ and compute the test statistic based on the ranks of those projected observations. Given a fixed nominal level α, the test function is constructed accordingly. This procedure is repeated for different random splits and the results are aggregated. One simple way of aggregation is to use a test function, which is an average of the test functions obtained for different random splits. But, this aggregated test function may take a fractional value in the open interval $(0, 1)$, and hence the implementation of the test may require randomization at the final stage. We avoid this final stage randomization by using an alternative method based on Bonferroni correction (see e.g., Dunn (1961)) or that based on false discovery rate (FDR) that ensures the level property for aggregation of tests with positively regression dependent test statistics (see e.g., Benjamini and Yekutieli (2001)). The same strategy based on one-dimensional linear projection is also adopted to construct multivariate distribution-free tests for matched pair data, where we consider a classification problem involving two data clouds $\{\boldsymbol{\xi}_i = \mathbf{x}_i - \mathbf{y}_i,\ i = 1, \ldots, n\}$ and $\{\boldsymbol{\eta}_i = \mathbf{y}_i - \mathbf{x}_i,\ i = 1, \ldots, n\}$ and use the SVM or the DWD classifier to estimate $\boldsymbol{\beta}$. Asymptotic results on the power properties of our proposed tests are derived when the sample size is fixed and the dimension of the data grows to infinity as well as for situations when the sample size grows while the dimension remains fixed. We also investigate the finite sample performance of our proposed tests by applying them to several high dimensional simulated and real data sets. The contents of this chapter are partially based on Ghosh and Biswas (2015).

In Chapter 3, we propose another nonparametric method based on shortest Hamiltonian path (SHP). In the case of two-sample problem involving two independent sets of observations, we consider the $n = n_1 + n_2$ observations from F and G as the vertices of an edge weighted complete graph, where the edge between two vertices has a cost equal to the Euclidean distance between two corresponding observations. For any Hamiltonian path (the path that visits each vertex exactly once), the sum of the costs corresponding its $n - 1$ edges is defined as the cost of the Hamiltonian path. We find the SHP (Hamiltonian path with the minimum cost) in this complete graph, and ranks are assigned to the sample observations following that path. These ranks are used to construct rank based tests for multivariate data, which retain the exact distribution-free property of their univariate analogs. Following this idea, we propose a multivariate generalization of the univariate run test, which can be conveniently used in HDLSS situations.

Using a similar idea, we also develop some distribution-free tests for matched pair data, where we assume the distribution of $\mathbf{X} - \mathbf{Y}$ to be symmetric and test whether it is symmetric about the origin (or any given $\boldsymbol{\theta}_0 \in \mathbb{R}^d$). Given a sample of n observations $\mathcal{X} = \{\boldsymbol{\xi}_i = \mathbf{x}_i - \mathbf{y}_i,\ i = 1, \ldots, n\}$, we consider another set of n observations $\mathcal{X}^* = \{\boldsymbol{\eta}_i = \mathbf{y}_i - \mathbf{x}_i,\ i = 1, \ldots, n\}$. We consider these $2n$ observations as vertices of an edge weighted complete graph as before. A path of length $n-1$ in this graph is called a covering path if it visits either $\boldsymbol{\xi}_i$ or $\boldsymbol{\eta}_i$ for each $i = 1, \ldots, n$. The shortest among all such distinct paths (i.e. the covering path with the minimum cost) is termed as the shortest covering path (SCP). Signs and ranks of the sample observations are defined along this path. If an observation on this path comes from $\mathcal{X}$ (respectively, $\mathcal{X}^*$) we consider its sign to be positive (respectively, negative). Using this idea, we develop two run tests, one based on the number of runs and the other based on the length of the longest run. These tests are distribution-free and they can be used in HDLSS situations.

Under appropriate regularity conditions, we prove the consistency of all these proposed tests in HDLSS asymptotic regime, where the sample size remains fixed and the dimension of the data grows to infinity. Several simulated and real data sets are also analyzed to evaluate their empirical performance. The contents of this chapter are partially based on Biswas et al. (2014, 2015).

In Chapter 4, we propose some multivariate two-sample tests based on nearest neighbor type coincidences. Unlike the tests proposed in Chapters 2 and 3, these tests are not distribution-free in finite sample situations. Therefore, we use the permutation principle to make them conditionally distribution-free. These proposed tests can be viewed as modifications over the existing two-sample test based on nearest neighbors proposed by Schilling (1986a) and Henze (1988). While investigating the high-dimensional behavior of some popular classifiers, Hall et al. (2005) derived some conditions under which the traditional nearest neighbor classifier fails in high dimension. We show that the nearest neighbor test of Schilling (1986a) and Henze (1988) fails under the same set of conditions. In such cases, its power may even converge to zero as the dimension increases. Our proposed tests overcome this limitation. Under fairly general conditions, we prove their consistency in HDLSS asymptotic regime, where the sample size remains fixed and the dimension grows to infinity. Several high dimensional simulated and real data sets are analyzed to study their empirical performance. We further investigate some theoretical properties of these tests in classical asymptotic regime, where the dimension remains fixed and the sample size tends to infinity. In such cases, they turn out to be asymptotically distribution-free and consistent under general alternatives. The contents of this chapter are partially based on Mondal et al. (2015).

In Chapter 5, we propose a two-sample test based on averages of inter-point distances. Consider two independent observations $\mathbf{X}_1, \mathbf{X}_2$ from F and $\mathbf{Y}_1, \mathbf{Y}_2$ from G. Under moment conditions on F and G, Baringhaus and Franz (2004) proved that $\mathbb{D}(F, G) = 2E\|\mathbf{X}_1 - \mathbf{Y}_1\| - E\|\mathbf{X}_1 - \mathbf{X}_2\| - E\|\mathbf{Y}_1 - \mathbf{Y}_2\| \geq 0$, where the equality holds if and only if $F = G$. They used an empirical analog of $\mathbb{D}(F, G)$ for testing $H_0 : F = G$ and rejected the null hypothesis for higher values of the test statistic. We point out some limitations of this test in HDLSS set up. In particular, we show that this test may have poor power in high dimension, especially when the scale difference between two distributions dominates the location difference. In order to overcome this problem, we derive another equivalent condition for $F = G$ and construct a two-sample test based on that criterion. Here also, we use the permutation principle to determine the cut-off. Under appropriate regularity conditions, this proposed test is found to be consistent in HDLSS asymptotic regime. We also investigate the behavior of this test in classical asymptotic regime, where it turns out to be asymptotically distribution-free and consistent under general alternatives. Several high-dimensional simulated and real data sets are analyzed to evaluate its empirical performance. The contents of this chapter are partially based on Biswas and Ghosh (2014).

Finally, Chapter 6 contains a comparative discussion among different nonparametric methods proposed in this thesis, and it ends with a brief discussion on possible directions for further research.

Chapter 2

Tests based on discriminating hyperplanes

We know that two d-dimensional random vectors $\mathbf{X}$ and $\mathbf{Y}$ follow the same distribution if and only if $\boldsymbol{\beta}^T\mathbf{X}$ has the same distribution as $\boldsymbol{\beta}^T\mathbf{Y}$ for all $\boldsymbol{\beta} \in \mathbb{R}^d$. Therefore, if $\mathbf{X} \sim F$ and $\mathbf{Y} \sim G$, the null hypothesis $H_0 : F = G$ can be viewed as an intersection of the hypotheses $H_{0,\boldsymbol{\beta}} : F_{\boldsymbol{\beta}} = G_{\boldsymbol{\beta}}$ for varying choices of $\boldsymbol{\beta} \in \mathbb{R}^d$, where $\boldsymbol{\beta}^T\mathbf{X} \sim F_{\boldsymbol{\beta}}$ (respectively, $\boldsymbol{\beta}^T\mathbf{Y} \sim G_{\boldsymbol{\beta}}$) if $\mathbf{X} \sim F$ (respectively, $\mathbf{Y} \sim G$). Similarly, the alternative hypothesis $H_A : F \neq G$ can be viewed as an union of the hypotheses $H_{A,\boldsymbol{\beta}} : F_{\boldsymbol{\beta}} \neq G_{\boldsymbol{\beta}}$ (see e.g., Roy (1953) for union-intersection principle). Therefore, under the alternative H_A, one can expect to have some choices of the direction vector $\boldsymbol{\beta}$ for which $F_{\boldsymbol{\beta}}$ differs from $G_{\boldsymbol{\beta}}$. In this chapter, we use some multivariate statistical methods to find one such $\boldsymbol{\beta}$. If the multivariate sample observations $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}, \mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ are projected along $\boldsymbol{\beta}$, we get two sets of univariate observations $\boldsymbol{\beta}^T\mathbf{x}_1, \ldots, \boldsymbol{\beta}^T\mathbf{x}_{n_1} \sim F_{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^T\mathbf{y}_1, \ldots, \boldsymbol{\beta}^T\mathbf{y}_{n_2} \sim G_{\boldsymbol{\beta}}$. So, after finding $\boldsymbol{\beta}$, one can use any suitable univariate distribution-free two-sample test like the Wilcoxon-Mann-Whitney (WMW) test or the Kolmogorov-Smirvov (KS) test on these projected observations.

Clearly, any test based on ranks of a fixed linear function of multivariate observations has the exact distribution-free property. However, in order to have good power properties of such a test based on linear projection, one should choose the direction vector $\boldsymbol{\beta}$ in such a way that the separation between the projected observations from the two populations is maximized along that direction in an appropriate sense. One possible way to achieve this is to use the direction vector of a suitable linear classifier that discriminates between two multivariate populations. The motivation for this choice partially comes from the fact that for two multivariate normal distributions with a common dispersion and different means, if one computes the univariate two-sample t-statistic based on linear projections of the data points along the director vector used in Fisher's linear discriminant function, where the mean vectors and the common covariance matrix for the two distributions are estimated from the data, it leads to the Hotelling's T^2 statistic. Further, for two independent normal random vectors $\mathbf{X}$ and $\mathbf{Y}$ with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and a common dispersion matrix $\boldsymbol{\Sigma}$, the power of the univariate t-test for testing $H_{0,\boldsymbol{\beta}} : \boldsymbol{\beta}^T\boldsymbol{\mu}_1 = \boldsymbol{\beta}^T\boldsymbol{\mu}_2$ based on $\boldsymbol{\beta}^T\mathbf{X}$ and $\boldsymbol{\beta}^T\mathbf{Y}$ is a monotonically increasing function of $\{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}^2/\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}$. So, the power of the test is maximized when $\boldsymbol{\beta}$ is chosen to be a scalar multiple of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, which is the coefficient vector of Fisher's linear discriminant function.

Even when the underlying distributions are not normal, we have some nice connections between classification and hypothesis testing problems. Consider a classification problem between two multivariate distributions F and G such that the prior probabilities of these two distributions are equal. Let us also consider a discriminating hyperplane $\{\mathbf{z} : \beta_0 + \boldsymbol{\beta}^T\mathbf{z} = 0,\ \mathbf{z} \in \mathbb{R}^d\}$ between these two distributions. Suppose that the classifier classifies $\mathbf{z}$ as an observation from F (respectively, G) if $\beta_0 + \boldsymbol{\beta}^T\mathbf{z} > 0$ (respectively, $\beta_0 + \boldsymbol{\beta}^T\mathbf{z} \leq 0$). Clearly, the average misclassification probability of this classifier is given by $0.5[1 - \{F_{\boldsymbol{\beta}}(-\beta_0) - G_{\boldsymbol{\beta}}(-\beta_0)\}]$, which is minimized if and only if $\boldsymbol{\beta}$ maximizes the Kolmogorov-Smirnov (KS) distance between $F_{\boldsymbol{\beta}}$ and $G_{\boldsymbol{\beta}}$. Further, when F and G are both elliptically symmetric unimodal distributions, which differ only in their locations, we have the following proposition, which yields an interesting insight into the connection between classifiers having the optimal misclassification rate and tests having the optimal power. The proof is given in Section 2.9.

PROPOSITION 2.1: *Suppose that F and G are elliptically symmetric unimodal multivariate distributions, which differ only in their locations. Then the one sided KS test as well as the one sided WMW test based on the ranks of* $\boldsymbol{\beta}^T\mathbf{x}_1, \ldots, \boldsymbol{\beta}^T\mathbf{x}_{n_1}, \boldsymbol{\beta}^T\mathbf{y}_1, \ldots, \boldsymbol{\beta}^T\mathbf{y}_{n_2}$

have the maximum power if and only if $\boldsymbol{\beta}$ coincides with the direction vector that determines the Bayes discriminating hyperplane associated with the classification problem involving distributions F and G with equal prior probabilities.

Note at this point that a linear classifier, which has its class boundary defined by the hyperplane $\{\mathbf{z} : \beta_0 + \boldsymbol{\beta}^T\mathbf{z} = 0,\ \mathbf{z} \in \mathbb{R}^d\}$ classifies $\mathbf{z}$ as an observation from the distribution F if it falls on one side of that hyperplane, and $\mathbf{z}$ is classified as an observation from G if it falls on the other side. Suppose that it classifies $\mathbf{z}$ as an observation from F (respectively, G) if $\beta_0 + \boldsymbol{\beta}^T\mathbf{z} > 0$ (respectively, $\beta_0 + \boldsymbol{\beta}^T\mathbf{z} \leq 0$). So, when we project the observations along the direction $\boldsymbol{\beta}$, projected observations from F are likely to have higher ranks than projected observations from G. Therefore, it is appropriate to consider the one sided KS test or the one sided WMW test (see e.g., Gibbons and Chakraborti (2003)) based on the ranks of $\boldsymbol{\beta}^T\mathbf{x}_1, \boldsymbol{\beta}^T\mathbf{x}_2, \ldots, \boldsymbol{\beta}^T\mathbf{x}_{n_1}, \boldsymbol{\beta}^T\mathbf{y}_1, \boldsymbol{\beta}^T\mathbf{y}_2, \ldots, \boldsymbol{\beta}^T\mathbf{y}_{n_2}$.

2.1 Adaptive determination of the direction vector

It is well-known that Fisher's linear discriminant function yields an optimal separation between two classes of observations when the underlying distributions are Gaussian having a common dispersion but different means. However, when one needs to estimate the dispersion and the means from the data, the estimated discriminant function performs poorly for high dimensional data. If the dimension exceeds the total sample size, the estimated dispersion becomes singular, and it cannot be used to construct Fisher's linear discriminant function. If one uses Fisher's linear discriminant function based on the Moore-Penrose generalized inverse of the pooled dispersion matrix in such situations, it usually yields poor performance in high dimensions (see e.g., Bickel and Levina (2004)).

Support vector machine (SVM) (see e.g., Vapnik (1998); Burges (1998)) is a well-known classification tool that can be used for linear classification between two distributions when the data are high dimensional. Suppose that we have a data set of the form $\{(\mathbf{z}_i, \omega_i);\ i = 1, 2, \ldots, n = n_1 + n_2\}$, where ω_i takes the value 1 and -1 if the observation $\mathbf{z}_i$ comes from the first population (i.e., $\mathbf{z}_i = \mathbf{x}_j$ for some j) and the second population (i.e., $\mathbf{z}_i = \mathbf{y}_j$ for some j), respectively. When the data clouds from the two distributions have perfect linear separation, SVM looks for two parallel hyperplanes $\beta_0 + \boldsymbol{\beta}^T\mathbf{z} = 1$ and $\beta_0 + \boldsymbol{\beta}^T\mathbf{z} = -1$ such that $(\beta_0 + \boldsymbol{\beta}^T\mathbf{z}_i)\omega_i \geq 1$ for all $i = 1, 2, \ldots, n$, and the distance between these two hyperplanes $2/\|\boldsymbol{\beta}\|$ is maximum. In practice, it finds the separating hyperplane $\beta_0 + \boldsymbol{\beta}^T\mathbf{z} = 0$ by minimizing $\frac{1}{2}\|\boldsymbol{\beta}\|^2$ subject to $(\beta_0 + \boldsymbol{\beta}^T\mathbf{z}_i)\omega_i \geq 1 \; \forall \; i = 1, 2, \ldots, n$. If the data clouds from the two distributions are not perfectly linearly separable, SVM introduces slack variables ζ_i $(i = 1, 2, \ldots, n)$ and modifies the objective function by adding a cost $C_0 \sum_{i=1}^{n} \zeta_i$ (C_0 is a cost parameter) to it. In such cases, SVM minimizes $\frac{1}{2}\|\boldsymbol{\beta}\|^2 + C_0 \sum_{i=1}^{n} \zeta_i$ subject to $(\beta_0 + \boldsymbol{\beta}^T\mathbf{z}_i)\omega_i \geq 1 - \zeta_i$ and $\zeta_i \geq 0 \; \forall \; i = 1, 2, \ldots, n$, and it uses the quadratic programming technique for this minimization. This optimization problem is often reformulated as the problem of minimizing $S_n(\beta_0, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[1 - \omega_i(\beta_0 + \boldsymbol{\beta}^T\mathbf{z}_i)]_+ + \frac{\lambda_0}{2}\|\boldsymbol{\beta}\|^2$, where $[t]_+ = \max\{t, 0\}$ and $\lambda_0 = 1/C_0$ is a regularization parameter (see e.g., Hastie et al. (2004)).

Marron et al. (2007) proposed another classification technique called distance weighted discrimination (DWD), which can also be used for linear classification in high dimensions. If the data clouds from the two distributions are perfectly linearly separable, DWD finds the separating hyperplane by minimizing $\sum_{i=1}^{n}\{(\beta_0 + \boldsymbol{\beta}^T\mathbf{z}_i)\omega_i\}^{-1}$ subject to $\|\boldsymbol{\beta}\| \leq 1$ and $(\beta_0 + \boldsymbol{\beta}^T\mathbf{z}_i)\omega_i \geq 0$ for all $i = 1, 2, \ldots, n$. When the data clouds are not linearly separable, DWD also introduces slack variables ζ_i to modify the objective function by adding a cost $C\sum_{i=1}\zeta_i$, where C is a cost parameter. In such cases, DWD finds the separating hyperplane $\beta_0 + \boldsymbol{\beta}^T\mathbf{z} = 0$ by minimizing $\sum_{i=1}^{n} 1/r_i + C\sum_{i=1}^{n}\zeta_i$ subject to $\|\boldsymbol{\beta}\| \leq 1$, $\zeta_i \geq 0$ and $r_i = (\beta_0 + \boldsymbol{\beta}^T\mathbf{z}_i)\omega_i + \zeta_i \geq 0$ for all $i = 1, 2, \ldots, n$. This is equivalent to minimization of $D_n(\beta_0, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[V_0\{\omega_i(\beta_0 + \boldsymbol{\beta}^T\mathbf{z}_i)\}]$, where

$$V_0(t) = \begin{cases} 2\sqrt{C} - Ct & \text{if } t \leq 1/\sqrt{C} \\ 1/t & \text{otherwise,} \end{cases}$$

(see e.g., Qiao et al. (2010)). DWD uses the interior point cone programming to minimize $D_n(\beta_0, \boldsymbol{\beta})$ and to estimate $\boldsymbol{\beta}$.

2.2 Construction of distribution-free two-sample tests

Clearly, for any fixed and non-random $\boldsymbol{\beta}$, the random variables $\boldsymbol{\beta}^T\mathbf{x}_1, \ldots, \boldsymbol{\beta}^T\mathbf{x}_{n_1}, \boldsymbol{\beta}^T\mathbf{y}_1, \ldots, \boldsymbol{\beta}^T\mathbf{y}_{n_2}$ form an exchangeable collection if $F = G$, and the ranks of these variables have the distribution-free property under H_0. Let $T_{\boldsymbol{\beta}}$ be a statistic based on the ranks of $\boldsymbol{\beta}^T\mathbf{x}_1, \ldots, \boldsymbol{\beta}^T\mathbf{x}_{n_1}, \boldsymbol{\beta}^T\mathbf{y}_1, \ldots, \boldsymbol{\beta}^T\mathbf{y}_{n_2}$ Assume that, for any specified level $0 < \alpha < 1$, the test for the null hypothesis $H_0 : F = G$ based on $T_{\boldsymbol{\beta}}$ is described by the test function

$$\phi_\alpha(T_{\boldsymbol{\beta}}) = \begin{cases} 1 & \text{if } T_{\boldsymbol{\beta}} > t_\alpha \\ \gamma_\alpha & \text{if } T_{\boldsymbol{\beta}} = t_\alpha \\ 0 & \text{otherwise,} \end{cases}$$

where one can choose t_α and γ_α in such a way that $E_{H_0}\{\phi_\alpha(T_{\boldsymbol{\beta}})\} = \alpha$. Because of the distribution-free property of $T_{\boldsymbol{\beta}}$, t_α and γ_α depend neither on (F, G) nor on $\boldsymbol{\beta}$. Further, for standard nonparametric tests (e.g., the KS test or the WMW test), one can obtain t_α and γ_α from standard statistical tables or softwares.

Note that if $\boldsymbol{\beta}$ is estimated based on the whole sample using the SVM or the DWD classifier, and then the multivariate observations are ranked after projecting them along that estimated direction $\widehat{\boldsymbol{\beta}}$, the resulting ranks do not have the distribution-free property. This is due to the fact that $\widehat{\boldsymbol{\beta}}$, which is constructed from a classification problem based on the two samples, is not a symmetric function of the observations in the combined sample, and the random variables $\widehat{\boldsymbol{\beta}}^T\mathbf{x}_1, \ldots, \widehat{\boldsymbol{\beta}}^T\mathbf{x}_{n_1}, \widehat{\boldsymbol{\beta}}^T\mathbf{y}_1, \ldots, \widehat{\boldsymbol{\beta}}^T\mathbf{y}_{n_2}$ do not form an exchangeable collection even if $F = G$. Therefore, in order to have a distribution-free test, we adopt a strategy, which is motivated by the idea of cross-validation techniques used in statistical model selection. In cross-validation, one splits the whole sample into subsamples and then uses one subsample to estimate the model by optimizing a suitable criterion, while another subsample is used to assess the adequacy of the estimated model. In a similar way, we randomly split each of the two samples into two disjoint subsamples. We use a suitable linear classifier (e.g., SVM or DWD) to construct $\widehat{\boldsymbol{\beta}}$ based on one subsample of containing m_1 $\mathbf{x}$'s and one subsample containing m_2 $\mathbf{y}$'s. Then we project the observations in the remaining two subsamples (of size $n_1 - m_1$ and $n_2 - m_2$) using that $\widehat{\boldsymbol{\beta}}$ and compute the test statistic $T_{\widehat{\boldsymbol{\beta}}}$ and the test function $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})$ based on the ranks of those projected observations. The following theorem shows the exact distribution-free property of ϕ_α. The proof of the theorem is given in Section 2.9.

THEOREM 2.1: *ϕ_α is a distribution-free test function in the sense that $E_{(F,G)}(\phi_\alpha) = \alpha$ for all (F, G) such that $F = G$.*

In practice, we repeat this procedure for several random splits and aggregate the results to come up with the final decision. For a given level α $(0 < \alpha < 1)$, one option for aggregation is to consider a test function ϕ^*_α, which is obtained by averaging the test functions $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})$ over M different random splits. From Theorem 2.1, it follows that $E_{H_0}(\phi^*_\alpha) = \alpha$. However, ϕ^*_α may take a fractional value in the interval $(0, 1)$ for a given data set. Therefore, the implementation of the test may require randomization at the final stage. We can avoid this final stage randomization by using the idea of union-intersection test based on Bonferroni correction (see e.g., Dunn (1961)). For each of the M random splits, we perform the nonparametric test (e.g., one-sided WMW or KS test) at level α/M, and finally accept the null hypothesis if and only if it is accepted for each random split. One can also use an alternative method based on the idea of controlling the false discovery rate (FDR) (see e.g., Benjamini and Hochberg (1995)). In this case, for each random split, we compute the p-value associated with the nonparametric test. Let $p_1, p_2, \ldots, p_M$ be the p-values associated with the tests based on M random splits, and $p_{(1)}, p_{(2)}, \ldots, p_{(M)}$ be the corresponding order statistics. For a given level α $(0 < \alpha < 1)$, we reject H_0 if the set $\{i : p_{(i)}/i \leq \alpha/M\}$ is non-empty. FDR was introduced by Benjamini and Hochberg (1995) for independent tests. Later, Benjamini and Yekutieli (2001) showed that the method of Benjamini and Hochberg also controls FDR for tests with positively regression dependent (PRD) test statistics. Benjamini and Yekutieli (2001) also developed a FDR procedure, which does not require the the test statistics to be positively regression dependent and works under arbitrary dependence structure. In this method, after finding the p-values, one rejects H_0 if the set $\{i : p_{(i)}/i \leq \alpha/M_0(i)\}$ is non-empty, where $M_0(i) = M \sum_{j=1}^{i} 1/j$ for $i = 1, 2, \ldots, M$. Since $M \leq M_0(i) \leq iM$ for all $i = 1, 2, \ldots, M$, this method is more conservative than the FDR procedure that assumes positive regression dependence structure, but less conservative than the Bonferroni method. So, in any given example, it is expected to yield power lying between those of the Bonferroni method and the Benjamini-Hochberg procedure. However, here we can safely assume that the tests corresponding to different random splits have PRD test statistics because they are based on the same initial data set (see also Cuesta-Albertos and Febrero-Bande (2010)), and since we are testing the same hypothesis over different partitions, in our case, FDR coincides with the level of the resulting test.

Therefore, the level of this resulting test can atmost be α (see Theorem 1.2 in Benjamini and Yekutieli (2001) and Proposition 2.3 in Cuesta-Albertos and Febrero-Bande (2010)). Henceforth, by FDR method, we will mean the FDR method proposed by Benjamini and Hochberg (1995), and we will use it for our all theoretical and numerical work.

Recently, Wei et al. (2015) proposed some two-sample tests based on linear projections, where they also used a linear classifier to select the projection direction and computed the test statistic based on the projected observations. But, they used the full sample to estimate the direction vector and to compute the test statistic. So, their tests were not distribution-free, and they had to use the permutation principle to make them conditionally distribution-free. However, their conditional tests based on the t-statistic and the MD statistic work well only for light tailed distributions. Unlike our rank based methods, they are not robust. Their test based on the AUC statistic is somewhat robust, but it does not have good power properties in HDLSS situations, where the observations from the two distributions are linearly separable (see Wei et al. (2015) for details). Our proposed tests based on WMW and KS statistics do not have such problems in high dimension, which we will see in the subsequent sections.

2.3 Power properties of proposed tests for HDLSS data

We have already mentioned that unlike most of the existing two-sample tests, our tests based on SVM and DWD can be used even when the dimension of the data is much larger than the sample size. Here, we carry out some theoretical analysis of the power properties of these tests when the sample size n is fixed, and the dimension d diverges to infinity. Throughout this section, we consider tests based on the one sided KS and the one sided WMW statistics. We consider all three aggregation methods, the method based on the average of test functions, the method based on Bonferroni correction and that based on FDR as discussed above. Henceforth, we will refer to them as the Avg-Method, the Bonf-Method and the FDR-Method, respectively. For our theoretical investigation, we assume the observations on $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})^T$ and $\mathbf{Y} = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(d)})^T$ to be independent, and they also satisfy the following assumptions.

(A1) *Fourth moments of $X^{(q)}$ and $Y^{(q)}$ ($q \geq 1$) are uniformly bounded.*

(A2) *Let* $\mathbf{X}_1, \mathbf{X}_2$ *be two independent copies of* $\mathbf{X}$*, and* $\mathbf{Y}_1, \mathbf{Y}_2$ *be two independent copies of* $\mathbf{Y}$*. For* $(\mathbf{U}, \mathbf{V}) = (\mathbf{X}_1, \mathbf{X}_2), (\mathbf{X}_1, \mathbf{Y}_1)$ *and* $(\mathbf{Y}_1, \mathbf{Y}_2)$*, the sum of all pairwise correlations,* $\sum_{q \neq q'} |corr\{(U^{(q)} - V^{(q)})^2, (U^{(q')} - V^{(q')})^2\}|$*, is of order* $o(d^2)$*.*

(A3) *There exist constants* $\sigma_1^2, \sigma_2^2 > 0$ *and* ν *such that* $d^{-1} \sum_{q=1}^{d} Var(X^{(q)}) \to \sigma_1^2$*,* $d^{-1} \sum_{q=1}^{d} Var(Y^{(q)}) \to \sigma_2^2$ *and* $d^{-1} \sum_{q=1}^{d} \{E(X^{(q)}) - E(Y^{(q)})\}^2 \to \nu^2$ *as* $d \to \infty$*.*

Under (A1) and (A2), the weak law of large number (WLLN) holds for the sequence $\{(U^{(q)} - V^{(q)})^2;\ q \geq 1\}$, i.e., $d^{-1} \left| \sum_{q=1}^{d} (U^{(q)} - V^{(q)})^2 - \sum_{q=1}^{d} E(U^{(q)} - V^{(q)})^2 \right| \overset{P}{\to} 0$ as $d \to \infty$ (the proof is straight forward and therefore omitted). Under (A3), one can compute the limiting value of $d^{-1} \sum_{q=1}^{d} E(U^{(q)} - V^{(q)})^2$ or that of $d^{-1} \sum_{q=1}^{d} (U^{(q)} - V^{(q)})^2$ as $d \to \infty$. This limiting value turns out to be $2\sigma_1^2$, $\sigma_1^2 + \sigma_2^2 + \nu^2$ and $2\sigma_2^2$ for $(\mathbf{U}, \mathbf{V}) = (\mathbf{X}_1, \mathbf{X}_2), (\mathbf{X}_1, \mathbf{Y}_1)$ and $(\mathbf{Y}_1, \mathbf{Y}_2)$, respectively.

Note that we need (A1) and (A2) to have WLLN for the sequence of dependent and non-identically distributed random variables. If the components of $\mathbf{X}$ and $\mathbf{Y}$ are independent and identically distributed (i.i.d.), WLLN holds under the existence of second order moments of $X^{(q)}$ and $Y^{(q)}$. In that case, (A2) and (A3) get automatically satisfied, and (A1) is not required.

Hall et al. (2005) looked at d-dimensional observations as infinite time series $\big(X^{(1)}, X^{(2)}, \ldots\big)$ truncated at length d and studied the high dimensional behavior of pairwise distances assuming a form of ρ-mixing (see e.g., Kolmogorov and Rozanov (1960)) for the time series. Assumption (A2) holds under that ρ-mixing condition. Jung and Marron (2009) assumed some weak dependence among measurement variables to study the high dimensional consistency of estimated principal component directions. Assumption (A2) also holds under those conditions. Andrews (1988) and de Jong (1995) also derived some sufficient conditions to have WLLN for the sequence of dependent and non-identically distributed random variables. Instead of (A1) and (A2), one can assume those conditions as well.

From our above discussion, it is quite transparent that under the assumptions (A1)-(A3), the Euclidean distance between any two observations, when divided by $d^{1/2}$, converges in probability to positive constant as d tends to infinity. If both of them are from the same distribution, it converges to $\sigma_1\sqrt{2}$ or $\sigma_2\sqrt{2}$ depending on whether they are from F or G. If one of them is from F and the other one is from G, it converges to $\sqrt{\sigma_1^2+\sigma_2^2+\nu^2}$. So, for large d, after re-scaling by a factor of $d^{-1/2}$, n sample observations tend to lie on the vertices of an n-polyhedron. Note that n_1 out of these n vertices are limits of n_1 i.i.d observations from F, and they form a regular simplex $\mathcal{S}_1$ of side length $\sigma_1\sqrt{2}$. The other n_2 vertices are limits of n_2 data points from G, and they form another regular simplex $\mathcal{S}_2$ of side length $\sigma_2\sqrt{2}$. The rest of the edges of the polyhedron connect the vertices of $\mathcal{S}_1$ to those of $\mathcal{S}_2$, and they are of length $\sqrt{\sigma_1^2+\sigma_2^2+\nu^2}$. Under H_0, when we have $\sigma_1^2=\sigma_2^2$ and $\nu^2=0$, and the whole polyhedron turns out to be a regular simplex on n points, while we may have $\nu^2>0$ under H_A. In a sense, (A1)-(A3) and $\nu^2>0$ ensure that the amount of information for discrimination between F and G grows to infinity as the dimension increases (see Hall et al. (2005) for further discussion). In conventional asymptotics, we get more information as the sample size increases, but here the sample size n is fixed and we expect the amount of information to diverge as the dimension d tends to infinity. In classical asymptotic regime, where d is fixed and n tends to infinity, consistency of a test is a rather trivial property. The power of any reasonable test converges to unity as the sample size increases. But when the sample size is fixed, and the dimension tends to infinity, consistency of a test is no longer a trivial property, and many well known and popular tests fail to have the consistency in this set up (see e.g., Wei et al. (2015)). The next theorem establishes the consistency of our proposed tests in this high dimensional asymptotic regime. The proof of the theorem is given in Section 2.9.

THEOREM 2.2: *Let $\widehat{\boldsymbol{\beta}}$ be computed using SVM or DWD applied to the two subsamples of sizes m_1 and m_2, and $T_{\widehat{\boldsymbol{\beta}}}$ be computed from the other two subsamples of sizes n_1-m_1 and n_2-m_2. Assume that $T_{\boldsymbol{\beta}}$ is either the one sided KS statistic or one sided WMW statistic such that $P_{H_0}(T_{\boldsymbol{\beta}}=t_{\max})<\alpha$, where $t_{\max}$ is the largest possible value of the statistic $T_{\boldsymbol{\beta}}$ computed based on two subsamples of sizes n_1-m_1 and n_2-m_2. Then under the assumptions (A1)-(A3), if $\nu^2>0$, the power of the proposed test based on the Avg-Method or the FDR-Method converges to unity. Under the same set of assumptions, the power of the proposed test based on the Bonf-Method also converges to unity if $P_{H_0}(T_{\boldsymbol{\beta}}=t_{\max})<\alpha/M$, where M is the number of random splits.*

It is appropriate to mention here that not only for the one sided KS and the one sided WMW statistics, the above result holds for any one sided linear rank statistic (see e.g., Hájek et al. (1999)) of the form $\sum_{i=1}^{m_1} a(R_i)$, where the R_is are the rank of the projected observations on $\mathbf{X}$ in the combined sample, and a is a monotonically increasing function. Also, in view of the results in Hall et al. (2005), the convergence of the powers of our tests to one actually holds even when both d and $m = (m_1+m_2)$ grow to infinity in such a way that m/d^2 tends to zero and $n-m$ is not too small. One should notice that depending on the values of σ_1^2, σ_2^2 and ν^2, both SVM and DWD need some additional conditions on m_1 and m_2 for perfect classification of future observations, otherwise they classify all observations to a single class (see Hall et al. (2005)). But, for our tests based on SVM and DWD directions, we do not need such conditions for the convergence of the power function to unity. Note also that the condition $\nu^2 > 0$ holds in the commonly used set up for two-sample testing problems, where the population distributions are assumed to have the same dispersion but different means. For the one sided KS statistics as well as any one sided linear rank statistic (as mentioned above), it is easy to see that $T_{\widehat{\boldsymbol{\beta}}}$ takes its maximum value $t_{\max}$ if and only if the rank of the linear function of any observation from F is smaller than that of any observation from G in the combined sample. Hence, $P_{H_0}(T_{\widehat{\boldsymbol{\beta}}} = t_{\max}) = (n_1-m_1)!(n_2-m_2)!/(n-m)!$ is smaller than α if $n-m$ is suitably large.

2.4 Results from the analysis of simulated data sets

We begin with a comparison among the powers of our tests based on three methods of aggregation discussed in Section 2.2. Note that in all these cases, we need to find $\widehat{\boldsymbol{\beta}}$ either using the SVM classifier or using the DWD classifier on the observations in the first subsample. For the SVM classifier, we used the R program 'svmpath' (see Hastie et al. (2004)), which automatically selects the regularization parameter λ_0. For the DWD classifier, we used the MATLAB codes of Marron et al. (2007) with the default penalty function. After finding $\widehat{\boldsymbol{\beta}}$, observations in the second subsample were projected along that direction to compute the test function and the p-value. This procedure was repeated 50 times, and the results were aggregated over these 50 random splits.

Unless mentioned otherwise, throughout this thesis, for all numerical work, all tests are considered to have 5% nominal level.

We considered some examples involving spherically symmetric multivariate normal and Cauchy distributions, where F and G had the same scatter matrix $\mathbf{I}_d$ (the $d \times d$ identity matrix) and differed only in their locations. Note that our proposed tests are invariant under a common location shift and a common orthogonal transformation of the data from F and G in view of the equivariance property of SVM and DWD classifiers under those transformations. For such an invariant two-sample test, its power is a function of the norm of the difference between the locations of two spherical distributions. We chose F to be symmetric around the origin and G to be symmetric around $(\Delta, 0, \ldots, 0)^T$. We considered two choices for d, and the value of Δ was chosen to be 1.5 and 2 for $d = 30$ and $d = 90$, respectively, so that all tests had powers appreciably different from the nominal level of 0.05. Note that while normal distributions have exponential tails and finite moments of all orders, Cauchy distributions have heavy polynomial tails and they do not have finite moments of any order. Assumptions (A1)-(A3) hold for normal distributions, but not for Cauchy distributions. We chose these two distributions in order to evaluate the performance of our tests not only when (A1)-(A3) hold, but also in situations when they fail to hold. In each of these examples, we generated 50 observations from each distribution to form the sample, which was then used to perform different tests. We carried out 1000 Monte-Carlo experiments, and for each test, we estimated its power by the proportion of times it rejected H_0.

Recall that for the implementation of our tests, we need to divide the whole sample into two subsamples. We carried out our experiment taking π proportion of observations in the first sub-sample, and computed the power of the corresponding test p_π for nine different choices of π ($\pi = 0.1, 0.2, \ldots, 0.9$). The relative power for a given value of π is computed as p_π / p_*, where $p_* = \max_\pi p_\pi$. Figure 2.1 shows these relative powers for our tests based on three methods. From this figure, it seems to be a good idea to use $\pi \in [0.2, 0.3]$. We carried out our experiment with different choices of F and G, but in most of the cases, our finding remained the same. Henceforth, we will use $\pi = 0.25$ for our tests.

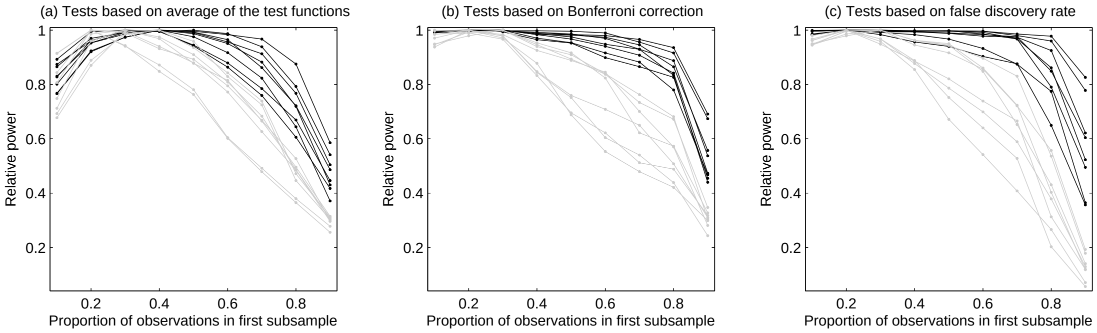


Figure 2.1: Relative powers of proposed tests for different sizes of first subsample. (black and grey curves shows the results for examples with normal and Cauchy distributions, respectively)

Table 2.1 shows the observed levels (when $\Delta = 0$) and powers of our proposed tests for $\pi = 0.25$. Recall that the FDR-Method controls the level of the test when the test statistics corresponding to different random splits are either independent or positively regression dependent (PRD) (see Benjamini and Hochberg (1995); Benjamini and Yekutieli (2001)). We computed correlation coefficients among these test statistics over 1000 Monte-Carlo simulations, and in all cases, all of them turned out to be positive. Observed levels of the tests based on the FDR-Method were below the nominal level. These give an indication that the test statistics corresponding to different random splits were PRD. Table 2.1 shows that while our tests based on the Avg-Method had observed levels quite close to 0.05 in all cases, tests based on the Bonf-Method and the FDR-Method had observed levels falling below their nominal levels in various cases. But in spite of their conservativeness, in all these cases, the Bonf-Method and the FDR-Method yielded powers significantly higher than those obtained using the Avg-Method. Sometimes, some of the random splits led to slightly lower values of the test statistic, and that affected the performance of the Avg-Method. However, the Bonf-Method and the FDR-Method did not get much affected by this fact because a very strong evidence in a single split is enough to reject H_0 in these cases. We carried out our experiment for different sample sizes and also for different choices of F and G, but the superiority of these two methods was evident in almost all cases. Between them, the latter had a slight edge. So, from now on, we will use tests based on the FDR-Method only.

Table 2.1: Observed levels and powers (in %) of proposed tests

		Normal				Cauchy			
		$d = 30$		$d = 90$		$d = 30$		$d = 90$	
		$\Delta = 0$	$\Delta = 1.5$	$\Delta = 0$	$\Delta = 2$	$\Delta = 0$	$\Delta = 1.5$	$\Delta = 0$	$\Delta = 2$
WMW-SVM	Avg-Method	5.3	79.2	5.3	93.3	4.4	39.1	5.2	59.2
	Bonf-Method	2.6	95.2	2.4	99.6	1.8	53.7	2.9	77.6
	FDR-Method	2.8	96.4	2.7	99.6	1.9	59.2	3.3	83.0
WMW-DWD	Avg-Method	5.0	93.1	4.5	97.3	4.1	41.6	4.4	57.3
	Bonf-Method	2.0	96.6	2.8	99.8	2.4	51.4	2.5	71.0
	FDR-Method	2.3	97.6	3.3	99.8	2.4	58.9	2.7	75.7
KS-SVM	Avg-Method	5.5	71.7	5.7	88.4	4.9	43.3	5.1	64.6
	Bonf-Method	2.8	94.2	2.8	99.6	2.0	70.7	2.6	87.8
	FDR-Method	3.4	95.2	3.8	99.3	2.7	70.8	3.2	88.0
KS-DWD	Avg-Method	5.2	88.9	4.7	94.7	5.6	46.8	4.5	64.0
	Bonf-Method	2.8	96.0	2.7	99.6	3.0	69.2	2.3	84.4
	FDR-Method	3.1	96.9	3.4	99.7	3.3	69.4	2.3	84.5

Next, we compared the performance of our tests (based on the FDR-Method) with some popular two-sample tests available in the literature. The test based on the Hotelling's T^2 statistic, spatial sign and rank tests (Sp-sign and Sp-rank) (see e.g., Möttönen and Oja (1995); Choi and Marden (1997)), Puri and Sen (1971)'s coordinate-wise sign and rank tests (PS-sign and PS-rank) were used for this comparison. For these sign and rank tests, we used both, the test based on the large sample distribution of the test statistic and the conditional test based on the permutation principle. In each case, the best one (which happened to be the conditional test in most of the cases) has been reported in Table 2.2. The codes for these tests are available in MNM (see Oja (2010) for details) and other packages in R. As we have mentioned before, Bai and Saranadasa (1996), Chen and Qin (2010), Park and Ayyala (2013) and Srivastava et al. (2013) proposed some Hotelling's T^2 type tests, which can be used even when the dimension is larger than the combined sample size, and the scatter matrix of the two distributions are different. We considered the last three for comparison.

Results are also reported for the test based on nearest neighbor (NN) type coincidences (see e.g., Schilling (1986a); Henze (1988)), the Cramer test (see Baringhaus and Franz (2004)), the multivariate run test (see Friedman and Rafsky (1979)) based on minimal spanning tree (MST) and Rosenbaum's Adjacency test (Rosenbaum (2005)) based on optimal non-bipartite matching (see e.g., Lu et al. (2011)). The codes for the NN test (we used the test based on three neighbors) and the Cramer test are available in R packages ‘MTSKNN’ and ‘cramer’, respectively. For the MST run test and the Adjacency test (based on Euclidean distance), we used our own codes. For comparison among different two sample tests, along with our previous examples involving 30 and 90 dimensional normal and Cauchy distributions, we also considered similar examples involving t distributions with 2 degrees of freedom. These three distributions were chosen because of varying degrees of heaviness of their tails. Cauchy distributions do not have finite first order moments, t_2 distributions have first order moments, but they do not have second order moments, and normal distributions have moments of all orders.

Table 2.2 shows that in the case of normal distributions, the Hotelling's T^2 test had observed levels close to the nominal level of 0.05, but in cases of t_2 (t wih 2 d.f.) and Cauchy distributions, they were marginally lower. Observed levels of Chen and Qin's test (CQ test) were slightly higher than 0.05 in some cases, whereas those of Park and Ayyala's test (PA test) were marginally below the nominal in the case of Cauchy distributions. Srivastava's test (SKK test) also had levels below 0.05 in cases of t_2 and Cauchy distributions, and in the case of Cauchy distributions, they were zero. However, all conditional tests had observed levels close to 0.05 in almost all examples.

In examples involving normal distributions, the Hotelling's T^2 test had good power properties for $d = 30$, but it did not perform well for $d = 90$. PS-sign, PS-rank, Sp-sign and Sp-rank tests also had similar behavior. The CQ test had the highest power in these two cases, while SKK, PA, Cramer and our proposed tests also had competitive performance. However, in cases of t_2 and Cauchy distributions, CQ, SKK, PA and Cramer tests did not have satisfactory performance at all. Note that CQ, SKK and PA tests were designed for testing the equality of two mean vectors. Although the location parameter is well defined in a Cauchy distribution, the mean vector does not exist. This was the main reason for the poor performance by these tests. Even in the case of heavy tailed t_2 distribution, these non-robust methods could not perform well. In cases of t_2 and Cauchy distributions, the Hotelling's T^2 test also had powers much lower than many of its nonparametric competitors. The NN test had the highest power for $d = 30$, but in the case of $d = 90$, our proposed tests outperformed their competitors when the KS test statistic was used. Among other tests, PS-sign, PS-rank, Sp-sign and Sp-rank tests could yield somewhat competitive results in the case of $d = 30$.

Table 2.2: Observed levels and powers (in %) of two-sample tests

	$\boldsymbol{\Sigma}$	p	Δ	Hotel	Sp	Sp	PS	PS	CQ	SKK	PA	Cram	NN	MST	Adj	WMW		KS	
				T^2	sign	rank	sign	rank					test	run		(S)	(D)	(S)	(D)
Normal	$\mathbf{I}$	30	0.0	5.1	4.2	4.3	6.4	6.7	5.0	3.8	3.7	4.7	4.5	5.2	5.7	2.8	2.3	3.4	3.1
			1.5	98.7	98.4	98.6	77.3	96.8	99.3	99.0	99.9	98.6	75.1	52.9	45.7	96.4	97.6	95.2	96.9
		90	0.0	5.4	4.9	4.5	4.4	6.2	5.7	4.0	3.9	5.3	4.3	5.4	5.1	2.7	3.3	3.8	3.4
			2.0	35.0	37.5	37.2	16.5	36.9	100	100	99.8	100	82.7	59.9	49.7	99.6	99.8	99.3	99.7
	$\boldsymbol{\Sigma}_0$	30	0.0	5.1	4.2	4.2	6.1	6.8	6.6	4.6	5.0	5.3	4.0	3.7	6.1	2.9	3.2	3.7	2.7
			1.5	98.4	97.0	97.5	79.7	96.4	76.7	86.0	82.2	25.5	45.8	33.6	42.8	95.9	97.1	94.0	94.7
		90	0.0	5.5	5.0	4.5	6.0	5.9	8.3	6.3	7.0	5.8	4.8	3.5	5.4	3.1	2.9	4.3	3.6
			2.0	33.9	32.4	33.1	16.8	39.9	37.9	46.0	40.0	13.1	40.7	31.0	35.3	98.8	98.5	98.1	97.6
t (2 df)	$\mathbf{I}$	30	0.0	2.9	4.6	4.6	5.8	6.3	5.2	0.9	4.0	5.0	3.6	3.7	4.9	3.0	2.7	3.2	3.6
			1.5	67.3	87.0	83.7	53.3	80.1	38.2	18.7	29.4	66.5	84.1	62.7	25.9	78.1	80.6	81.6	82.2
		90	0.0	3.8	5.3	5.1	5.2	4.9	5.5	0.1	3.9	4.3	3.7	3.9	5.0	3.7	3.3	3.5	3.9
			2.0	29.3	21.4	28.6	12.6	32.4	34.8	10.1	31.9	64.6	88.6	55.8	21.0	91.7	92.4	92.5	94.1
	$\boldsymbol{\Sigma}_0$	30	0.0	3.0	4.2	4.7	6.9	7.4	6.1	3.5	5.2	5.0	3.4	3.9	3.8	2.8	2.7	3.1	3.3
			1.5	67.4	87.4	84.9	50.2	81.9	14.0	12.4	14.3	23.2	59.5	40.4	25.2	82.3	81.0	83.1	83.0
		90	0.0	3.9	5.0	5.6	5.7	5.3	7.0	3.9	6.3	4.4	3.4	3.7	4.6	3.8	3.5	3.6	3.2
			2.0	29.5	21.3	29.0	11.9	31.8	9.2	7.1	9.6	11.7	49.6	34.0	20.9	94.6	93.9	93.4	94.5
Cauchy	$\mathbf{I}$	30	0.0	1.9	3.9	4.8	5.6	5.4	4.6	0.0	1.0	5.4	5.7	4.9	4.8	1.9	2.4	2.7	3.3
			1.5	30.0	73.3	63.0	39.4	57.7	6.0	0.0	1.2	12.0	80.6	56.7	24.9	59.2	58.9	70.7	69.2
		90	0.0	3.6	4.7	4.1	4.6	5.0	6.3	0.0	1.8	6.0	5.0	5.4	6.5	3.3	2.7	3.2	2.3
			2.0	29.7	17.4	26.7	11.5	24.8	7.6	0.0	1.4	8.4	80.1	63.1	27.5	76.6	75.7	84.0	82.5
	$\boldsymbol{\Sigma}_0$	30	0.0	1.9	4.0	4.7	5.6	5.8	6.8	0.8	2.8	4.8	4.3	4.0	5.5	2.0	2.4	3.7	2.3
			1.5	29.9	72.0	63.4	37.3	60.6	7.9	3.4	1.2	7.7	37.0	25.3	22.9	58.7	56.0	71.2	64.7
		90	0.0	3.7	3.5	3.9	5.5	6.1	7.2	0.6	3.2	5.5	5.7	4.4	5.0	3.4	2.3	3.4	3.0
			2.0	29.2	15.5	24.3	12.8	29.8	8.4	3.6	0.6	7.0	23.7	17.9	20.1	86.2	75.1	89.3	82.6

We carried out our experiment also with elliptically symmetric normal, t_2 and Cauchy distributions, where $\boldsymbol{\Sigma}_0$ was used as the common scatter matrix of the two distributions. The first half of the diagonal elements of $\boldsymbol{\Sigma}_0$ were unity, and the rest were 2. All off-diagonal elements of $\boldsymbol{\Sigma}_0$ were 0.5 except those in the first row and the first column, which were taken to be 0. This choice of $\boldsymbol{\Sigma}_0$ led to the same Mahalanobis distance between the locations of F and G as it was in the case with the common dispersion matrix $\mathbf{I}$. In this set up, in the presence of high correlations among the measurement variables, the power of the CQ test dropped down drastically. Performance of the Cramer test, the MST run test and the NN test also deteriorated substantially. In the example involving 30-dimensional normal distribution, the Hotelling's T^2 test had the highest power, while nonparametric sign and rank tests also performed well. But, in the case of $d = 90$, our proposed tests outperformed all other tests considered here. They outperformed their all competitors in cases of 90-dimensional t_2 and Cauchy distributions as well. In the case of $d = 30$, only the powers of the spatial sign test were marginally higher.

From Table 2.2, it seems to be a good idea to use our proposed tests based on the WMW statistic and the KS statistic when the underlying distributions have light tails and heavy tails, respectively. Even if the underlying distributions are normal, the Hotelling's T^2 statistic should be used only when the dimension is not large compared to the sample size. The CQ test and the Cramer test can yield good performance in high dimensions, but they have poor power properties when the measurement variables are highly correlated and/or the underlying distributions have heavy tails. SKK and PA tests also had similar problems. Nonparametric sign and rank tests and the NN test are good options if the data dimension is not very large.

Tests based on linear functions of multivariate observations have also been proposed by many other authors. We have already discussed about the tests proposed by Wei et al. (2015) and their limitations. Rousson (2002) suggested to use the first principle component direction for linear projection to develop a distribution-free test. Lopes et al. (2011) proposed tests based on several random projections. These tests are applicable even when the dimension of the data exceeds the sample size. We used some high dimensional simulated data sets to compare the performance our proposed tests with these tests based on random projections and principal component direction. Some of the tests considered in Table 2.2, which can be applied to HDLSS data, were also used for comparison. We carried out our experiment with two d-dimensional normal (and Cauchy) distributions having the same scatter matrix $\mathbf{I}_d$ but different location parameters $(0, 0, \ldots, 0)^T$ and $(0.15, 0, 15, \ldots, 0.15)^T$. We considered samples of size 50 from each distribution to perform different tests, and this procedure was repeated 1000 times as before. Observed powers of different tests are shown in Figure 2.2 for different values of d starting from 3 to 600.

In the case of normal distribution, CQ, SKK, PA and Cramer tests had better performance than their competitors. Our proposed tests had powers much higher than rest of the tests considered here. In the case of Cauchy distribution, when CQ, SKK, PA and Cramer tests failed, our proposed tests significantly outperformed all of their competitors. This is consistent with what we observed before. Note that in these examples, separation between the two distributions increases with the dimension. So, one would expect that the power of a test should tend to unity as the dimension increases. But that did not happen for tests based on random projections and principle component directions. In each of these two cases, we used both KS and WMW tests after finding the direction vector, and the best one is reported here. From Figure 2.2, it is quite evident that both of them had miserable performance in high dimensions.

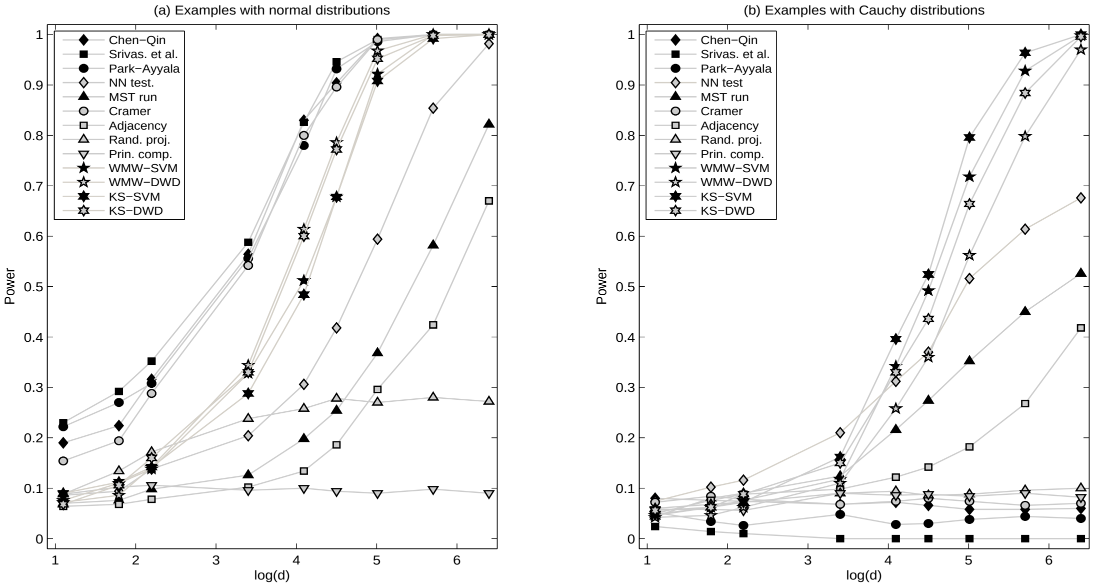


Figure 2.2: Powers of two-sample tests for varying choices of data dimension.

2.5 Results from the analysis of benchmark data sets

We analyzed four benchmark data sets for further evaluation of our proposed methods. Three of them, Sonar data, Arcene data, Hill and valley data, and their descriptions are available at the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/). The Colon data set is available in R package ‘rda’. Description of this microarray gene expression data set can be found in Alon et al. (1999). Several researchers have extensively investigated these data sets, mainly in the context of classification. It is well known that in all these examples, we have reasonable separation between two competing classes. So, in each of these cases, we can assume the alternative hypothesis to be the true, and different tests can be compared on the basis of their power functions. Note that if we use the whole data set for testing, any test will either reject H_0 or accept it. Based on that single experiment, it is difficult to compare among different test procedures. So, in each of these cases, we repeated the experiment 1000 times based on 1000 different subsets chosen from the data (by taking equal number of observations from the two classes), and the results are reported in Table 2.3. For each data set, we report the results for three different choices of the sample (subset) size. Since the dimension of the data was larger than the sample size in most of the cases, here we report the results only for those tests which are applicable in HDLSS situations.

The Sonar data set was used by Gorman and Sejnowski (1988) in their study of classification of sonar signals using a neural network. It contains 111 patterns obtained by bouncing sonar signals off a 'metal cylinder' and 97 patterns obtained from 'rocks' at various angles and under various conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. Signals were obtained from various aspect angles, spanning 90 degrees for cylinder and 180 degrees for rocks. Each number in a 60-dimensional pattern represents the energy within a particular frequency band integrated over a certain period of time, where the integration aperture for higher frequencies occur late. In this data set, the SKK test had the best performance, but our proposed tests outperformed the rest of the tests considered here. In the case of $n = 80$, PA, Cramer, MST run and NN tests had competitive performance, but in cases smaller sample sizes, powers of our proposed tests, especially those based on the WMW statistic, were higher than their competitors.

In the Hill and Valley data set, each record represents 100 points on a two-dimensional graph. When these points are plotted in order (from 1 to 100) as the Y coordinate, they create either a Hill (a bump in the terrain) or a Valley (a dip in the terrain). There are two versions of this data set at the UCI machine learning repository; a noisy version and a noise-free version. Each version has both training and test sets. For our analysis, we considered the training set of the noise-free version consisting of 305 instances of 'Hill' and 301 instances of 'Valley'. In this example, while all other two-sample tests had powers close to the nominal level of 0.05, our proposed tests had excellent performance. In the case of $n = 100$, they rejected H_0 in all cases, and even in the case of $n = 70$, their powers were close to unity. We also analyzed the noisy version of the data, but the results were almost similar. So, here we do not report them.

The Colon data set contains gene expression patterns for two types of cells. Gene expressions in 40 tumor and 22 normal colon tissue samples were analyzed with an Affymetrix oligonucleotide array for more than 6,500 human genes. Out of them, 2000 genes with highest minimal intensity across the samples were chosen, and for each of them, the intensity score was normalized. so that the average intensity across the tissues was 0, and its standard deviation was 1. The data set contains these normalized intensities for 2000 genes for each of these 62 samples (see Alon et al. (1999) for details). Though all these samples were not independent, we considered them to be independent to carry out our analysis. In this data set, the CQ test had the highest power closely followed by SKK and Cramer tests. Our proposed tests, the PA test and the NN test had similar performance, and their powers were much higher than that of MST run and Adjacency tests.

Table 2.3: Observed powers (in %) of two-sample tests in benchmark data sets

	Sonar			Hill & Valley			Colon			Arcene		
	($d = 60$)			($d = 100$)			($d = 2000$)			($d = 10000$)		
Sample size	40	60	80	40	70	100	20	24	30	40	50	60
Chen-Qin	39.8	65.3	86.9	1.3	1.0	0.2	93.9	97.7	100.0	51.1	54.8	67.7
Srivas. et al.	86.8	99.8	100.0	5.6	7.6	7.8	88.6	96.4	100.0	****	****	****
Park-Ayyala	73.2	95.0	99.8	3.8	4.0	6.6	67.0	87.0	99.0	****	****	****
Cramer	43.3	75.6	95.7	3.9	6.7	5.4	89.7	97.0	99.8	34.0	50.0	62.1
NN test	69.9	90.9	99.8	6.7	7.8	5.9	77.1	86.7	96.2	88.1	99.3	100.0
MST run	48.4	84.5	97.3	3.2	5.3	7.4	60.6	68.5	81.9	67.4	86.5	98.2
Adjacency	17.8	26.7	36.1	4.9	7.5	4.1	34.2	44.7	54.9	60.2	74.2	90.5
WMW-SVM	76.1	97.8	100.0	55.0	100.0	100.0	74.4	82.8	96.6	52.5	79.1	99.0
WMW-DWD	73.6	98.4	100.0	43.9	95.1	100.0	75.9	83.3	99.6	47.8	75.7	95.3
KS-SVM	72.9	97.3	100.0	55.5	100.0	100.0	71.2	82.1	95.7	63.0	86.3	94.9
KS-DWD	71.2	98.3	100.0	44.7	95.9	100.0	70.7	82.6	99.0	55.9	83.4	92.4

**** Because of computational problem, these tests could not be used.

The Arcene data set is one of five data sets of the NIPS 2003 feature selection challenge. It was obtained by merging three mass-spectrometry data sets. All data consist of mass-spectra obtained with the SELDI technique. The samples include cancer patients (ovarian or prostate cancer) and healthy patients. There were 7000 original features indicating the abundance of proteins in human sera having a given mass value. In addition to that, 3000 features with no predictive power were added to increase the number of features to 10000 (see Guyon et al. (2007) for details). There were separate training, test and validation sets in the UCI repository. For our analysis, we use random subsets from the training set consisting of 44 cancer patients and 56 healthy patients. In this high dimensional data set, because of computational difficulty, we could not compute the powers of SKK and PA tests over 1000 subsets. The NN test had the highest power in this data set. The MST run test and our proposed tests also had competitive performance. They outperformed the other three tests considered here.

2.6 Tests for high dimensional matched pair data

Instead of having two independent sets of observations from F and G, we may have n observations $\binom{\mathbf{x}_1}{\mathbf{y}_1}, \binom{\mathbf{x}_2}{\mathbf{y}_2}, \ldots, \binom{\mathbf{x}_n}{\mathbf{y}_n}$ from a $2d$-variate distribution with d-dimensional marginals F and G for $\mathbf{X}$ and $\mathbf{Y}$. In such cases, it is a common practice to consider it as a one-sample problem, where $\{\boldsymbol{\xi}_i = \mathbf{x}_i - \mathbf{y}_i;\ i = 1, 2, \ldots, n\}$ are used as sample observations. Note that if we assume $F(\mathbf{x}) = G(\mathbf{x} - \boldsymbol{\theta})$ for all $\mathbf{x}$ and some $\boldsymbol{\theta} \in \mathbb{R}^d$, the distribution of $\boldsymbol{\xi} = \mathbf{X} - \mathbf{Y}$ is symmetric about $\boldsymbol{\theta}$, and testing the equality of the locations of F and G is equivalent to test $H_0 : \boldsymbol{\theta} = 0$. In such situations, one needs to develop multivariate versions of one sample linear rank tests. Here we propose some methods based on linear projection of observations that lead to multivariate generalizations of univariate sign, signed rank and other one-sample linear rank tests, which are distribution-free, and they can be conveniently used even when the dimension of the data is much larger than the sample size.

Here also, we split the whole sample into two subsamples. The first subsample of size m is used to estimate the projection direction. For this estimation, we consider a classification problem between two data clouds $\{\boldsymbol{\xi}_i,\ i = 1, 2 \ldots, m\}$ and $\{\boldsymbol{\eta}_i = -\boldsymbol{\xi}_i,\ i = 1, 2 \ldots, m\}$, and use the SVM or the DWD classifier to find the separating hyperplane. The direction vector perpendicular to this hyperplane is used as $\widehat{\boldsymbol{\beta}}$. Note that if the distribution of $\boldsymbol{\xi}$ is elliptically symmetric (or the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ is elliptically symmetric), the separating hyperplane $\{\mathbf{z} : \beta_0 + \boldsymbol{\beta}^T\mathbf{z} = 0\}$ leads to the best classification between the distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ if and only if the expected values of the one sided univariate sign and signed rank statistics are maximized when the observations are projected along $\boldsymbol{\beta}$ (follows from arguments similar to that used in the proof of Proposition 2.1). Motivations for this classification approach also follow from the interesting result given below. The proof is given in Section 2.9.

PROPOSITION 2.2: *Suppose that* Π *is an elliptically symmetric multivariate distribution having a non-zero location. Then, the sign and the signed rank tests based on linear projections of the data will have the maximum power if and only if the observations are projected along the direction vector of the Bayes discriminating hyperplane associated with the classification problem involving distributions* Π *and* Π^* *with equal prior probabilities, where* $\boldsymbol{\xi} \sim \Pi^*$ *if and only if* $-\boldsymbol{\xi} \sim \Pi$.

After finding $\widehat{\boldsymbol{\beta}}$ using the SVM or the DWD classifier, all $n-m$ observations in the second subsample are projected along $\widehat{\boldsymbol{\beta}}$ to compute the univariate test statistic (i.e., sign or signed rank statistic) and the corresponding test function. This procedure is repeated for several random splits, and the results can be aggregated using either of the three methods discussed in Section 2.2. Following the same argument as used in the proof of Theorem 2.1, one can verify that these proposed tests are distribution-free.

To carry out a theoretical investigation on the power properties of our tests, we assume the observations on $\boldsymbol{\xi}$ to be independent, and we also consider the following regularity conditions on the distribution of $\boldsymbol{\xi} = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(d)})$, which are similar to (A1)-(A3) stated before.

(B1) *Fourth moments of* $\xi^{(q)}$ $(q \geq 1)$ *are uniformly bounded.*

(B2) *Let* $\boldsymbol{\xi}_1$ *and* $\boldsymbol{\xi}_2$ *be two independent copies of* $\boldsymbol{\xi}$. *For* $\mathbf{V} = \boldsymbol{\xi}_2, -\boldsymbol{\xi}_2$ *and* $\mathbf{0}$, $\sum_{q \neq q'} |corr\{({\xi_1}^{(q)} - V^{(q)})^2, ({\xi_1}^{(q')} - V^{(q')})^2\}|$ *is of order* $o(d^2)$.

(B3) *There exist constants* $\sigma_0^2 > 0$ *and* ν_0 *such that (i)* $d^{-1}\sum_{q=1}^{d}[E(\xi^{(q)})]^2 \rightarrow \nu_0^2$ *and (ii)* $d^{-1}\sum_{q=1}^{d} Var(\xi^{(q)}) \rightarrow \sigma_0^2$ *as* $d \rightarrow \infty$.

Note that (A1) and (B1) are equivalent, and under (A3), $\nu_0^2 = \lim_{d\to\infty} d^{-1}\sum_{q=1}^{d}[E(\xi^{(q)})]^2 = \nu^2$. Theorem 2.3 below shows that under (B1)-(B3), the powers of the proposed sign and signed rank tests converge to 1 as d tends to infinity. The proof of this theorem is similar to the proof of Theorem 2.2, and it is given in Section 2.9.

THEOREM 2.3: *Let* $\widehat{\boldsymbol{\beta}}$ *be computed using the SVM or the DWD classifier on the first subsample of size* m, *and the univariate distribution-free test statistic* T *(e.g., sign or* *signed rank statistic) is computed based on observations in the second sample projected along $\widehat{\boldsymbol{\beta}}$. Under H_0, if this test statistic takes its maximum value with probability smaller than α, under the regularity conditions (B1)-(B3), the powers of the Avg-Method and the FDR-Method converge to 1 as d diverges to infinity. Similar result holds for the Bonf-Method if under H_0, the test statistic takes its maximum value with probability smaller than α/M, where M is the number of random splits.*

We carried out simulation studies to compare the level and the power properties of our proposed tests based on sign and signed rank statistics with some existing methods. In particular, we used one-sample versions of Hotelling's T^2, CQ and PA tests, Puri and Sen (1971)'s coordinate-wise sign (PS-sign) and signed rank (PS-rank) tests and tests based on spatial signs (Sp-sign) and ranks (Sp-rank) (see e.g., Möttönen et al. (1997)). For all these tests, we use the same abbreviations as in the two-sample case. The test of Srivastava (2009) (referred to as the SR test) was also used for comparison. For sign and rank tests, we report the results of conditional tests based on permutations since they outperformed corresponding large sample tests. For our proposed methods based on DWD, we used the MATLAB codes as before, but due to singularity of matrices in the regularization path of SVM, the R program 'svmpath' could not be used. Instead we used the SVM toolbox in MATLAB, where the regularization parameter was chosen based on a pilot study.

Again, we considered some examples involving high dimensional (d= 30 and 90) normal, t_2 and Cauchy distributions with $(0, 0, \ldots, 0)^T$ and $(\Delta, 0, \ldots, 0)^T$ as the locations for $\mathbf{X}$ and $\mathbf{Y}$, respectively. To study the level properties of different tests, we used $\Delta = 0$, while for studying their powers properties, we chose the value of Δ depending on the problem (Δ=0.75 and 1.0 for d=30 and 90, respectively) such that most of the competing tests had power appreciably different from 0.05. In all these testing problems, we chose $Var(\mathbf{X}) = Var(\mathbf{Y}) = 0.5\ \mathbf{I} + 0.5\ \mathbf{11}'$ and $Cov(\mathbf{X}, \mathbf{Y}) = 0.5\ \mathbf{11}'$, where $\mathbf{1} = (1, \ldots, 1)'$. We generated 100 observations from the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ to constitute the sample, and each experiment was repeated 1000 times to compute the levels and the powers of different tests, which are reported in Table 2.4.

In the case of normal distribution, the Hotelling's T^2 test had observed levels close to 0.05, but they were slightly lower in cases of t_2 and Cauchy distributions. The SR test also had low levels for t_2 and Cauchy distributions. The CQ test, the PA test and all conditional tests had observed levels close to 0.05 in almost all cases, but for our proposed tests based on FDR, they were substantially lower than the nominal level.

Table 2.4: Observed levels and powers (in %) of paired sample tests

		Hotel.	CQ	SR	PA	Sp.	Sp.	PS	PS	Sign	Sign	S.Rank	S.Rank
	Δ	T^2				sign	rank	sign	rank	SVM	DWD	SVM	DWD
Normal	0.00	4.5	5.9	5.8	5.6	4.3	4.6	4.0	3.8	1.5	1.6	2.7	2.6
$p = 30$	0.75	98.6	99.9	99.8	99.8	98.1	98.3	84.9	96.4	86.4	98.4	92.6	99.2
Normal	0.00	4.9	4.6	3.8	4.2	4.9	5.0	4.7	4.0	2.7	3.0	2.6	2.9
$p = 90$	1.00	41.2	100.0	100.0	100.0	42.8	43.2	35.7	42.2	86.2	99.4	94.2	99.6
t (2 d.f.)	0.00	3.5	4.3	1.2	4.4	4.6	5.2	4.4	4.5	3.2	3.0	2.4	2.8
$p = 30$	0.75	68.5	37.4	19.8	28.7	89.7	84.5	62.8	69.1	80.7	86.7	81.0	85.8
t (2 d.f.)	0.00	3.6	5.4	0.1	4.9	3.8	4.3	4.1	3.9	2.7	2.3	3.1	2.9
$p = 90$	1.00	35.3	39.6	10.0	31.9	26.9	35.6	23.4	32.0	93.2	95.0	91.6	92.3
Cauchy	0.00	2.3	6.0	0.0	5.2	4.4	4.5	4.9	5.1	2.4	2.5	3.2	2.3
$p = 30$	0.75	30.4	7.6	0.0	6.4	76.2	67.6	54.2	55.8	69.6	74.8	58.4	62.0
Cauchy	0.00	3.4	5.5	0.0	4.0	5.9	5.5	4.0	5.9	1.3	1.7	2.2	1.9
$p = 90$	1.00	33.4	8.3	0.0	5.4	14.4	24.2	12.7	22.4	85.2	87.8	77.0	78.2

In terms of power properties, the overall performance of our proposed tests, particularly those based on the DWD classifier, was better than most of the existing methods, especially for $d = 90$. In the case of normal distribution, though CQ, PA and SR tests had better performance than other methods, these three tests did not have satisfactory performance in the examples involving t_2 and Cauchy distributions. For $d = 30$, Sp-sign and Sp-rank tests also had competitive performance, but for $d = 90$, our proposed tests outperformed all other rank based tests considered here. In the case of 90-dimensional normal distribution, while all other rank based tests had powers less than 0.45, our proposed tests based on DWD had powers more than 0.99. We observed similar phenomenon also in the case of 90-dimensional Cauchy distribution, where our proposed tests had much higher powers than their competitors. This is consistent with our findings in Section 2.4.

We also analyzed the Colon data set to investigate the performance of different tests for high dimensional data. In this data set, there were observations on normal as well as cancer tissues from 22 individuals. For each of these tissue samples, there were expression levels for 2000 genes. From these 22 pairs, we chose 18 pairs at random and performed tests based on that. These experiment was repeated 500 times to compute the powers of different tests. Since the dimension of the data set was much larger than the sample size, in this example, along with our proposed tests, we could use CQ, PA and SR tests only. The CQ test had the best performance in this example, it rejected the null hypothesis in 99.8% cases. The SR test and the PA test had powers 0.908 and 0.828 only. Our proposed tests also had comparable performance. While the sign and the signed rank tests based on SVM had powers 0.818 and 0.880, those for the tests based on DWD were 0.826 and 0.894, respectively.

2.7 Large sample properties of proposed tests

So far, we have proved the exact distribution-free property of our proposed tests and shown the convergence of their power functions when the sample size is fixed and the dimension grows to infinity. In this section, we will study their power properties when the sample grows to infinity and the dimension of the data is not large. For studying the large sample properties of our proposed two-sample tests based on SVM and DWD, we assume that as the first subsample size m tends to infinity, m_1/m converges to $1/2$. Otherwise, one has to make some adjustments for the unbalancedness in the data (see e.g., Qiao et al. (2010)). However, if the dimension of the data is not large, especially relative to the sample size, in addition to SVM and DWD, there are many other ways to estimate the projection direction $\boldsymbol{\beta}$. Unlike what happens for high dimensional data, we can use very simple classifiers like Fisher's linear discriminant rule to estimate $\boldsymbol{\beta}$. Alternatively, if $T_{\boldsymbol{\beta}}$ is the univariate KS or WMW statistic computed using the observations in the first subsample projected along $\boldsymbol{\beta}$, one can also find $\widehat{\boldsymbol{\beta}}$ by maximizing $T_{\boldsymbol{\beta}}$ over the set $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\| = 1\}$. In cases of KS and WMW statistics, this maximization leads to linear classifiers based on regression depths and half-space depths, respectively (see, e.g., Ghosh and Chaudhuri (2005)). Clearly, the finite sample distribution-free property established in Theorem 2.2 remains valid irrespective of the classification procedure so long as $\widehat{\boldsymbol{\beta}}$ is computed from one subsample and univariate distribution-free tests are implemented on linear projections of data points in the other subsample. Note that in the case of multivariate two-sample location problem, where $F(\mathbf{x}) = G(\mathbf{x} - \boldsymbol{\theta})$ with $\boldsymbol{\theta} \neq \mathbf{0}$, if $\widehat{\boldsymbol{\beta}} \notin Q = \{\boldsymbol{\beta} : \boldsymbol{\beta}^T\boldsymbol{\theta} = 0\}$, the power of the univariate test (WMW or KS test) applied on the projected observations converges to 1 as the size of the second subsample tends to infinity. For instance, if the distribution of $\widehat{\boldsymbol{\beta}}$ is absolutely continuous, we have consistency of the resulting tests because the set Q has Lebesgue measure zero. Further, even in the case of general alternative $H_A : F \neq G$, if F and G satisfy Carleman condition, (i.e., $E(\|\mathbf{X}\|^r) < \infty \ \forall \ r \geq 1$ and $\sum_{r \geq 1}(E(\|\mathbf{X}\|^r))^{-1/r} = \infty$), the set $Q_0 = \{\boldsymbol{\beta} : F_{\boldsymbol{\beta}} = G_{\boldsymbol{\beta}}\}$ has Lebesgue measure 0 (see Corollary 3.3 in Cuesta-Albertos et al. (2007)), and consequently, the powers of our tests constructed using the KS statistic converge to 1 as the size of the second subsample tends to infinity.

Suppose that F and G are elliptically symmetric, and they differ only in their locations $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. From Proposition 2.1, we know that in this case, KS and WMW tests based on linear projections of the data have the maximum power if and only if the observations are projected along $\boldsymbol{\beta}_* = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, where $\boldsymbol{\Sigma}$ is the common scatter matrix of F and G. Let $\widehat{\boldsymbol{\beta}}_D, \widehat{\boldsymbol{\beta}}_S, \widehat{\boldsymbol{\beta}}_F$ and $\widehat{\boldsymbol{\beta}}_M$ be the estimates of $\boldsymbol{\beta}$ obtained from the first subsample using DWD, SVM, Fisher's linear discrimination and maximization of $T_{\boldsymbol{\beta}}$ (as described above), respectively. Then, the following theorem provides useful insights into asymptotic power properties of the proposed multivariate two-sample tests.

THEOREM 2.4: *For a fixed size of the second subsample, let $\gamma(\boldsymbol{\beta})$ be the power of the univariate KS (or WMW) test when the observations in the second subsample are projected along $\boldsymbol{\beta}$. Define $\gamma_* = \sup_{\boldsymbol{\beta}} \gamma(\boldsymbol{\beta})$. If F and G are elliptically symmetric, and they differ only in their locations, $\gamma(\widehat{\boldsymbol{\beta}}_M)$ converges to γ_* as the first subsample size m tends to infinity. If F and G have finite second moments, we also have this convergence for $\gamma(\widehat{\boldsymbol{\beta}}_F)$, $\gamma(\widehat{\boldsymbol{\beta}}_D)$ and $\gamma(\widehat{\boldsymbol{\beta}}_S)$ when the regularization parameter λ_0 used in SVM is of the order $o(m^{-1/2})$. Under the same set of conditions, the powers of these tests converge to 1 if the sizes of both of the first and the second subsamples tend to infinity.*

In the case of matched pair data, we may consider the situation when the $2d$-dimensional joint distribution of $(\mathbf{X}, \mathbf{Y})$ is elliptically symmetric and the corresponding d-dimensional marginals of $\mathbf{X}$ and $\mathbf{Y}$ (F and G, respectively) differ only in their locations. Otherwise, we may assume the distribution of $\boldsymbol{\xi} = \mathbf{X} - \mathbf{Y}$ to be symmetric about a non-zero location. In that case, we can either find $\widehat{\boldsymbol{\beta}}$ using classification methods like SVM and DWD applied to $\boldsymbol{\xi}$'s and $\boldsymbol{\eta}$'s as described in the previous section. As an alternative, one can also construct $\widehat{\boldsymbol{\beta}}$ using Fisher's linear discrimination or maximization of univariate sign $SG_m(\boldsymbol{\beta}) = \frac{1}{m}\sum_{i=1}^{m} I\{\boldsymbol{\beta}^T(\mathbf{x}_i - \mathbf{y}_i) > 0\}$ or signed rank $SGR_m(\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{j=i+1}^{m} I\{\boldsymbol{\beta}^T(\mathbf{x}_i - \mathbf{y}_i) + \boldsymbol{\beta}^T(\mathbf{x}_j - \mathbf{y}_j) > 0\}/\binom{m}{2}$ statistic. Here $I\{\cdot\}$ denotes the indicator function. Results, which are analogous to Theorem 2.4, concerning asymptotic powers of multivariate paired sample tests constructed using sign and signed rank statistics based on data points projected along $\widehat{\boldsymbol{\beta}}$ can be derived under appropriate conditions. We omit the mathematical details as those details are very similar to those in two-sample problems (see our comments at the end of Section 2.9).

2.8 Tests based on real valued functions of the data

Recall now the statement of Theorem 2.1 and discussion preceding the theorem. It is straight forward to verify that the distribution-free property asserted in Theorem 2.1 remains valid if the statistic T is computed based on any real valued function h of the data corresponding to second subsample, where such a h may be chosen based on the first subsample. Linearity of h is not required for Theorem 2.1 to hold. Consequently, if one constructs a nonlinear classifier based on the first subsample and use the corresponding discriminant function to form the test statistic based on the second subsample, one can get a distribution-free test with power properties depending on the choice of the discriminant function. If the distributions of two multivariate samples are elliptic and unimodal differing only in their location, the optimal Bayes classifier discriminating between the two distributions happens to be linear when the prior probabilities are equal. So, in such cases, it is reasonable to construct tests based on only linear functions of the data. Also, if F and G both belong to the exponential family, the Bayes classifier turns out to be a linear function of the sufficient statistics. So, after finding the sufficient statistics for that family, the same method based on linear projection can be used there as well. However, in more general situations, the Bayes classifier may not be a linear function of the data, and there it is more appropriate to consider tests based on suitable nonlinear functions of the data.

PROPOSITION 2.3: *Suppose that h is a real valued measurable transformation of d-dimensional observations, and it is chosen from the first subsample. Consider a univariate rank statistic T (e.g., KS or WMW statistic), which is computed on the transformed* *observations in the second subsample. Then the resulting multivariate two-sample test has the distribution-free property. Define $\gamma_0(h)$ as the power of the univariate test when it is implemented on the observations transformed using the transformation function h. If f and g are density function corresponding to the two distributions F and G, respectively, $\gamma_0(h)$ is maximized when $h(\cdot) = g(\cdot)/f(\cdot)$, which is the likelihood ratio.*

Therefore, in practice, one can construct consistent estimates $\hat{f}$ and $\hat{g}$ for f and g from the first subsample, and transform the observations in the second subsample using the transformation function $\widehat{\mathcal{T}}(\cdot) = \hat{g}(\cdot)/\hat{f}(\cdot)$. Kernel density estimates (see e.g., Silverman (1986); Scott (2015)) or nearest neighbor density estimates (see e.g., Loftsgaarden and Quesenberry (1965)) can be used for this purpose. Results analogous to Theorem 2.4 can be proved for these transformations as well. However, nonparametric estimation of f and g makes the convergence of the estimates rather slow, especially in high dimension. Another option is to use nonlinear SVM classifier based on radial basis or other basis functions (see e.g., Burges (1998)). Note that these nonlinear SVM classifiers can be used even when the dimension is larger than the sample size.

2.9 Proofs and mathematical details

PROOF OF PROPOSITION 2.1: Without loss of generality, let us assume that F and G have locations $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}$, respectively. Also, it is enough to consider only those $\boldsymbol{\beta}$'s for which $\boldsymbol{\beta}^T\boldsymbol{\mu} > 0$ and $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} = 1$, where $\boldsymbol{\Sigma}$ is the common scatter matrix of F and G. For all such choices of $\boldsymbol{\beta}$, the distribution of $\boldsymbol{\beta}^T\mathbf{X}$ remains the same with location 0 and scatter 1. The distribution $\boldsymbol{\beta}^T\mathbf{Y}$ also remains the same except for its location $\boldsymbol{\beta}^T\boldsymbol{\mu} > 0$. Now, consider two direction vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ such that $\boldsymbol{\beta}_1^T\boldsymbol{\mu} > \boldsymbol{\beta}_2^T\boldsymbol{\mu} > 0$. Clearly, this implies that ${\boldsymbol{\beta}_1}^T\mathbf{Y}$ is stochastically larger than ${\boldsymbol{\beta}_2}^T\mathbf{Y}$. Consequently, the ranks of the corresponding linear functions of the observations have a similar stochastic ordering. Hence, the powers of the one sided KS test and the one sided WMW test are higher if the data are projected along $\boldsymbol{\beta}_1$. Therefore, in order to maximize the power of any such test, one needs to maximize $\boldsymbol{\beta}^T\boldsymbol{\mu}$ subject to $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} = 1$. Since $(\boldsymbol{\beta}^T\boldsymbol{\mu})^2 \leq (\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})(\boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$, this maximum is achieved when $\boldsymbol{\beta}$ is a positive scalar multiple of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, the direction vector corresponding to the Bayes classifier. □

PROOF OF THEOREM 2.1: For any split of the data into two independent subsamples, $\widehat{\boldsymbol{\beta}}$ is independent of the data points from which $T_{\widehat{\boldsymbol{\beta}}}$ and $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})$ are computed. As a consequence, for any given $\widehat{\boldsymbol{\beta}}$, the conditional size of the test, which is same as the conditional expectation of the test function $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})$, is α for any (F, G) such that $F = G$, in view of the distribution-free property of the test statistic $T_{\boldsymbol{\beta}}$ for any fixed $\boldsymbol{\beta}$. Since this conditional expectation does not depend on $\widehat{\boldsymbol{\beta}}$ nor on the specific split involved, we must have $E_{H_0}(\phi_\alpha) = \alpha$. □

PROOF OF PROPOSITION 2.2: If Π is elliptically symmetric with a non-zero location, so is Π^*, and it differs from Π only in its location. So, the result can be proved using arguments based on stochastic ordering as in the proof of Proposition 2.1. □

PROOF OF THEOREM 2.2: Under (A1)-(A3), as $d \to \infty$, $\|\mathbf{x}_i - \mathbf{x}_j\|/\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2}$ for $1 \leq i < j \leq n_1$, $\|\mathbf{y}_i - \mathbf{y}_j\|/\sqrt{d} \xrightarrow{P} \sigma_2\sqrt{2}$ for $1 \leq i < j \leq n_2$, and $\|\mathbf{x}_i - \mathbf{y}_j\|/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2}$ for $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$. So, after re-scaling, m_1 observations from F tend to lie on the vertices of a regular simplex $\mathcal{S}_1$ and m_2 observations from G tend to lie on the vertices of another regular simplex $\mathcal{S}_2$, while each vertex of $\mathcal{S}_1$ is equidistant from all vertices of $\mathcal{S}_2$ and vice versa. Because of this symmetric nature of data geometry, the discriminating surface constructed by SVM applied to the subsamples with sizes m_1 and m_2 consisting of observations on $\mathbf{X}$ and $\mathbf{Y}$ bisects each of the $m_1 m_2$ lines joining the vertices of $\mathcal{S}_1$ and $\mathcal{S}_2$. So, if $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the means of these m_1 and m_2 observations on $\mathbf{X}$ and $\mathbf{Y}$, and $\widehat{\boldsymbol{\beta}}$ is the projection direction estimated by SVM, $\widehat{\boldsymbol{\beta}}$ tends to be proportional to $\bar{\mathbf{y}} - \bar{\mathbf{x}}$ in the sense that $\left\| \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|} - \frac{\bar{\mathbf{y}}-\bar{\mathbf{x}}}{\|\bar{\mathbf{y}}-\bar{\mathbf{x}}\|} \right\| \xrightarrow{P} 0$ as $d \to \infty$. Similar results hold if $\boldsymbol{\beta}$ is estimated using the DWD classifier as well (see proofs of Theorems 1 and 2 in Hall et al. (2005)). So, both for SVM and DWD, the linear transformations $\mathbf{z} \to \widehat{\boldsymbol{\beta}}^T\mathbf{z}$ and $\mathbf{z} \to (\bar{\mathbf{y}} - \bar{\mathbf{x}})^T\mathbf{z}$ asymptotically (as $d \to \infty$) lead to the same ranking among the $n - m$ projected observations of the second subsample. Since $\mathbf{z}_1^T(\bar{\mathbf{y}} - \bar{\mathbf{x}}) > \mathbf{z}_2^T(\bar{\mathbf{y}} - \bar{\mathbf{z}}) \Leftrightarrow \|\mathbf{z}_1 - \bar{\mathbf{x}}\|^2 - \|\mathbf{z}_1 - \bar{\mathbf{y}}\|^2 > \|\mathbf{z}_2 - \bar{\mathbf{x}}\|^2 - \|\mathbf{z}_2 - \bar{\mathbf{y}}\|^2$, the transformation $\mathbf{z} \to \|\mathbf{z} - \bar{\mathbf{x}}\|^2 - \|\mathbf{z} - \bar{\mathbf{y}}\|^2$ also leads to the same ranking.

Now, for any $\mathbf{x}_i$ from the subsample of size m_1, $d^{-1}(\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 - \|\mathbf{x}_i - \bar{\mathbf{y}}\|^2) \xrightarrow{P} -(\sigma_1^2/m_1 + \sigma_2^2/m_2 + \nu^2)$, and for any $\mathbf{y}_i$ from the subsample of size m_2, $d^{-1}(\|\mathbf{y}_i - \bar{\mathbf{x}}\|^2 - \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2) \xrightarrow{P} (\sigma_1^2/m_1 + \sigma_2^2/m_2 + \nu^2)$. Therefore, after finding $\widehat{\boldsymbol{\beta}}$ using the SVM classifier or the DWD classifier, we consider the one-sided alternative that suggests $G_{\widehat{\boldsymbol{\beta}}}$ to be stochastically larger than $F_{\widehat{\boldsymbol{\beta}}}$.

Now $T_{\widehat{\boldsymbol{\beta}}}$ (e.g., the one sided KS statistic or the one sided WMW statistic) is computed from two subsamples consisting of $n_1 - m_1$ observations on $\mathbf{X}$ and $n_2 - m_2$ observations on $\mathbf{Y}$ using the ranks of those observations projected along $\widehat{\boldsymbol{\beta}}$, where ranking is done after combining the two subsamples. Now, for each $\mathbf{x}$ from these first $n_1 - m_1$ observations, we have $d^{-1}(\|\mathbf{x}-\bar{\mathbf{x}}\|^2 - \|\mathbf{x}-\bar{\mathbf{y}}\|^2) \xrightarrow{P} (\sigma_1^2/m_1 - \sigma_2^2/m_2) - \nu^2$ and for each $\mathbf{y}$ from the next $n_2 - m_2$ observations, we have $d^{-1}(\|\mathbf{y}-\bar{\mathbf{x}}\|^2 - \|\mathbf{y}-\bar{\mathbf{y}}\|^2) \xrightarrow{P} (\sigma_1^2/m_1 - \sigma_2^2/m_2) + \nu^2$. So, when $\nu^2 > 0$, $T_{\widehat{\boldsymbol{\beta}}}$ attains its maximum value $t_{\max}$. If $P_{H_0}(T_{\widehat{\boldsymbol{\beta}}} = t_{\max}) < \alpha$, the limiting p-value (as $d \to \infty$) becomes smaller than α for each of the M partitions (where M is assumed to be finite), and hence the test based on the FDR-Method rejects H_0 with probability tending to one as d tends to infinity. Also, $T_{\widehat{\boldsymbol{\beta}}} \xrightarrow{P} t_{\max}$ and $P_{H_0}(T_{\widehat{\boldsymbol{\beta}}} = t_{\max}) < \alpha$ imply that $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}}) \xrightarrow{P} 1$. Since any test function is bounded, this proves that $E_{H_A}\{\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})\} \to 1$. Consequently the power of the test based on the Avg-Method, $E_{H_A}(\phi^*_\alpha)$, converges to unity as $d \to \infty$. If $P_{H_0}(T_{\widehat{\boldsymbol{\beta}}} = t_{\max}) < \alpha/M$, the null hypothesis is rejected by a test of level α/M. So, the power of the test based on the Bonf-Method also converges to 1 as d tends to infinity. □

PROOF OF THEOREM 2.3: Under (B1) and (B2), WLLN holds for the sequence of $\{\xi_1^{(q)2}, q \geq 1\}$, $\{(\xi_1^{(q)} - \xi_2^{(q)})^2, q \geq 1\}$ and $\{(\xi_1^{(q)} + \xi_2^{(q)})^2, q \geq 1\}$. So, using (B1)-(B3), one can show that for all $i = 1, \ldots, m$, $d^{-1/2}\|\boldsymbol{\xi}_i\| = d^{-1/2}\|\boldsymbol{\eta}_i\| \xrightarrow{P} (\nu_0^2 + \sigma_0^2)^{1/2}$ as $d \to \infty$ (here $\boldsymbol{\eta}_i = -\boldsymbol{\xi}_i$ for all i). This implies that $d^{-1/2}\|\boldsymbol{\xi}_i - \boldsymbol{\eta}_i\| \xrightarrow{P} 2(\nu_0^2 + \sigma_0^2)^{1/2} = \kappa_1$ (say) for $i = 1, 2, \ldots, m$. Again, for any $i \neq j$, we have $d^{-1/2}\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\| = d^{-1/2}\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_j\| \xrightarrow{P} (2\sigma_0^2)^{1/2}$ and $d^{-1/2}\|\boldsymbol{\xi}_i - \boldsymbol{\eta}_j\| \xrightarrow{P} (2\sigma_0^2 + 4\nu_0^2)^{1/2} = \kappa_2$ (say). So, as d tends to infinity, after re-scaling by a factor of $d^{-1/2}$, $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_m$ tend to lie on the vertices of a regular m-simplex $\mathcal{S}_1^\circ$ and $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_m$ tend to lie on the the vertices of another regular m-simplex $\mathcal{S}_2^\circ$, which is obtained from $\mathcal{S}_1^\circ$ by reflecting it along all coordinate axes.

Now, from any vertex $\boldsymbol{\xi}_i$ on $\mathcal{S}_1^\circ$, the distances of the $(m-1)$ vertices of $\mathcal{S}_2^\circ$ are equal to κ_2 and that of one other vertex $\boldsymbol{\eta}_i$ is κ_1. This is easy to visualize for $m = 2$ (see Figure 2.3). Note that $\kappa_2 < \kappa_1$. Naturally, SVM chooses the hyperplane that passes through the origin and bisects each of these lines of length κ_2.

Now, consider a new observation $\boldsymbol{\xi}_0$. As d tends to infinity, after rescaling by a factor of $d^{-1/2}$, it tends to be equidistant from all vertices of $\mathcal{S}_1^\circ$, and that common distance is $(2\sigma_0^2)^{1/2}$. So, its squared distance from the centroid of the first set of observations is given by $2\sigma_0^2 - \sigma_0^2(1-m^{-1})$. Similarly, its (re-scaled) distances from all vertices of $\mathcal{S}_2^\circ$ tend to be κ_1. Hence, its squared distance from the centroid of $\mathcal{S}_2^\circ$ turns out to be $\kappa_2^2 - \sigma_0^2(1-m^{-1})$. So, SVM correctly classifies $\boldsymbol{\xi}_0$ if $2\sigma_0^2 < \kappa_1^2$ i.e. $\nu_0^2 > 0$. Also note that here we have same number of observations in each of the two classes. So, SVM and DWD have the same limiting behavior as $d \to \infty$ (see Hall et al. (2005)).

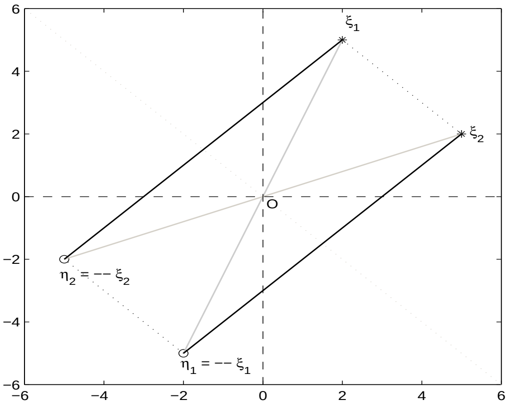


Figure 2.3: Geometry of high dimensional data.

Let $\widehat{\boldsymbol{\beta}}$ be the direction perpendicular to the separating hyperplane chosen by SVM or DWD. From the above discussion it is clear that $P\{\widehat{\boldsymbol{\beta}}^T\boldsymbol{\xi} > 0\} = P\{\widehat{\boldsymbol{\beta}}^T(\mathbf{X}-\mathbf{Y}) > 0\} \to 1$ as $d \to \infty$. Now, using the same argument as used in the proof of Theorem 2.2, we can show that the powers of sign and signed rank tests converge to 1 as d increases. □

LEMMA 2.1: Let $\mathbf{x}_1, \ldots, \mathbf{x}_{m_1} \overset{i.i.d.}{\sim} F$ and $\mathbf{y}_1, \ldots, \mathbf{y}_{m_2} \overset{i.i.d.}{\sim} G$. Define the WMW statistic $\mathcal{U}_{m_1,m_2}(\boldsymbol{\beta}) = \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} I\{\boldsymbol{\beta}^T(\mathbf{x}_i - \mathbf{y}_j) > 0\}$, the KS statistic $\mathcal{K}_{m_1,m_2}(\boldsymbol{\beta}) = \sup_{\beta_0} |\frac{1}{m_1}\sum_{i=1}^{m_1} I\{\boldsymbol{\beta}^T\mathbf{x}_i \leq \beta_0\} - \frac{1}{m_2}\sum_{i=1}^{m_2} I\{\boldsymbol{\beta}^T\mathbf{y}_i \leq \beta_0\}|$, and their population analogs $\mathcal{U}(\boldsymbol{\beta}) = P(\boldsymbol{\beta}^T(\mathbf{X}-\mathbf{Y}) > 0)$ and $\mathcal{K}(\boldsymbol{\beta}) = \sup_{\beta_0} |F_{\boldsymbol{\beta}}(\beta_0) - G_{\boldsymbol{\beta}}(\beta_0)|$, where $\mathbf{X} \sim F$ and $\mathbf{Y} \sim G$. As $\min\{m_1, m_2\}$ tends to infinity, $\sup_{\boldsymbol{\beta}} |\mathcal{U}_{m_1,m_2}(\boldsymbol{\beta}) - \mathcal{U}(\boldsymbol{\beta})|$ and $\sup_{\boldsymbol{\beta}} |\mathcal{K}_{m_1,m_2}(\boldsymbol{\beta}) - \mathcal{K}(\boldsymbol{\beta})|$ both converge to 0 almost surely.

PROOF: Using the arguments based on Hoeffding's inequality (for U-statistic) and Vapnik-Chervonenkis (VC) dimension (see e.g., Vapnik (1998)), we can show the almost sure convergence of $\sup_{\boldsymbol{\beta}} |\mathcal{U}_{m_1,m_2}(\boldsymbol{\beta}) - \mathcal{U}(\boldsymbol{\beta})|$ to 0 (see Theorem 3.1(i) in Ghosh and Chaudhuri (2005) for details) as $\min\{m_1, m_2\} \to \infty$.

Now, define $\mathcal{E}_{m_1,m_2}(\beta_0, \boldsymbol{\beta}) = \frac{1}{m_1}\sum_{i=1}^{m_1} I\{\boldsymbol{\beta}^T \mathbf{x}_i < \beta_0\} + \frac{1}{m_2}\sum_{i=1}^{m_2} I\{\boldsymbol{\beta}^T \mathbf{y}_i \geq \beta_0\}$ and $\mathcal{E}(\beta_0, \boldsymbol{\beta}) = P(\boldsymbol{\beta}^T \mathbf{X} < \beta_0) + P(\boldsymbol{\beta}^T \mathbf{Y} \geq \beta_0)$. Using the arguments based on Hoeffding's inequality and VC dimension, we have $\sup_{\beta_0,\boldsymbol{\beta}} |\mathcal{E}_{m_1,m_2}(\beta_0, \boldsymbol{\beta}) - \mathcal{E}(\beta_0, \boldsymbol{\beta})| \xrightarrow{a.s.} 0$. (see Theorem 3.1(ii) in Ghosh and Chaudhuri (2005)). One can show that $\sup_{\beta_0} \mathcal{E}(\beta_0, \boldsymbol{\beta}) = 1 - \mathcal{K}(\boldsymbol{\beta})$ and $\sup_{\beta_0} \mathcal{E}_{m_1,m_2}(\beta_0, \boldsymbol{\beta}) = 1 - \mathcal{K}_{m_1,m_2}(\boldsymbol{\beta})$. So, $\sup_{\boldsymbol{\beta}} |\mathcal{K}_{m_1,m_2}(\boldsymbol{\beta}) - \mathcal{K}(\boldsymbol{\beta})| = \sup_{\beta_0,\boldsymbol{\beta}} |\mathcal{E}_{m_1,m_2}(\beta_0, \boldsymbol{\beta}) - \mathcal{E}(\beta_0, \boldsymbol{\beta})| \xrightarrow{a.s.} 0$ as $\min\{m_1, m_2\} \to \infty$. □

REMARK: Lemma 2.1 holds even when d increases with the sample size at the rate of $\min\{m_1, m_2\}^t$ for some $t \in (0, 1)$ (see Section 3 in Ghosh and Chaudhuri (2005)).

LEMMA 2.2: Consider the objective function $D_m(\beta_0, \boldsymbol{\beta})$ used in DWD classification as discussed in Section 2.1. Suppose that m_1 (respectively, m_2) out of m observations are from F (respectively, G), and m_1/m tends to 1/2 as $m \to \infty$. Define $D(\beta_0, \boldsymbol{\beta}) = 0.5\ \{E[V_0(\beta_0 + \boldsymbol{\beta}^T \mathbf{X})] + E[V_0(-\beta_0 - \boldsymbol{\beta}^T \mathbf{Y})]\}$, where V_0 is defined as in Section 2.1. Let $(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D)$ be a minimizer of $D_m(\beta_0, \boldsymbol{\beta})$, and $(\beta_0^D, \boldsymbol{\beta}^D)$ be the unique minimizer of $D(\beta_0, \boldsymbol{\beta})$. If F and G have finite second moments, $\widehat{\boldsymbol{\beta}}_m^D$ converges to $\boldsymbol{\beta}^D$ almost surely as m tends to infinity.

PROOF: Note that $D_m(\beta_0, \boldsymbol{\beta})$ can be expressed as $D_m(\beta_0, \boldsymbol{\beta}) = \frac{m_1}{m}\frac{1}{m_1}\sum_{i=1}^{m_1} V_0(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) + \frac{m_2}{m}\frac{1}{m_2}\sum_{i=1}^{m_2} V_0(-\beta_0 - \boldsymbol{\beta}^T \mathbf{y}_i)$. For any fixed β_0 and $\boldsymbol{\beta}$, $V_0(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)$ $(i = 1, 2, \ldots, m_1)$ are i.i.d. bounded random variables. So, using Hoeffding's inequality, we can find a constant A_0 such that for every $\epsilon > 0$, $P\{|\frac{1}{m_1}\sum_{i=1}^{m_1} V_0(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - E[V_0(\beta_0 + \boldsymbol{\beta}^T \mathbf{X})]| > \epsilon\} < 2e^{-A_0 m_1 \epsilon^2}$. Since V_0 is Lipschitz continuous, for any $\boldsymbol{\beta}_{+1} = \binom{\beta_{01}}{\boldsymbol{\beta}_1}$ and $\boldsymbol{\beta}_{+2} = \binom{\beta_{02}}{\boldsymbol{\beta}_2}$, we have $|V_0(\beta_{01} + \boldsymbol{\beta}_1^T \mathbf{x}) - V_0(\beta_{02} + \boldsymbol{\beta}_2^T \mathbf{x})| \leq \|\mathbf{x}\|\ \|\boldsymbol{\beta}_{+1} - \boldsymbol{\beta}_{+2}\|$. Therefore, under the existence of second moments of F and G, following Theorem 19.4 and Example 19.7 of Van der Vaart (2000, pp. 270-71), one can show that the class of functions $\{V_0(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}) :\ (\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^d\}$ has finite VC dimension d_v (say). So, using the results on probability inequalities (see e.g., Devroye et al. (1996); Van der Vaart (2000)), we get

$$P\{\sup_{\beta_0,\boldsymbol{\beta}} \Big|\frac{1}{m_1}\sum_{i=1}^{m_1} V_0(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - E[V_0(\beta_0 + \boldsymbol{\beta}^T \mathbf{X})]\Big| > \epsilon\} < 2m_1^{d_v} e^{-A_0 m \epsilon^2}.$$

Since $\sum_{m_1 \geq 1} m_1^{d_v} e^{-A_0 m_1 \epsilon^2} < \infty$, using the Borel-Cantelli Lemma, one can show that $\sup_{\beta_0, \boldsymbol{\beta}} |\frac{1}{m_1} \sum_{i=1}^{m_1} V_0(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - E[V_0(\beta_0 + \boldsymbol{\beta}^T \mathbf{X})]| \xrightarrow{a.s} 0$ as $m_1 \to \infty$. Similarly, we have $\sup_{\beta_0, \boldsymbol{\beta}} |\frac{1}{m_2} \sum_{i=1}^{m_2} V_0(-\beta_0 - \boldsymbol{\beta}^T \mathbf{y}_i) - E[V_0(-\beta_0 - \boldsymbol{\beta}^T \mathbf{Y})]| \xrightarrow{a.s} 0$ as $m_2 \to \infty$. Now, m_1/m and m_2/m both converge to 1/2 as $m \to \infty$. So, $\sup_{\beta_0, \boldsymbol{\beta}} |D_m(\beta_0, \boldsymbol{\beta}) - D(\beta_0, \boldsymbol{\beta})| \xrightarrow{a.s} 0$ as $m \to \infty$.

Now, from the definition of $(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D)$ and $(\beta_0^D, \boldsymbol{\beta}^D)$, it is quite transparent that $|D_m(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D) - D((\beta_0^D, \boldsymbol{\beta}^D)| \leq \sup_{\beta_0, \boldsymbol{\beta}} |D_m(\beta_0, \boldsymbol{\beta}) - D(\beta_0, \boldsymbol{\beta})| \xrightarrow{a.s} 0$ as $m \to \infty$. Again, we have $D_m(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D) \leq D_m((\beta_0^D, \boldsymbol{\beta}^D)$ and $D(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D) \geq D((\beta_0^D, \boldsymbol{\beta}^D)$ for all m. This implies that $|D(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D) - D((\beta_0^D, \boldsymbol{\beta}^D)|$ converges to 0 on a set of probability one. Now, on the same set, if $(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D)$ converges, it has to converge to $(\beta_0^D, \boldsymbol{\beta}^D)$ in view of uniqueness of $(\beta_0^D, \boldsymbol{\beta}^D)$ and the continuity of the function $D(\beta_0, \boldsymbol{\beta})$. Here without loss of generality, we can assume that for all m, $(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D)$ lies in the compact surface of the unit ball in R^{d+1}. So, any subsequence of the sequence of these estimates has a convergent subsequence converging to $(\beta_0^D, \boldsymbol{\beta}^D)$ on that set of probability one. Hence, $(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D)$ also converges to $(\beta_0^D, \boldsymbol{\beta}^D)$ almost surely. □

LEMMA 2.3: *If F and G are elliptically symmetric, and they differ only in their location, $\widehat{\boldsymbol{\beta}}_m^M$ converges almost surely to a constant multiple of $\boldsymbol{\beta}_*$ (defined in Section 2.7) as m tends to infinity. If F and G have finite second moments, we also have this almost sure convergence for $\widehat{\boldsymbol{\beta}}_m^F$ and $\widehat{\boldsymbol{\beta}}_m^D$ and probability convergence for $\widehat{\boldsymbol{\beta}}_m^S$ when λ_0, the regularization parameter used in SVM, is of the order $o(m^{-1/2})$.*

PROOF: If F and G are elliptically symmetric and they differ in only in their locations, the Bayes discriminant function is linear with direction vector proportional to $\boldsymbol{\beta}_*$. Since $\boldsymbol{\beta}_*/\|\boldsymbol{\beta}_*\|$ is the unique maximizer of $\mathcal{U}(\boldsymbol{\beta})$ and $\mathcal{K}(\boldsymbol{\beta})$ (see Proposition 2.1 and note that we maximize $\mathcal{U}(\boldsymbol{\beta})$ and $\mathcal{K}(\boldsymbol{\beta})$ over $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| = 1$), from Lemma 2.1, we have $\widehat{\boldsymbol{\beta}}_m^M \xrightarrow{a.s} \boldsymbol{\beta}_*/\|\boldsymbol{\beta}_*\|$ both for WMW and KS statistics.

From Fisher consistency (see Qiao et al. (2010)) of DWD classifier, we have $\boldsymbol{\beta}^D \propto \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^D$ is as defined in Lemma 2.2. So, if F and G have finite second moments, the almost sure convergence of $\widehat{\boldsymbol{\beta}}_m^D$ to a constant multiple of $\boldsymbol{\beta}_*$ follows from Lemma 2.2.

The Fisher discriminant function computed from the data is given by $\widehat{\boldsymbol{\beta}}_m^F = \hat{\boldsymbol{\Sigma}}^{-1}(\mathbb{M}_1 - \mathbb{M}_2)$, where $\mathbb{M}_1$ and $\mathbb{M}_2$ are sample means for $\mathbf{X}$ and $\mathbf{Y}$, and $\hat{\boldsymbol{\Sigma}}$ is the moment based estimate of the pooled dispersion matrix. Now, under the assumption of existence of second order moments of F and G, we have $\widehat{\boldsymbol{\beta}}_m^F \xrightarrow{a.s.} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{\beta}_*$.

Now, consider the case of SVM. First note that if F and G have disjoint supports, there is nothing to prove. So, we assume that they have some overlapping regions. Now, if F and G have finite second moments, they satisfy the assumptions (A1)-(A4) of Koo et al. (2008). So, if λ_0 is of the order $o(m^{-1/2})$, $\widehat{\boldsymbol{\beta}}_m^S$, the minimizer of $S_m(\beta_0, \boldsymbol{\beta})$ converges (in probability) to $\boldsymbol{\beta}^S$, the minimizer of $S(\beta_0, \boldsymbol{\beta}) = 0.5(E[1-(\beta_0+\boldsymbol{\beta}^T\mathbf{X})_+] + E[1-(-\beta_0-\boldsymbol{\beta}^T\mathbf{Y})_+])$ (follows from Theorem 1 of Koo et al. (2008)). Now, due to Fisher consistency of SVM (see e.g., Lin (2002)), we have $\boldsymbol{\beta}^S$ proportional to $\boldsymbol{\beta}^*$. □

PROOF OF THEOREM 2.4: If we can show the continuity of $\gamma(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, we can say that $\exists\, \boldsymbol{\beta}_*$ such that $\gamma_* = \gamma(\boldsymbol{\beta}_*)$. Then, the convergence of the power function to γ_* follows from Lemma 2.3. Consider any fixed sequence $\{\boldsymbol{\beta}_m, m \geq 1\}$ that converges to $\boldsymbol{\beta}$. So, for any fixed size of the second subsample, $(\boldsymbol{\beta}_m^T\mathbf{x}_1, \ldots, \boldsymbol{\beta}_m^T\mathbf{x}_{n_1-m_1}, \boldsymbol{\beta}_m^T\mathbf{y}_1, \ldots, \boldsymbol{\beta}_m^T\mathbf{y}_{n_2-m_2})$ converges to $(\boldsymbol{\beta}^T\mathbf{x}_1, \ldots, \boldsymbol{\beta}^T\mathbf{x}_{n_1-m_1}, \boldsymbol{\beta}^T\mathbf{y}_1, \ldots, \boldsymbol{\beta}^T\mathbf{y}_{n_2-m_2})$ almost surely and hence in distribution. Now, note that $Q_{\boldsymbol{\beta}} = \{(\mathbf{x}, \mathbf{y}) : \boldsymbol{\beta}^T(\mathbf{x}-\mathbf{y}) > 0\}$ is an open set in $\mathbb{R}^d$ with boundary having probability measure zero. Also, for any fixed $(\mathbf{x}, \mathbf{y})$, the set $Q^{\mathbf{x},\mathbf{y}} = \{\boldsymbol{\beta} : \boldsymbol{\beta}^T(\mathbf{x}-\mathbf{y}) > 0\}$ is open in $\mathbb{R}^d$. Since $\boldsymbol{\beta}_m^T(\mathbf{x}-\mathbf{y}) \to \boldsymbol{\beta}^T(\mathbf{x}-\mathbf{y})$, for any $(\mathbf{x}, \mathbf{y}) \in Q_{\boldsymbol{\beta}}$, we have $\boldsymbol{\beta}_m^T(\mathbf{x}-\mathbf{y}) > 0$ for sufficiently large m. If $T_{\boldsymbol{\beta}}$ denotes the value of a univariate rank statistic (e.g., the KS statistic or the WMW statistic) computed using the observations projected along $\boldsymbol{\beta}$, the event $\{T_{\boldsymbol{\beta}} = r\}$ can be expressed in terms of finite unions and intersections of the sets $Q_{\boldsymbol{\beta}}^{ij} = \{(\mathbf{x}_i, \mathbf{y}_j) : \boldsymbol{\beta}^T(\mathbf{x}_i - \mathbf{y}_j) > 0\}$; $1 \leq i \leq n_1 - m_1$, $1 \leq j \leq n_2 - m_2$. So, $P\{T_{\boldsymbol{\beta}_m} = r\} \to P\{T_{\boldsymbol{\beta}} = r\}$ for all r, and hence we have the continuity of $\gamma(\boldsymbol{\beta})$.

Since $\boldsymbol{\beta}_m$ converges to $\boldsymbol{\beta}_*$, and $\boldsymbol{\beta}_*^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \neq 0$, for any $\epsilon > 0$, there exist an integer M_0 such that for all $m \geq M_0$, $P(\boldsymbol{\beta}_m \notin Q) > 1 - \epsilon$, where $Q = \{\boldsymbol{\beta} :\ \boldsymbol{\beta}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0\}$. Now, if $\boldsymbol{\beta}_m \notin Q$, powers of the tests based on WMW and KS statistics converge to 1 as the size of the second subsample tends to infinity. So, when sizes of both the first and the second subsamples tend to infinity, powers of these tests convergence to 1. □

PROOF OF PROPOSITION 2.3: The distribution-free property of the resulting test follows from the arguments used in the proof of Theorem 2.1.

Assume that $\mathbf{X} \sim F$, $\mathbf{Y} \sim G$ and correspondingly $h(\mathbf{X}) \sim F_h$ and $h(\mathbf{Y}) \sim G_h$. Since the most powerful tests are unbiased, we have G_h stochastically larger than F_h. Now, consider any other transformation $h'(.)$, and let us assume that $h'(\mathbf{X}) \sim F_{h'}$ and $h'(\mathbf{Y}) \sim G_{h'}$. Now, if $F_{h'}(t) = F_h(t) \;\; \forall t$, from the properties of the most powerful test, we have $G_{h'}(t) \leq G_h(t) \;\; \forall t$. So, the power of the resulting test gets maximized when the likelihood ratio is used as the transformation function (follows from arguments similar to that used in the proof of Proposition 2.1).

If $F_{h'}(t) \neq F_h(t)$ for at least one t, one can find a monotone transformation $\psi(.)$, such that $\psi \circ h'(\mathbf{X}) \sim F_h$. Since the power of the rank test remains invariant under monotone transformation, $h'(\cdot)$ and $\psi \circ h'(\cdot)$ lead to the same power. So, we have $F_{\psi \circ h'}(t) = F_h(t) \;\; \forall t$. Now, using the same argument as above, we can claim that the transformation $h(\cdot)$ leads to more power than the transformation $h'(\cdot)$ or $\psi \circ h'(\cdot)$. $\square$

COMMENTS: In the case of matched pair data, we have $m_1 = m_2$. Though $\boldsymbol{\xi}_i$ and $\boldsymbol{\eta}_j$ are independent for $i \neq j$, we have dependency between $\boldsymbol{\xi}_i$ and $\boldsymbol{\eta}_i$. However, note that the convergence of $\widehat{\boldsymbol{\beta}}_m^F$ does not require independence of observations on $\mathbf{X}$ and $\mathbf{Y}$. So, similar result holds for the matched pair data. In the proof of Lemma 2.1 and Lemma 2.2, we did not use independence between observations on $\mathbf{X}$ and $\mathbf{Y}$. Therefore, analogous convergence results can be proved for $\widehat{\boldsymbol{\beta}}_m^M$ and $\widehat{\boldsymbol{\beta}}_m^D$ in the case of matched pair data as well. The convergence result similar to that of $|S_m(\beta_0, \boldsymbol{\beta}) - S(\beta_0, \boldsymbol{\beta})|$ for the matched pair data can be proved by writing $S_m(\beta_0, \boldsymbol{\beta})$ as a sum of the functions of the $\mathbf{x}_i$'s and that of $\mathbf{y}_i$'s (like the alterative expression for $D_m(\beta_0, \boldsymbol{\beta})$ used in Lemma 2.2) and then repeating the arguments used in the proof of Theorem 1 of Koo et al. (2008). So, analogous convergence result for $\widehat{\boldsymbol{\beta}}_m^S$ can also be proved.

Chapter 3

Tests based on shortest path algorithms

In Chapter 2, we proposed a general method based on linear projection of multivariate observations for distribution-free multivariate generalizations of univariate rank based two-sample tests. We have seen that the resulting WMW and KS tests based on SVM and DWD classifiers work well if F and G have reasonable linear separation between them, particularly when they differ in their locations. However, the implementation of that generalization method requires splitting of the whole sample into two subsamples, and the performance of the resulting tests depend on those subsample sizes. In this chapter, we propose another method for distribution-free multivariate generalizations of univariate rank based two-sample tests, which does not require any splitting of the whole sample and works well even when the two distributions differ only in their scatters or shapes. Here we find the shortest path that passes through all sample observations from F and G, and the tests are constructed by ranking the observations along that path. If we consider these sample observations as the vertices of a complete graph, and the distance between each pair of observations is considered as the cost of the edge connecting them, this shortest path connecting all observations is called the shortest Hamiltonian path (SHP). Note that SHP is a spanning tree, but not necessarily the minimal spanning tree (MST). The MST of a complete graph may or may not be a path. If it is a path (i.e., no vertex has degree bigger than 2), it is the SHP. But in practice, it often contains some vertices with degrees larger than 2, and in such cases, it differs from the SHP. Detailed descriptions of our SHP based two-sample tests are given in the following sections.

3.1 SHP and multivariate two-sample tests

Consider a graph $\mathcal{G}$ on n vertices. A Hamiltonian path $\mathcal{H}$ in $\mathcal{G}$ is defined as a connected, acyclic sub-graph of $\mathcal{G}$ with $n-1$ edges, where no vertex has degree bigger than two. In other words, $\mathcal{H}$ is a path in $\mathcal{G}$ that visits each vertex of $\mathcal{G}$ exactly once. For any given $\mathcal{G}$, a Hamiltonian path may or may not exist, but if $\mathcal{G}$ is a complete graph on n vertices, there are $n!$ Hamiltonian paths. However, for every path, there is another path in the reverse order. So, if we consider them as the same path, there are $n!/2$ distinct Hamiltonian paths. Now, consider $\mathcal{G}$ to be a complete graph on n vertices, where each of the $n(n-1)/2$ edges has a cost (e.g., the distance between the two vertices of the edge) associated with it. For each $\mathcal{H}$ in $\mathcal{G}$, one can compute the sum of the costs corresponding to its $n-1$ edges, which is defined to be the cost of $\mathcal{H}$. The Hamiltonian path having the minimum cost is defined as the shortest Hamiltonian path (SHP) $\mathcal{H}^*$. For a graph $\mathcal{G}$, $\mathcal{H}^*$ may not be unique, but if the costs corresponding to different edges come from continuous distributions, it becomes unique with probability one. Figure 3.1 shows a complete graph on four vertices $\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4\}$ along with the costs corresponding to different edges. There are 12 distinct Hamiltonian paths in this graph, where the path $\mathbf{z}_2 \rightarrow \mathbf{z}_1 \rightarrow \mathbf{z}_3 \rightarrow \mathbf{z}_4$ (or $\mathbf{z}_4 \rightarrow \mathbf{z}_3 \rightarrow \mathbf{z}_1 \rightarrow \mathbf{z}_2$) is the shortest Hamiltonian path.

In a two-sample problem, where we have n_1 independent observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}$ from F and n_2 independent observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$ from G, define $\mathbf{z}_i = \mathbf{x}_i$ for $i = 1, \ldots, n_1$ and $\mathbf{z}_{n_1+i} = \mathbf{y}_i$ for $i = 1, \ldots, n_2$. Now, consider a complete graph on $n = n_1 + n_2$ vertices $\mathbf{z}_1, \ldots, \mathbf{z}_n$, where the edge between $\mathbf{z}_i$ and $\mathbf{z}_j$ $(1 \leq i < j \leq n)$ has the cost $\|\mathbf{z}_i - \mathbf{z}_j\|$, the Euclidean distance between $\mathbf{z}_i$ and $\mathbf{z}_j$. We find the SHP $\mathcal{H}^*$ in this graph and rank the observations along $\mathcal{H}^*$. For instance, if we consider $\mathbf{z}_2 \rightarrow \mathbf{z}_1 \rightarrow \mathbf{z}_3 \rightarrow \mathbf{z}_4$ as the SHP, ranks of $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$ and $\mathbf{z}_4$ are taken as $2, 1, 3$ and 4, respectively. Univariate rank based tests (e.g., the WMW test or the KS test) can be constructed using these ranks. In Fig. 3.1, if we assume that $\mathbf{z}_1, \mathbf{z}_2$ are from F and $\mathbf{z}_3, \mathbf{z}_4$

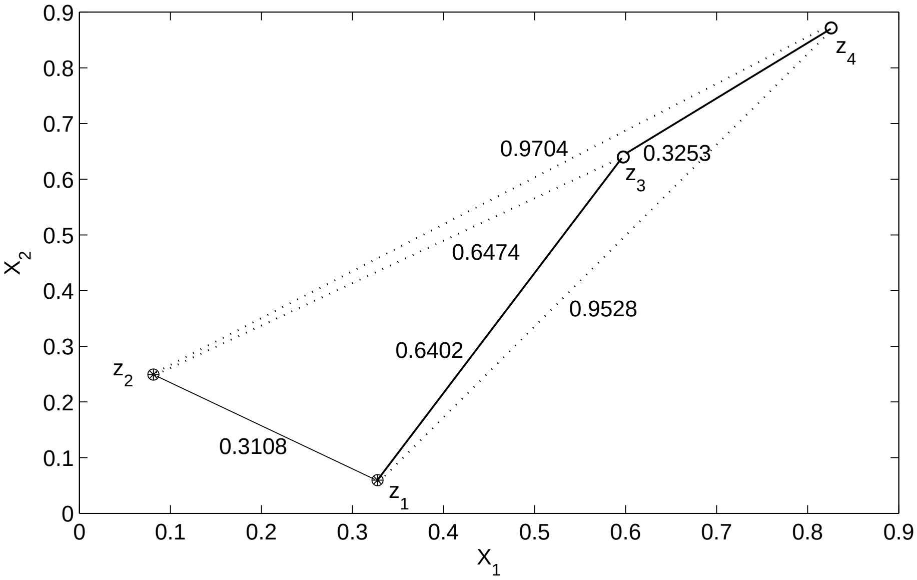


Figure 3.1: Shortest Hamiltonian path in a complete graph on four vertices.

are from G, the two-sided WMW statistic and the two-sided KS statistic both take the value 1. One should note that if the path is traversed in the reverse order (i.e., if we consider $\mathbf{z}_4 \rightarrow \mathbf{z}_3 \rightarrow \mathbf{z}_1 \rightarrow \mathbf{z}_2$ as the SHP), though the ranks of the observations change, the values of the WMW and the KS statistics remain unchanged, and hence the resulting tests lead to the same decisions as before. Similarly, one can construct a test based on the number of runs along $\mathcal{H}^*$. In Fig. 3.1, the number of runs turns out to be 2 ($\mathbf{z}_2 \rightarrow \mathbf{z}_1$ and $\mathbf{z}_3 \rightarrow \mathbf{z}_4$ or $\mathbf{z}_4 \rightarrow \mathbf{z}_3$ and $\mathbf{z}_1 \rightarrow \mathbf{z}_2$). Note that all these tests are based on pairwise distances between the sample observations. So, when the Euclidean distance is used, the corresponding test statistics become invariant under location change, rotation and homogeneous scale transformation of the data though they may not have the maximal invariance property for any of these transformations. These tests are fairly simple, and they can be conveniently used for HDLSS data or even for functional data taking values in a Banach space. Clearly, this proposed method based on SHP can be used for the multivariate generalization of any univariate rank based two-sample test, but here we concentrate only on multivariate generalization of the Wald and Wolfowitz (1940) run test. This is chosen because of its better empirical performance.

3.2 Multivariate run test based on SHP

As we have mentioned above, here we find $\mathcal{H}^*$ in the complete graph $\mathcal{G}$ consisting of $n_1 + n_2$ vertices each representing a sample observation and count the number of runs $T^{SHP}_{n_1,n_2}$ along $\mathcal{H}^*$. Note that $T^{SHP}_{n_1,n_2}$ can be expressed as $T^{SHP}_{n_1,n_2} = 1+\sum_{i=1}^{n-1} \Lambda_i^{\mathcal{H}^*}$, where $\Lambda_i^{\mathcal{H}^*}$ is an indicator variable that takes the value 1 if and only if the i-th edge of $\mathcal{H}^*$ connects two observations from two different distributions. If F and G are widely separated, one would expect $T^{SHP}_{n_1,n_2}$ to be small, while under $H_0 : F = G$, it is expected to be large. So, we reject H_0 for small values of $T^{SHP}_{n_1,n_2}$. Friedman and Rafsky (1979) constructed a similar multivariate run test based on graphs (referred as the MST run test in Chapter 2). They found the MST ($\mathcal{M}$) in $\mathcal{G}$ and used the test statistic $T^{MST}_{n_1,n_2} = 1 + \sum_{i=1}^{n-1} \Lambda_i^{\mathcal{M}}$, where $\Lambda_i^{\mathcal{M}}$ denotes the indicator variable that takes the value 1 if and only if the i-th edge of $\mathcal{M}$ connects two observations from two different distributions. Naturally, H_0 is rejected for small values of $T^{MST}_{n_1,n_2}$. Note that if F and G are one-dimensional distributions, both the SHP and the MST are obtained by joining the observations $z_1, \ldots, z_n$ either in increasing or in decreasing order, and in that case, $T^{SHP}_{n_1,n_2}$ and $T^{MST}_{n_1,n_2}$ match with the univariate run statistic. Therefore, the MST run test and our proposed run test, both can be viewed as multivariate generalizations of the univariate run test.

From the above discussion, it is quite transparent that the MST run test and the proposed run test both have the distribution-free property in one dimension. But, the MST run test fails to retain this distribution-free property in higher dimension. Unlike what happens in the case of SHP, the distribution of degrees of the vertices in the MST is not fixed, and for a given data set, the conditional distribution of the MST run statistic under H_0 depends on the distribution of these degrees and the configuration of the MST (see Friedman and Rafsky (1979) for details). However, our proposed run test successfully retains this distribution-free property in higher dimensions. Note that $T^{SHP}_{n_1,n_2}$ and the univariate run statistic are the same function of ranks of the observations from the two distributions; while $T^{SHP}_{n_1,n_2}$ uses the ranks computed along $\mathcal{H}^*$. Now, under H_0, because of the exchangeability of $\mathbf{z}_1, \ldots, \mathbf{z}_n$, irrespective of the underlying distribution and data dimension, this rank vector has the same distribution as in the univariate case. Therefore, $T^{SHP}_{n_1,n_2}$ has the distribution-free property, and its null distribution exactly matches with that of the univariate run statistic, which is given by

$$\begin{array}{ll} P_{H_0}(T^{SHP}_{n_1,n_2} = 2r) & = 2\binom{n_1-1}{r-1}\binom{n_2-1}{r-1}/\binom{n}{n_1} \text{ and} \\ P_{H_0}(T^{SHP}_{n_1,n_2} = 2r+1) & = \left[\binom{n_1-1}{r}\binom{n_2-1}{r-1} + \binom{n_1-1}{r-1}\binom{n_2-1}{r}\right]/\binom{n}{n_1} \end{array}$$

for $r = 1, 2, \dots, \min\{n_1, n_2\}$, where $\binom{a}{b} = 0$ if $a < b$ (see e.g., Wald and Wolfowitz (1940); Gibbons and Chakraborti (2003)). In that sense, $T^{SHP}_{n_1,n_2}$ can be viewed as the most natural multivariate generalization of the univariate run statistic. However, unlike the univariate case, this multivariate rank-based test may not have the semi-parametric optimality discussed in Hallin and Werker (2003). If n_1 and n_2 are small, in order to carry out our test, we can use the statistical table available for the univariate run test. However, because of the discrete nature of $T^{SHP}_{n_1,n_2}$, one may need to use randomization at the cut-off point to match the size of the test with the level of significance. Note that the multisample extension of the proposed test is quite straight forward, and the null distribution of the test statistic can be found in Mood (1940).

If n_1 and n_2 are large, one can also use the test based on the asymptotic null distribution of $T^{SHP}_{n_1,n_2}$. Under H_0, the expectation and the variance of $T^{SHP}_{n_1,n_2}$ are $E_{H_0}(T^{SHP}_{n_1,n_2}) = 2n_1n_2/n + 1$ and $Var_{H_0}(T^{SHP}_{n_1,n_2}) = 2n_1n_2(2n_1n_2 - n)/n^2(n-1)$, respectively. Let us assume that as $n \to \infty$, $n_1/n \to \lambda$ for some $\lambda \in (0,1)$. Under this condition, $E_{H_0}(T^{SHP}_{n_1,n_2}/n) \to 2\lambda(1-\lambda)$ and $Var_{H_0}(T^{SHP}_{n_1,n_2}/\sqrt{n}) \to 4\lambda^2(1-\lambda)^2$ as $n \to \infty$. In this case, one can show that (see e.g., Wald and Wolfowitz (1940)), under H_0, $T^{SHP*}_{n_1,n_2} = \sqrt{n}\left[T^{SHP}_{n_1,n_2}/n - 2\lambda(1-\lambda)\right] \xrightarrow{d} N(0, 4\lambda^2(1-\lambda)^2)$.

However, unless n_1 and n_2 are very small, finding $\mathcal{H}^*$ in a complete graph $\mathcal{G}$ is a computationally hard problem. While one can easily find $\mathcal{M}$ in $\mathcal{G}$ in polynomial time, finding $\mathcal{H}^*$ is equivalent to the well-known travelling salesman's problem, which is NP-complete (see e.g., Garey and Johnson (1979)). However, there are some good heuristic search algorithms available in literature (see e.g., Lawler et al. (1985)). In this article, we have adopted a popular method based on Kruskal's algorithm (see e.g., Kruskal (1956)). First, it sorts the edges of $\mathcal{G}$ in increasing order of their costs. Next, it starts from the edge with the minimum cost and selects the edges one by one according to their costs. However, if an edge along with the previously chosen edges makes a cycle or if it makes the degree of a vertex more than two, we do not select that edge. The algorithm terminates when $n-1$ edges are chosen. The Hamiltonian path formed by these $n-1$ edges is considered as the shortest Hamiltonian path. This algorithm worked well for our test, and the reasons for its success are discussed in detail in Section 3.4.

3.3 An illustrative example with high dimensional data

We have already mentioned that our proposed run test and the MST run test both can be used even when the dimension of the data exceeds the sample size. Now, we consider two simple examples to investigate how these two tests perform in HDLSS situations. Let us assume that the observations in F and G are distributed as $N_d((\mu,\ldots,\mu)^T,\sigma^2\mathbf{I}_d)$ and $N_d((0,\ldots,0)^T,\mathbf{I}_d)$, respectively. Here, N_d stands for a d-variate normal distribution, and $\mathbf{I}_d$ denotes the $d\times d$ identity matrix. We consider two choices of μ and σ^2, namely, ($\mu = 0.3$, $\sigma^2 = 1$) and ($\mu = 0$, $\sigma^2 = 1.3$), which lead to a location problem and a scale problem, respectively. In each case, we generated 20 observations from each distribution to test $H_0 : F = G$ against $H_A : F \neq G$. Each experiment was repeated 500 times, and the proportion of times a test rejected H_0 was considered as an estimate of its power. In the case of MST run test, which is not distribution-free, we used the conditional test based on 500 permutations. We used different values of d ranging from 3 to 3000, and the results are presented in Fig. 3.2. Like our proposed run test, the Adjacency test of Rosenbaum (2005) is also distribution-free. So, we have also used it for comparison. To make it applicable to HDLSS data, we used the Euclidean metric for distance computation as before. For this test, we used both, the distances between the observations and the distances between the coordinatewise rank vectors to perform the test. Since the former one yielded better results, in Fig. 3.2 we have reported the estimated powers for that test.

Both in location and scale problems, as d increases, the separability between F and G also increases. So, one should expect the powers of these tests to tend to unity as the dimension increases. We observed that in the location problem (see Fig. 3.2(a)), but not in the scale problem (see Fig. 3.2(b)). In the location problem, all three tests had comparable performance, though our proposed test had an edge. But, the result was

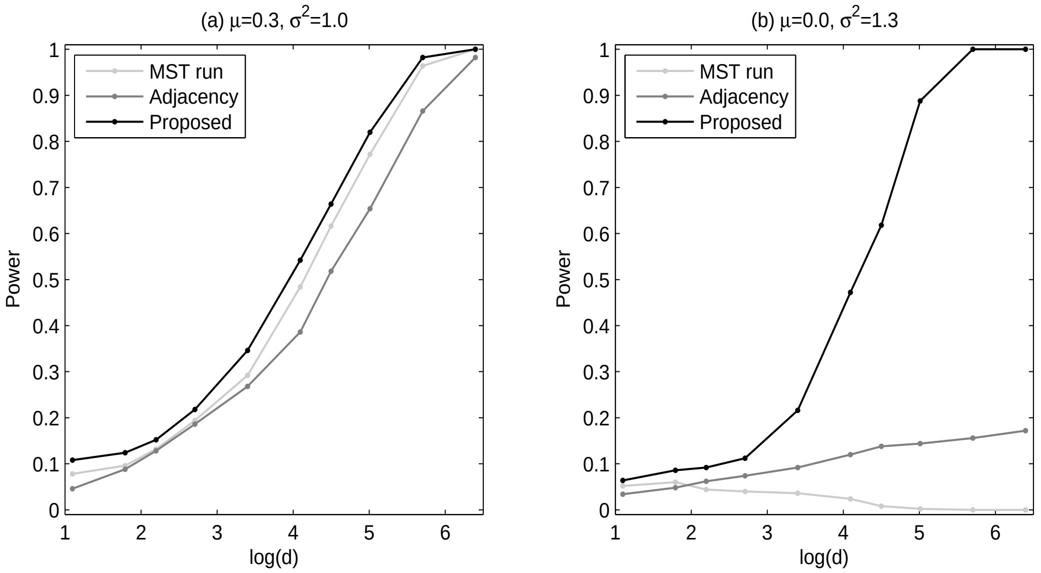


Figure 3.2: Powers of two run tests and the Adjacency test for varying choices of d.

more interesting in the case of scale problem. In this case, the powers of the proposed test and the Adjacency test increased with d, but latter increased at a very, very slow rate. While the power of the proposed run test rapidly increased to unity, that of the MST run test surprisingly dropped down to zero as the dimension increased. In the next section, we investigate the reasons behind such diametrically opposite behavior of these two multivariate run tests for high dimensional data.

3.4 Behavior of multivariate run tests in high dimensions

To carry out a theoretical investigation on the behavior of our proposed test and the MST run test for high dimensional data, here also, we assume to have n_1 independent observations on $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})^T$ from F and n_2 independent observations on $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(d)})^T$ from G, while the observations on $\mathbf{X}$ and $\mathbf{Y}$ are also considered to be independent. We consider the same set of assumptions (A1)-(A3) as in Chapter 2 and study the limiting behavior of the power functions of these two run tests when n_1 and n_2 are fixed as d diverges to infinity.

We have seen that, under (A1)-(A3), the pairwise distance between any two observations, when divided by $d^{1/2}$, converges in probability to a positive constant. If both of them are from F (respectively, G), it converges to $\sigma_1\sqrt{2}$ (respectively, $\sigma_2\sqrt{2}$). If one of them is from F and the other one is from G, it converges to $(\sigma_1^2+\sigma_2^2+\nu^2)^{1/2}$. Here σ_1^2, σ_2^2 and ν^2 have the same meaning as in Chapter 2. However, if the components of $\mathbf{X}$ and $\mathbf{Y}$ vectors are independent and identically distributed, as they were in the examples involving normal distributions in Section 3.3, we only need the existence of second order moments of the component variables for these above convergence results. Under (A1)-(A3), if $\nu^2>0$ or $\sigma_1^2\neq\sigma_2^2$, the power of the proposed test converges to unity as the dimension increases.

THEOREM 3.1: *Assume that F and G both satisfy the assumptions (A1)-(A3). Also assume that n_1 and n_2 are such that $n_1!\ n_2!/(n_1+n_2-1)!\leq\alpha$. If $\nu^2>0$ or $\sigma_1^2\neq\sigma_2^2$, the power of the proposed run test of level α converges to 1 as d tends to infinity.*

The proof of the theorem is given in Section 3.8. Note that $n_1!\ n_2!/(n_1+n_2-1)!<0.05$ for all $n_1,n_2\geq 5$. So, for the large dimensional consistency of the proposed test with 5% nominal level, it is enough to have 5 observations from each distribution. Box plots in Fig. 3.3(b) show the distributions of $T^{SHP}_{n_1,n_2}$ for different choices of d in the location problem discussed in Section 3.3. This figure clearly shows that $T^{SHP}_{n_1,n_2}$ converged to 2 as d increased. This happens when ν^2 exceeds $|\sigma_1^2-\sigma_2^2|$. But, if we have $\nu^2<|\sigma_1^2-\sigma_2^2|$, $T^{SHP}_{n_1,n_2}$ converges (in probability) to 3 as d tends to infinity. We observed it in the scale problem (see Fig. 3.3(d)). If $\sigma_1^2>\sigma_2^2$ (respectively, $\sigma_1^2<\sigma_2^2$), $\mathcal{H}^*$ starts and ends with observations from F (respectively, G) with all observations from G (respectively, F) in the middle. One can appreciate this by looking at Figure 3.4, which shows the MST and the SHP for a two-class location and scale problems in dimension 3000, when we have five observations from each distribution. This whole phenomenon can be mathematically explained in the proof of Theorem 3.1.

One should also note that Theorem 3.1 holds even for the implemented version of the test, where $T^{SHP}_{n_1,n_2}$ is computed along the path obtained by Kruskal's algorithm. If $\nu^2>|\sigma_1^2-\sigma_2^2|$, this algorithm first selects (n_1-1) '**XX**'-type edges and (n_2-1) '**YY**'-type edges to form two disjoint paths before joining them by an '**XY**'-type edge. As a result, we have $T^{SHP}_{n_1,n_2}=2$. In the case of $\nu^2\leq|\sigma_1^2-\sigma_2^2|$, under the conditions of Theorem 3.1, we have $|\sigma_1^2-\sigma_2^2|>0$. Without loss of generality, let us assume $\sigma_1^2<\sigma_2^2$, which implies $2\sigma_1^2 < \sigma_1^2 + \sigma_2^2 + \nu^2 \leq 2\sigma_2^2$. So, Kruskal's algorithm first selects $(n_1 - 1)$ '**XX**'-type edges to form a path on n_1 nodes corresponding to n_1 observations from F. Only two of these n_1 nodes will have degree 1 and the rest will have degree 2. Since all nodes in $\mathcal{H}$ have degrees less than or equal to 2, it cannot have more than two '**XY**'-type edges, and hence T_{n_1,n_2}^{SHP} cannot exceed 3.

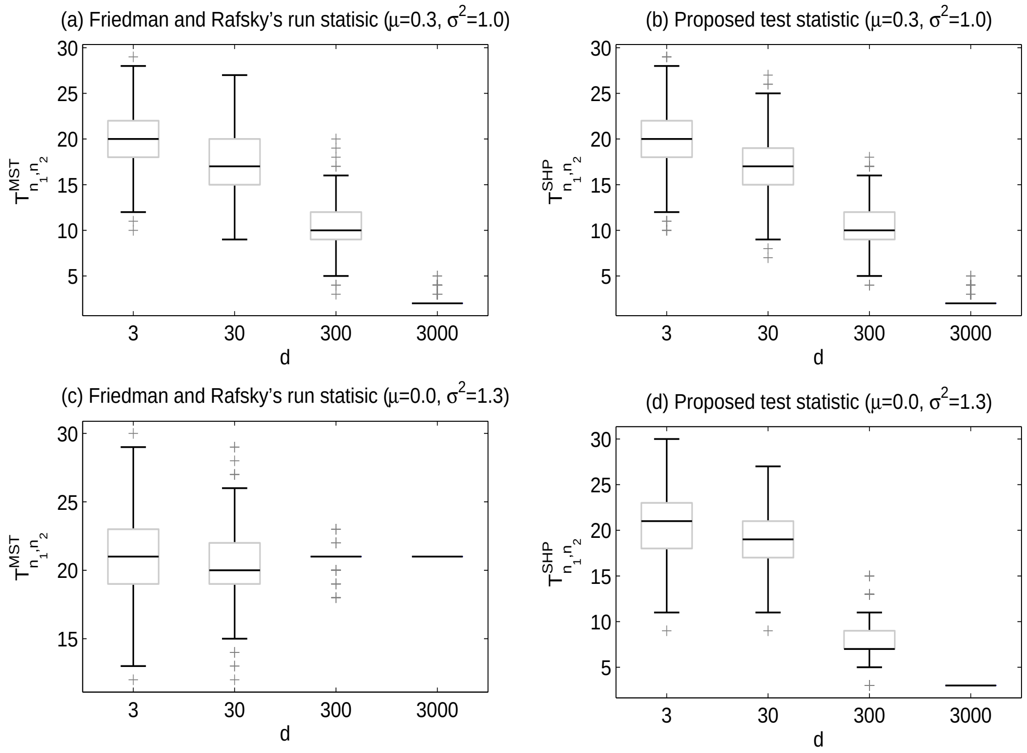
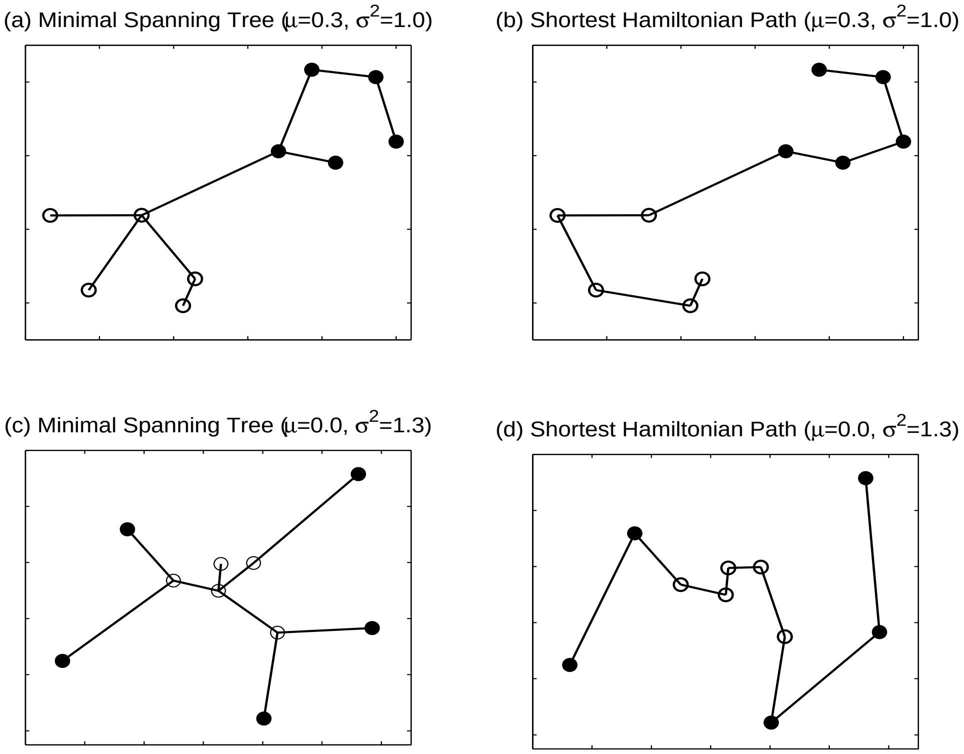


Figure 3.3: Distributions of T_{n_1,n_2}^{MST} and T_{n_1,n_2}^{SHP} for varying choices of d.

However, instead of leading to the actual $\mathcal{H}^*$, Kruskal's algorithm sometimes yields a sub-optimal path in terms of its cost. But, the test does not get affected if the number of runs along that path remains the same. In order to study the behavior of Kruskal's algorithm, we carried out an experiment with the location problem considered in Section 3.3. We chose $n_1 = n_2 = 5$ so that we could compute the actual $\mathcal{H}^*$ by complete enumeration. In the case of $d = 3000$, most of the times, we had two runs along the actual $\mathcal{H}^*$, where all observations from one distribution (F, say) were followed by all observations from the other distribution (G, say). Clearly, any re-arrangement among the observations from F (or from G) can change the cost of the path, but not the value of the run statistic. In many cases, Kruskal's algorithm led to such a re-arrangement.



(a) Minimal Spanning Tree (μ=0.3, σ^2=1.0)

(b) Shortest Hamiltonian Path (μ=0.3, σ^2=1.0)

(c) Minimal Spanning Tree (μ=0.0, σ^2=1.3)

(d) Shortest Hamiltonian Path (μ=0.0, σ^2=1.3)

Figure 3.4: Minimal spanning trees and shortest Hamiltonian paths for $d = 3000$.

Fig. 3.5(a) shows the box plots for efficiency scores of Kruskal's algorithm computed as the ratio of the cost of the actual $\mathcal{H}^*$ and that of the Kruskal path (path obtained by Kruskal's algorithm) for different dimensions, and Fig. 3.5(b) shows the distribution of the difference between the test statistics computed along these two paths. These figures clearly show that Kruskal's algorithm worked well, and its performance improved as the dimension increased. For $d = 3000$, the test statistics computed along the two paths were the same in more than 95% cases. We observed similar phenomenon for the scale problem as well, and this can also be explained using a similar argument.

However, under (A1)-(A3), the performance of the MST run test depends on the ordering '**XX**'-type, '**XY**'-type and '**YY**'-type distances. If $\nu^2 > |\sigma_1^2 - \sigma_2^2|$ (i.e., $\sigma_1\sqrt{2}, \sigma_2\sqrt{2} < (\sigma_1^2 + \sigma_2^2 + \nu^2)^{1/2}$), for large d, all '**XY**'-type distances become larger than all '**XX**'-type and '**YY**'-type distances. In that case, each observation from F (respectively, G) tends to have its first $n_1 - 1$ (respectively, $n_2 - 1$) neighbors from F (respectively, G) itself. As a result, $T^{MST}_{n_1,n_2}$ attains its minimum value 2 with probability tending to one. We observed it in the location problem in Section 3.3 (see Fig. 3.3(a)), where we had $\sigma_1^2 = \sigma_2^2 = 1$ and $\nu^2 = 0.09$. So, in this case, unless n_1 and n_2 are very small, the power of the MST run test converges to 1 as d tends to infinity. However, the situation gets completely changed if $\nu^2 < |\sigma_1^2 - \sigma_2^2|$ (i.e., either $\sigma_1\sqrt{2}$ or $\sigma_2\sqrt{2}$ exceeds $(\sigma_1^2 + \sigma_2^2 + \nu^2)^{1/2}$). Without loss of generality, let us assume $\sigma_2^2 - \sigma_1^2 > \nu^2$ as it was the case in the scale problem in Section 3.3. In this case, each observation from F has its first $n_1 - 1$ neighbors from F as before, but each observation from G has all of its first n_1 neighbors from F as well. So, T_{n_1,n_2}^{MST} converges (in probability) to $n_2 + 1$ (see Fig. 3.3(c) and Fig. 3.4), which is equal to (even bigger than) its expected value under H_0 if $n_1 = n_2$ ($n_1 < n_2$), and much higher than the cut-off. This is the reason why this test yielded poor performance in the scale problem. In fact, in such cases, depending on n_1 and n_2, the power of this test may even tend to zero as d tends to infinity.

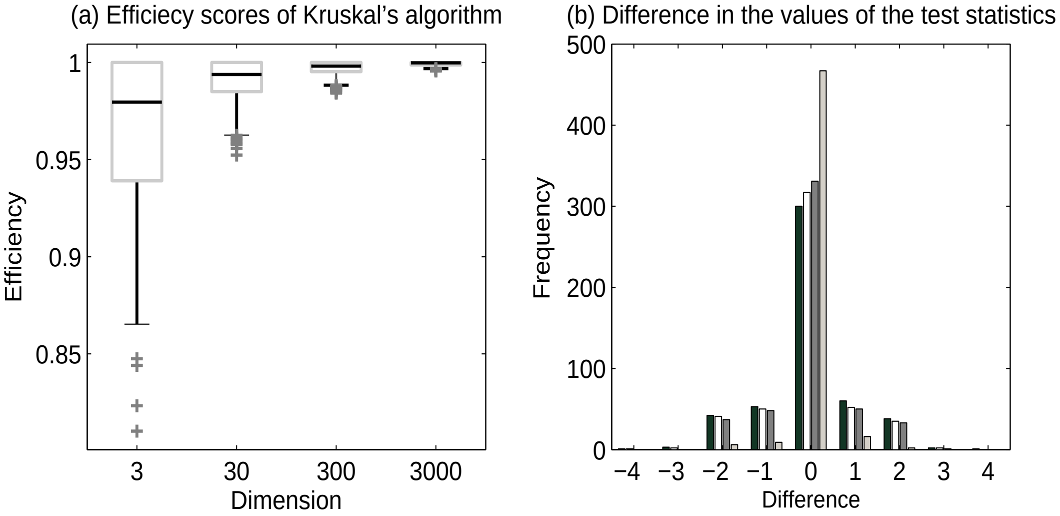


Figure 3.5: Performance of Kruskal's algorithm in different dimensions.
(d= 3 (black), d= 30 (white), d= 300 (dark grey) and d= 3000 (light grey))

THEOREM 3.2: *Suppose that F and G both satisfy the assumptions (A1)-(A3).*
(i) If $\nu^2 > |\sigma_1^2 - \sigma_2^2|$ and $\max\{\lfloor n/n_2 \rfloor, \lfloor n/n_1 \rfloor\}/\binom{n_1+n_2}{n_1} \le \alpha$, the power of the MST run test of level α converges to 1 as $d \to \infty$ (Here, $\lfloor r \rfloor$ denotes the highest integer $\le r$).
(ii) If $\nu^2 < \sigma_1^2 - \sigma_2^2$ and $n_1/n_2 > (1+\alpha)/(1-\alpha)$ (interchange σ_1^2 and σ_2^2, if required, and in that case, interchange n_1 and n_2, accordingly), the power of the MST run test of level α converges to 0 as $d \to \infty$.

The proof of the theorem is given in Section 3.8. Note that part (*ii*) of Theorem 3.2 gives only a sufficient condition when the MST run test fails. This test may fail in many other cases. For instance, in the scale problem in Section 3.3, we had $n_1 = n_2 = 20$ (i.e., $n_1/n_2 = 1$), but the power of this test dropped down to 0 as d increased.

3.5 Results from the analysis of simulated data sets

We analyzed some simulated data sets to compare the performance of the proposed test with some popular nonparametric two-sample tests available in the literature. Along with the MST run test and the Adjacency test, we also considered the NN test (see, e.g., Schilling (1986a); Henze (1988)) based on three neighbors and the Cramer test (see e.g., Baringhaus and Franz (2004)) for comparison. We carried out our analysis for $n_1 = n_2 = 20$ and $n_1 = n_2 = 50$. Unlike the proposed run test and the Adjacency test, the other three tests do not have the distribution-free property. For them, we used the conditional tests based on 500 permutations. Each experiment was repeated 500 times, and the estimated powers of different tests are reported in Table 3.1 for two choices of d (30 and 90).

As Example-1 and Example-2, we considered the location and the scale problems discussed in Section 3.3. In Example-1 (location problem), Cramer test had the best performance followed by the NN test. The proposed test had the third best performance in this example. But, in Example-2 (scale problem), this proposed test outperformed all of its competitors. In view of Theorems 3.1 and 3.2, good performance of the proposed test and poor performance of the MST run test were expected in this example. Like MST run test, the power of the NN test also dropped down to zero as the dimension increased. The reason behind its poor performance will be discussed in Chapter 4.

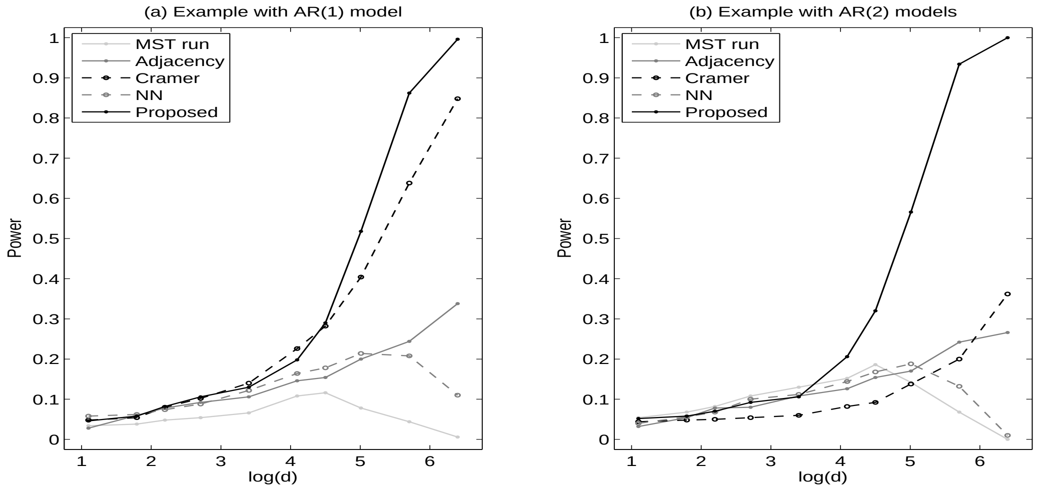
In the next four examples (Example-3 to Example-6), we had $\nu^2 = 0$ and $\sigma_1^2 = \sigma_2^2$, where ν^2, σ_1^2 and σ_2^2 have the same meanings as in (A3). We used these examples to investigate how the proposed test performs when the assumptions of Theorem 3.1 do not hold. Example-3 and Example-4 deal with two multivariate normal distributions, where F and G differ only in their correlation structures. In Example-3, F and G had the scatter matrices $\mathbf{\Sigma}_F = (((0.35)^{|i-j|}))_{d\times d}$ and $\mathbf{\Sigma}_G = (((-0.35)^{|i-j|}))_{d\times d}$, respectively.

In Example-4, while all off-diagonal elements of $\mathbf{\Sigma}_F$ were 0.1, those of $\mathbf{\Sigma}_G$ were 0.3. Note that (A1)-(A3) were valid in Example-3, but (A2) was violated in Example-4. In Example-3, the NN test had the best performance followed by the proposed test. In this example, the Cramer test failed to compete with other methods. In Example-4, the proposed test clearly outperformed all of its competitors. The Adjacency test had the next best performance, but even its power was not at all comparable to that of the proposed test. In Example-5, F (multivariate normal distribution $N_d((0,0,\ldots,0)^T, 3\mathbf{I}_d)$) and G (standard multivariate t-distribution with 3 degrees of freedom) had the same mean vector and the same dispersion matrix, but they differed in their shapes. The proposed test had excellent performance in this example as well. While the MST run test and the NN test both failed to reject H_0 even on a single occasion, the proposed test could reject it in almost all cases. In Example-6, F was an equal mixture of two normal distributions $N_d(0.3\ \mathbf{1}_d,\ \mathbf{I}_d)$ and $N_d(-0.3\ \mathbf{1}_d,\ 4\mathbf{I}_d)$, and G was also an equal mixture of two normal distributions $N_d(0.3\ (\mathbf{1}_{d/2}^T, -\mathbf{1}_{d/2}^T)^T,\ \mathbf{I}_d)$ and $N_d(0.3\ (-\mathbf{1}_{d/2}^T, \mathbf{1}_{d/2}^T)^T,\ 4\mathbf{I}_d)$. Here $\mathbf{1}_d = (1,\ldots,1)^T$ denotes the d-dimensional vector with all elements unity. Again, in this example, the proposed test outperformed its competitors.

Table 3.1: Observed powers (in %) of two-sample tests in simulated data sets

		Ex. 1		Ex. 2		Ex. 3		Ex. 4		Ex. 5		Ex. 6	
n	d	30	90	30	90	30	90	30	90	30	90	30	90
40	MST run	29	62	04	01	39	44	07	03	00	00	09	09
	Adjacency	27	52	09	14	32	39	12	15	31	43	14	39
	Cramer	83	100	10	15	08	05	06	07	49	67	04	02
	NN test	49	87	07	04	48	59	10	09	01	00	18	25
	Proposed	35	66	22	62	44	50	18	55	85	99	28	56
100	MST run	62	96	06	02	94	97	13	10	00	00	18	19
	Adjacency	50	87	11	16	87	93	18	27	71	85	42	88
	Cramer	99	100	14	42	15	13	09	09	92	99	06	06
	NN coin.	84	99	08	04	99	100	23	20	01	00	35	54
	Propsed	69	98	39	95	98	99	42	94	99	100	67	98

Finally, we considered two examples with auto-regressive processes of order 1 and order 2 (AR(1) and AR(2)). In one example, the observations $\mathbf{X} = (X^{(1)},\ldots,X^{(500)})$ in F were generated using the AR(1) model $X^{(t)} = 0.25 + 0.3X^{(t-1)} + U_t$ for $t = 1,\ldots,500$, and the observations $\mathbf{Y} = (Y^{(1)},\ldots,Y^{(500)})$ in G were generated using another AR(1) model $Y^{(t)} = 0.25 + 0.5Y^{(t-1)} + V_t$, where $X^{(0)}$,$Y^{(0)}$, $U_1 \ldots, U_{500}$,$V_1, \ldots, V_{500}$ are i.i.d. $N(0,1)$ variates. In the other example, the observations in F were generated using the AR(2) model $X^{(t)} = 0.3X^{(t-1)} + 0.2X^{(t-2)} + U_t$ for $t = 1, \ldots, 500$, and those in G were generated using the model $Y^{(t)} = 0.4Y^{(t-1)} + 0.3Y^{(t-2)} + V_t$ for $t = 1, \ldots, 500$, where $X^{(0)}, X^{(-1)}, Y^{(0)}, Y^{(-1)}, U_1, U_2, \ldots, U_{500}, V_1, V_2, \ldots, V_{500}$ are i.i.d. $N(0,1)$. In both cases, we repeated the experiment 500 times taking $n_1 = n_2 = 20$. We performed this experiment for various choices of d starting from 3 to 3000, and the results are presented in Fig. 3.6. The superiority of the proposed test in high dimension is quite transparent from this figure, especially in the second example.


(a) Example with AR(1) model

(b) Example with AR(2) models

Figure 3.6: Powers of different two-sample tests for varying choices of d.

3.6 Results from the analysis of benchmark data sets

We analyzed five benchmark data sets for further assessment of the proposed method. Three of these data sets, Sonar data, Colon data and Arcene data, were analyzed in Chapter 2. The Trace data set is obtained from the UCR time series classification/clustering page (http://www.cs.ucr.edu/∼eamonn/time_series_data/), and the Ionosphere data set is taken from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/). Detailed descriptions of these data sets are available at these sources. All these data sets have been extensively used in the literature of supervised classification, and in each of these cases, there is a reasonable separation between the competing classes. So, here also different tests can be compared on the basis of their power functions. However, as we have mentioned before, it is difficult to compare among different test procedures using a single experiment based on the whole data set. Therefore, in each of these cases, we repeated the experiment several times taking random subsets of the same size chosen from the whole data set. We will use the same strategy for the analysis of benchmark data sets in the subsequent chapters. In this section, we formed these subsets taking equal number of observations from the two classes, and each experiment was repeated 500 times to compute the powers of different tests. The results for different subset sizes (sample sizes) are shown in Fig. 3.7.

The Ionosphere data set contains 34-dimensional observations from two classes, which correspond to 'Good' and 'Bad' radar returns. Radar data were collected by a system in Goose Bay that consisted of a phased array of 16 high-frequency antennas. The targets were free electrons in the ionosphere. Radar returns showing evidence of some type of structure in the ionosphere are termed as 'Good', and the returns which do not show any evidence are termed as 'Bad'. There are 126 instances of 'Good' and 225 instances of 'Bad' radar returns (see also Sigillito et al. (1989) for details). In this data set, the proposed test and the Cramer test had better performance than their competitors (see Fig. 3.7(a)). For sample size less than 20, the latter had a slight edge, but the proposed test had higher power afterwards. These two tests had power 1 for samples of size 40 or higher. The performances of other three tests were also comparable. Among them, the NN test yielded better performance.

Description of the Sonar data set has been given in Chapter 2. In this data set, the NN test had the best overall performance closely followed by the proposed run test (see Fig. 3.7(b)). In all cases, the difference between their powers was less than 0.02. The MST run test also had comparable performance. The Cramer test had the highest power for sample size 10, but it was outperformed by the NN test and the proposed test for larger sample size. The Adjacency test did not have satisfactory performance in this data set. For instance, when all other tests reached the maximum power 1, it had power less than 0.3.

The Trace data set was designed to simulate instrumentation failures in a nuclear power plant. The original data set consists of 16 classes each containing 50 instances,

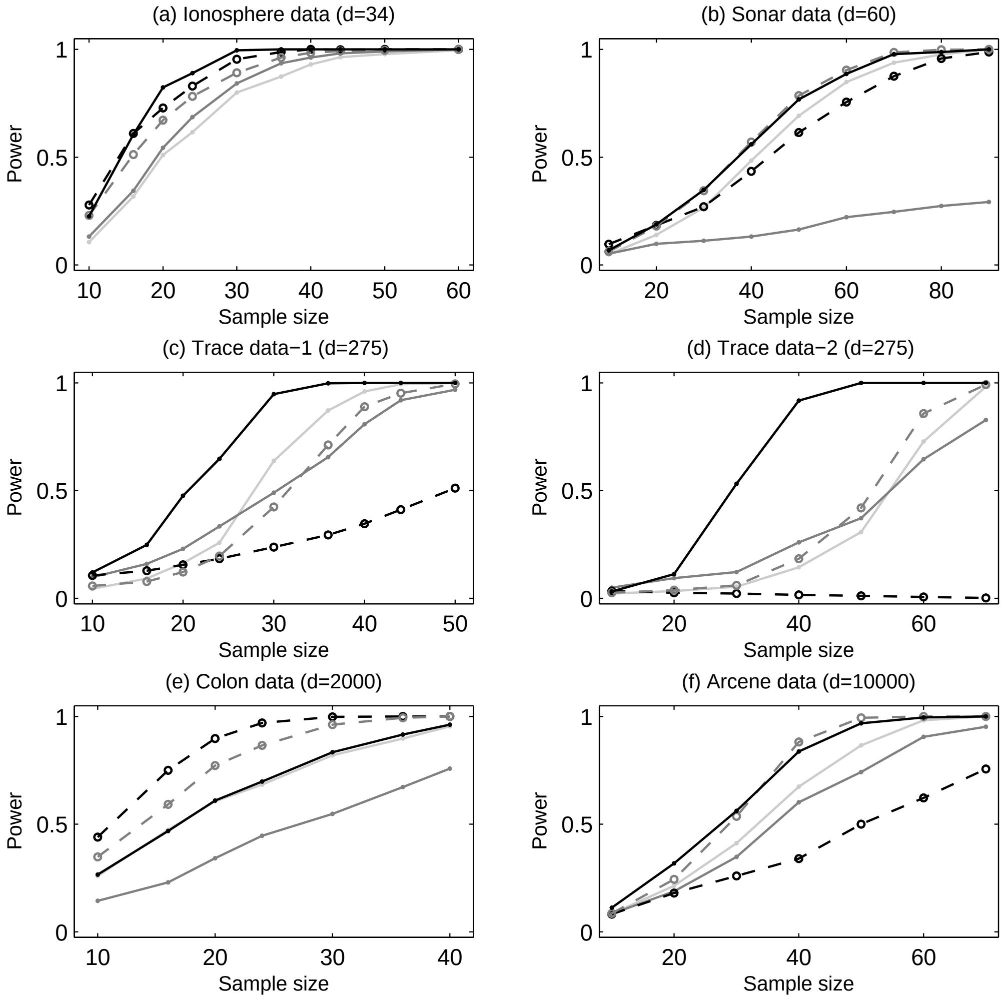


Figure 3.7: Powers of different two-sample tests in benchmark data sets. (MST run test (light grey), Adjacency test (dark grey), Cramer test (black dashed), NN test (dark grey dashed), proposed test (black) in benchmark data sets.)

where each instance has four features. For our analysis, we used a subset of this data set, which is available at the UCI machine learning repository. It contains the second feature of class 2 and 6, and the third feature of class 3 and 7, which are considered as the four new classes. There are 200 instances, 50 for each class, where all instances are linearly interpolated to have the same length of 275 data points. We considered all $\binom{4}{2}$ pairs of classes separately for testing, but in four out of these six cases, because of high separability between two classes, almost all tests attained power 1 even when very small samples were used. So, here we report the results only for two testing problems, one between the first and the second classes (referred to as Trace data-1), and the other between the third and the fourth classes (referred to as Trace data-2). In both of these cases, our proposed test had substantially higher power than all other tests considered here (see Fig. 3.7(c) and Fig. 3.7(d)). The Cramer test had very poor performance in these data sets, especially in Trace data-2.

Next, we analyzed Colon and Arcene data sets, where the data dimensions are larger than 1000. Descriptions of these data sets have been given in Chapter 2. In Colon data set, the Cramer test yielded the best performance, while the NN test had the second position (see Fig. 3.7(e)). The MST run test and the proposed run test had almost similar performance, and they performed better than the Adjacency test. In Arcene data set, the proposed test and the NN test outperformed all other tests considered here (see Fig. 3.7(f)). The proposed test had an edge over the NN test for small sample sizes, but for samples of size 40 and 50, the latter had the highest power. Both of them had power 1 for samples of size 60 or higher.

3.7 Tests for matched pair data

Like Section 2.6, here we deal with n paired observations $\binom{\mathbf{x}_1}{\mathbf{y}_1}, \binom{\mathbf{x}_2}{\mathbf{y}_2}, \ldots, \binom{\mathbf{x}_n}{\mathbf{y}_n}$ from a $2d$-variate distribution with d-dimensional marginals F and G for $\mathbf{X}$ and $\mathbf{Y}$, respectively. In such cases, it is common practice to consider $\{\boldsymbol{\xi}_i = \mathbf{x}_i - \mathbf{y}_i;\ i = 1, 2, \ldots, n\}$ as sample observations and perform a one-sample test. Here also, we consider the distribution of $\boldsymbol{\xi} = \mathbf{X} - \mathbf{Y}$ to be symmetric about some $\boldsymbol{\theta} \in \mathbb{R}^d$ and test the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ against the alternative $H_A : \boldsymbol{\theta} \neq 0$. First consider the case $d = 1$ and assume that $\xi_1, \ldots, \xi_n$ are i.i.d. univariate continuous random variables with a distribution symmetric about 0. In this case, if we define $S_i = sign(\xi_i)$ and R_i as the rank of $|\xi_i|$ in $\{|\xi_1|, \ldots, |\xi_n|\}$ for all $i = 1, 2, \ldots, n$, it is easy to check that.

(a) $P\{(S_1, \ldots, S_n) = (s_1, \ldots, s_n)\} = 2^{-n}$ for all $(s_1, \ldots, s_n) \in \{-1, 1\}^n$,

(b) $P\{(R_1, \ldots, R_n) = (r_1, \ldots, r_n)\} = 1/n!$ for all permutations $(r_1, \ldots, r_n)$ of $\{1, \ldots, n\}$,

(c) $(S_1, \ldots, S_n)$ and $(R_1, \ldots, R_n)$ are independent.

So, if the test statistic is a function of $(S_1, \ldots, S_n)$ and $(R_1, \ldots, R_n)$ (e.g., linear rank statistic), the resulting test becomes distribution-free in finite sample situations. To construct distribution-free tests for multivariate data, we extend the notions of signs $S_1, \ldots, S_n$ and ranks $R_1, \ldots, R_n$ in such a way that the results (a)-(c) hold under H_0.

We define $\boldsymbol{\eta}_i = -\boldsymbol{\xi}_i$ for $i = 1, \ldots, n$. Note that under H_0, $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ and $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ have the same distribution, while under H_A they differ in their locations. Now, consider a complete graph $\mathcal{G}_0$ on $2n$ vertices $\mathbf{z}_1, \ldots, \mathbf{z}_{2n}$, where $\mathbf{z}_i = \boldsymbol{\xi}_i$ and $\mathbf{z}_{n+i} = \boldsymbol{\eta}_i$ for $i = 1, \ldots, n$. Also, assume that each edge of $\mathcal{G}_0$ has a cost associated with it. For instance, the Euclidean distance between the two vertices of an edge can be considered as its cost. Now, consider a path $\mathcal{P}$ of length $n-1$ in $\mathcal{G}_0$ such that for every $i = 1, \ldots, n$, $\mathcal{P}$ covers either $\boldsymbol{\xi}_i$ or $\boldsymbol{\eta}_i$. Clearly, there are $2^n n!$ such paths in $\mathcal{G}_0$. However, for every path, there is another path in the reverse order. Again, for any path and its reverse path, two other equivalent paths can be obtained if we replace all $\mathbf{z}_i$'s by $\mathbf{z}_{n+i}$ (respectively, $\mathbf{z}_{i-n}$) if $i \leq n$ (respectively, $i > n$). For each of these four paths, the total cost of the $n-1$ edges remains the same. If we consider these four equivalent paths as the same path, the number of distinct covering paths (i.e., the paths that cover either $\boldsymbol{\xi}_i$ or $\boldsymbol{\eta}_i$ for all $i = 1, \ldots, n$) reduces to $2^{n-2}n!$. For each of these distinct covering paths, the sum of the costs corresponding to its $n-1$ edges is defined as its cost. Among these distinct paths, we choose the one having the minimum cost, and we call it the shortest covering path $\mathcal{P}_0$. This shortest covering path (SCP) may not be unique, but if the costs corresponding to different edges come from continuous distributions, just like SHP, it becomes unique with probability one.

Figure 3.8 shows a complete graph on $2n = 6$ vertices in two-dimension along with the costs corresponding to different edges. There are 12 distinct covering paths in this graph, where the path $\mathbf{z}_1 \rightarrow \mathbf{z}_3 \rightarrow \mathbf{z}_5$ (or $\mathbf{z}_5 \rightarrow \mathbf{z}_3 \rightarrow \mathbf{z}_1$, or equivalently, $\mathbf{z}_4 \rightarrow \mathbf{z}_6 \rightarrow \mathbf{z}_2$ or $\mathbf{z}_2 \rightarrow \mathbf{z}_6 \rightarrow \mathbf{z}_4$) is the SCP.

We define $S_1, \ldots, S_n$ and $R_1, \ldots, R_n$ along $\mathcal{P}_0$. For each $i = 1, \ldots, n$, S_i takes the value 1 (respectively, -1) if $\boldsymbol{\xi}_i$ (respectively, $\boldsymbol{\eta}_i$) appears on $\mathcal{P}_0$, and R_i is defined as the position of $\boldsymbol{\xi}_i$ (or $\boldsymbol{\eta}_i$) along $\mathcal{P}_0$. Between the two terminal nodes of $\mathcal{P}_0$, as a starting point, we choose the one which is closer to $\mathbf{0}$. Since $\boldsymbol{\xi}$ and $\boldsymbol{\eta} = -\boldsymbol{\xi}$ have the same distribution under H_0, and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ form an exchangeable collection, it is easy to check that $S_1, \ldots, S_n$ and $R_1, \ldots, R_n$ defined in this way satisfy properties (a)-(c) mentioned earlier. So, if we construct a test statistic, which is a function of $S_1, \ldots, S_n$

and $R_1, \ldots, R_n$, the resulting test becomes distribution-free. Like the univariate case, we can use the linear rank statistic of the form $T_0 = \sum_{i=1}^{n} I\{S_i = 1\}a(R_i)$, where $I\{\cdot\}$ is the indicator function, and $a : \{1, \ldots, n\} \to \mathbb{R}$ is a score function. Using $a(i) = 1$ and $a(i) = i$ for $i = 1, \ldots, n$, one obtains the sign statistic $\sum_{i=1}^{n} I\{S_i = 1\}$ and the signed-rank statistic $\sum_{i=1}^{n} R_i \; I\{S_i = 1\}$, respectively. Under H_0, since $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ and $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ have the same distribution, we expect almost equal numbers of $\boldsymbol{\xi}_i$'s and $\boldsymbol{\eta}_i$'s on $\mathcal{P}_0$. But, under H_A, one would expect a dominance of either the $\boldsymbol{\xi}_i$'s or the $\boldsymbol{\eta}_i$'s on $\mathcal{P}_0$. So, we should reject H_0 for very small or very large values of T_0, or in other words, H_0 is to be rejected for large values of $T_0^* = \max\{T_0, \sum_{i=1}^{n} a(i) - T_0\}$.

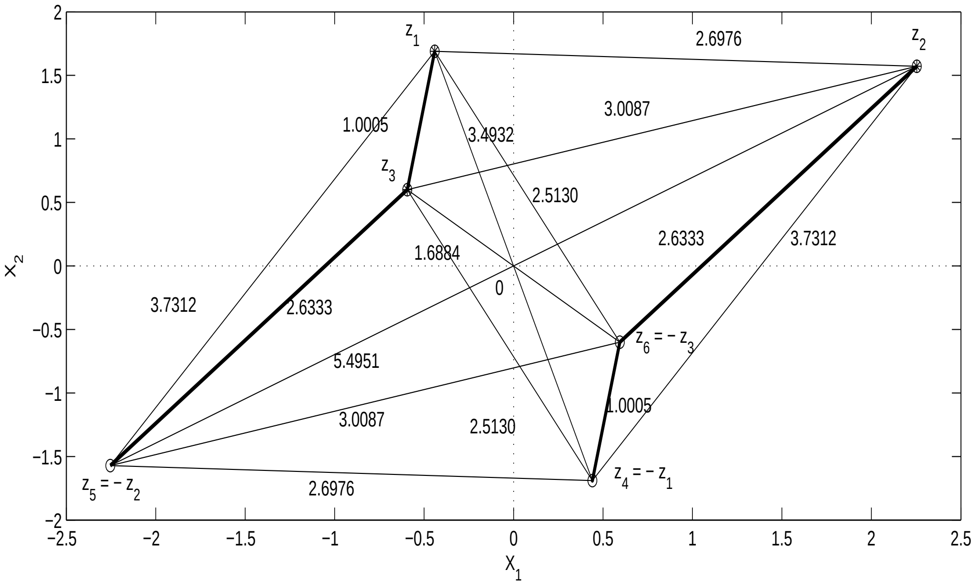


Figure 3.8: A complete graph on $2n = 6$ vertices and the shortest covering path.

One can also construct a test based on the number of runs or that based on the length of the longest run along $\mathcal{P}_0$. The number of runs can be expressed as $T_1 = 1 + \sum_{i=1}^{n-1} \Lambda_i$, where Λ_i is an indicator variable that takes the value 1 if and only if the i-th edge of $\mathcal{P}_0$ connects two observations with different S-values. The length of the longest run is given by $T_2 = \max_{0 \le i < j \le n} \; (j - i) \; I\{\Lambda_i = 1, \Lambda_{i+1} = \ldots = \Lambda_{j-1} = 0, \Lambda_j = 1\}$, where $\Lambda_0 = \Lambda_n = 1$ and for $i = 1, \ldots, n-1$, the Λ_is are defined as above. Under H_0, when two data clouds $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ and $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ are well mixed, T_1 is expected to be large, while T_2 is expected to be small. But under H_A, when they are well separated, we expect small values of T_1 and large values of T_2. So, here we use one-sided cut-offs.

In Fig. 3.8, along the path $\mathbf{z}_1 = \boldsymbol{\xi}_1 \to \mathbf{z}_3 = \boldsymbol{\xi}_3 \to \mathbf{z}_5 = \boldsymbol{\eta}_2$, both T_1 and T_2 take the value 2, while the values of the sign statistic and the signed rank statistic are 2 and 3, respectively. Barring the signed rank statistic, the values of the other three test statistics do not depend on the choice of the starting point of $\mathcal{P}_0$, and they remain the same if the path is traversed in the reverse order. The values of T_1 and T_2 also remain the same along an equivalent path, where all $\mathbf{z}_i$'s are replaced by $\mathbf{z}_{n+i}$ (respectively, $\mathbf{z}_{i-n}$) if $i \leq n$ (respectively, $i > n$). Along this path, though T_0 becomes $\sum_{i=1}^{n} a(i) - T_0$, the value of T_0^* remains the same. Therefore, all these tests lead to the same results as before if this equivalent path is chosen. Here also, in the univariate case, where $\mathcal{P}_0$ is obtained by joining the observations ξ_i (or η_i, if $\xi_i < 0$) in increasing order of magnitudes, T_0 coincides with the univariate linear rank statistic. Usually, we do not use run tests for the univariate one-sample location problem. But, alternative distribution-free tests for that problem can also be constructed using univariate analogs of T_1 and T_2.

Note that the test statistics constructed in this way are the same functions of $(S_1, \ldots, S_n)$ and $(R_1, \ldots, R_n)$ as their univariate analogs. So, irrespective of the underlying distribution and the data dimension, null distributions of these test statistics exactly match with those of their univariate counter parts, and the statistical tables available for the univariate tests can be used to determine the cut-offs in multivariate cases as well. Under H_0, T_0 is distributed as $\sum_{i=1}^{n} W_i$, where $P_{H_0}(W_i = 0) = P_{H_0}(W_i = a(i)) = 1/2$ for each $i = 1, \ldots, n$, and they are independent (see e.g., Gibbons and Chakraborti (2003)). One can check that under H_0, $T_1 - 1$ follows a binomial distribution with parameters $n-1$ and $1/2$. The null distribution of T_2 is given in Fu and Koutras (1994). For the construction of a linear rank test with the nominal level α $(0 < \alpha < 1)$, we consider a test function of the form $\phi_\alpha(t) = I\{t > t_\alpha\} + \gamma_\alpha I\{t = t_\alpha\}$, where t_α and γ_α $(0 \leq \gamma_\alpha < 1)$ are chosen in such a way that $E_{H_0}(\phi_\alpha(T_0^*)) = \alpha$. For run tests, we reject H_0 when T_1 is small or T_2 is large. Because of the discrete nature of T_1 and T_2, here also we need randomization at cut-off points so that the sizes of these tests match the level of significance α.

If the sample size is large, we can also use the tests based on the asymptotic null distributions of the test statistics. Asymptotic normality of T_1 under H_0 is obtained using normal approximation to the binomial distribution, and that of T_0 can be shown using a central limit theorem for independent random variables $W_1, W_2 \ldots, W_n$ (see e.g., Gibbons and Chakraborti (2003)). The large sample distribution of T_2 can be found in Gordon et al. (1986).

3.7.1 Computation of test statistics

Unless the sample size is very small, finding $\mathcal{P}_0$ is also computationally difficult, and it is an NP-complete problem (see e.g., Garey and Johnson (1979)). Here we use a heurustic method based on Prim's algorithm (see, Prim (1957)) for this purpose, where the distance between two observations is used as the cost of the edge connecting them. First we select the pair $\mathbf{z}_i$ and $\mathbf{z}_j$ ($|j-i| \neq n$) having the minimum distance between them and define a set $\Omega = \{i, j\}$. We join $\mathbf{z}_i$ and $\mathbf{z}_j$ by an edge to get a path of unit length with $\mathbf{z}_i$ and $\mathbf{z}_j$ as its two ends. From each of these two ends, we calculate the distance of $\mathbf{z}_l$, where $l \notin \Omega$ and $|l - l'| \neq n$ for any $l' \in \Omega$. If the minimum of these distances is observed between $\mathbf{z}_i$ and $\mathbf{z}_r$, we join $\mathbf{z}_i$ and $\mathbf{z}_r$ to get a path of length 2 ($\mathbf{z}_j \rightarrow \mathbf{z}_i \rightarrow \mathbf{z}_r$) with $\mathbf{z}_j$ and $\mathbf{z}_r$ as its two terminal nodes. We also update Ω by adding r to it. Next, we consider the distances of all $\mathbf{z}_l$ ($l \notin \Omega$ and $|l - l'| \neq n$ for any $l' \in \Omega$) from these two terminal nodes and choose a new edge in the same way to get a path of length 3. The set Ω is also updated by adding the index of the new selected node. We proceed in this way until a path of length $(n-1)$ is chosen. Clearly, this path contains either $\boldsymbol{\xi}_i$ or $\boldsymbol{\eta}_i$ for all $i = 1, \ldots, n$, and it is considered as the SCP. Test statistics are computed using the signs and the ranks (as defined before) of the observations along this path. Though this path finding algorithm sometimes leads to a sub-optimal solution in terms of cost, the test statistic computed along this path often remains the same as that computed along the actual $\mathcal{P}_0$, especially in high dimensions. As a consequence, the resulting tests generally perform well for HDLSS data. We will discuss this in detail in the next subsection to make it more transparent.

3.7.2 Power properties of constructed tests in HDLSS set up

Let $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ be n independent realizations of a d-dimensional random vector $\boldsymbol{\xi} = \big(\xi^{(1)}, \ldots, \xi^{(d)}\big)$ that follows a symmetric distribution with the location $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(d)})$

and the scatter matrix $\mathbf{\Sigma}$. Here we study the power properties of our tests where n is fixed and the d grows to infinity. For our theoretical investigation, we consider a more general cost function of the form $\rho_\psi^h(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = h\left(\sum_{q=1}^{d} \psi(|\xi_1^{(q)} - \xi_2^{(q)}|)\right)$, where $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ and $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ are continuous, monotonically increasing functions with $h(0) = \psi(0) = 0$ such that ρ_ψ^h is a distance in $\mathbb{R}^d$. Clearly, this class of distance functions include all l_p distances with $p \geq 1$. We modify the assumptions (B1)-(B3) (stated in Section 2.6) accordingly and assume the following.

(B1°) *For* $\mathbf{V} = \boldsymbol{\xi}_2$ *or* $-\boldsymbol{\xi}_2$*, second moments of* $\psi(|\xi_1^{(q)} - V^{(q)}|)$*'s are uniformly bounded.*

(B2°) *For* $\mathbf{V} = \boldsymbol{\xi}_2$ *or* $-\boldsymbol{\xi}_2$*,* $\sum_{q \neq q'} Corr\{\psi(\xi_1^{(q)}, V^{(q)}), \psi(\xi_1^{(q')}, V^{(q')})\}$ *is of order* $o(d^2)$*.*

Note that if ψ is bounded, (B1°) holds automatically. If ρ_ψ^h is the l_p distance, (B1°) holds when the $2p$-th moment of the $\xi^{(i)}$'s are uniformly bounded. Like (B2), the assumption (B2°) implies a form of weak dependence among the measurement variables. Now, define $\tau_d(\boldsymbol{\theta}) = d^{-1} \sum_{q=1}^{d} E\Big[\psi(|\xi_1^{(q)} + \xi_2^{(q)}|) - \psi(|\xi_1^{(q)} - \xi_2^{(q)}|)\Big]$ and $\tau = \liminf_{d \to \infty} \tau_d(\boldsymbol{\theta})$. In the case of Euclidean distance, (i.e., $\psi(t) = t^2$), one can show that $\tau_d(\boldsymbol{\theta}) = d^{-1} \sum_{q=1}^{d} (\theta^{(q)})^2 \geq 0$, where the equality holds if and only if $\theta^{(q)} = 0$ for $q = 1, 2, \ldots, d$. Also, for any ψ, where $\psi^{'}(t)/t$ is a non-constant monotone function in $(0, \infty)$, from Baringhaus and Franz (2010) it follows that $E\Big[\psi(|\xi_1^{(q)} + \xi_2^{(q)}|) - \psi(|\xi_1^{(q)} - \xi_2^{(q)}|)\Big] \geq 0$, where the equality holds if and only if $\theta^{(q)} = 0$ for $q = 1, 2, \ldots, d$. So, the result $\tau_d(\boldsymbol{\theta}) \geq 0$ also holds for such functions (e.g., $\psi(t) = t$ or $\psi(t) = t/(1+t)$), and there also $\tau_d(\boldsymbol{\theta}) = 0$ implies $\boldsymbol{\theta} = \mathbf{0}$. Therefore, under H_0, while we have $\tau = 0$, τ is expected to be positive under H_A. The following theorem shows that in such cases, the powers of our distribution-free tests based on T_0, T_1 and T_2 converge to unity as d increases.

THEOREM 3.3: *Assume that the distribution of* $\boldsymbol{\xi} = \mathbf{X} - \mathbf{Y}$ *satisfies (B1°) and (B2°). If* $\tau = \liminf_{d \to \infty} \tau_d(\boldsymbol{\theta}) > 0$ *and* 2^{n-1} *is larger than* $1/\alpha$*, the powers of the level* α *tests based on* T_0*,* T_1*, and* T_2 *converge to unity as d grows to infinity.*

The proof of the theorem is given in Section 3.8. This theorem shows that for a test of 5% level, it is enough to have six observations for its high-dimensional consistency.

Though our path finding method based on Prim's algorithm may fail to select the actual shortest covering path $\mathcal{P}_0$ in some of the cases, the above theorem holds even for the implemented versions of the tests based on that algorithm. From the arguments given in the proof of Theorem 3.3, one can check that under H_A, as $d \to \infty$, all $\rho_{\psi}^{h}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$ distances ($i \neq j$) become smaller than all $\rho_{\psi}^{h}(\boldsymbol{\xi}_i, \boldsymbol{\eta}_j)$ distances with probability tending to one. So, this algorithm first selects an edge connecting either two $\boldsymbol{\xi}_i$s or two $\boldsymbol{\eta}_i$s. Now, if it selects an edge connecting two $\boldsymbol{\xi}_i$s (respectively, $\boldsymbol{\eta}_i$s), because of this ordering of distances, only $\boldsymbol{\xi}_i$s (respectively, $\boldsymbol{\eta}_i$s) get selected in the subsequent stages. So, for large d, the covering path selected by the algorithm contains either all $\boldsymbol{\xi}_i$s or all $\boldsymbol{\eta}_i$s with probability tending to 1. Note that along this path, the arrangement of the $\boldsymbol{\xi}_i$s (or the $\boldsymbol{\eta}_i$s) can differ from that in actual $\mathcal{P}_0$, but that re-arrangement only changes the cost of the covering path, not the values of the resulting test statistics.

3.7.3 Analysis of simulated data sets

We analyzed some simulated data sets to compare the performance of our distribution-free tests with some existing one-sample tests. For this comparison, we considered the one-sample Hotelling's T^2 test, PS-sign test, PS-rank test (see, Puri and Sen (1971)), Sp-sign test and Sp-rank test (see e.g., Möttönen et al. (1997)). Codes for these tests are available in different R packages. However, these tests are not applicable when the dimension exceeds the sample size. So, in addition to them, we also considered the SR test (see Srivastava (2009)), the CQ test (see Chen and Qin (2010)) and the PA test (see Park and Ayyala (2013)), which can be used even in HDLSS situations. Note that all these tests were used in Section 2.6. For nonparametric sign and rank tests, we used both, the large sample test and the conditional test based on the permutation principle. In each case, the best one (which happened to be the permutation test in almost all cases) has been reported in Table 3.2.

For our proposed tests, we used three types of distance function: the Euclidean distance, the l_1 distance, and a bounded distance function with $\psi(t) = t/(1+t)$ and $h(t) = t$. Among them, the tests based on the Euclidean distance had the best overall performance. Also, the tests based on T_1 and T_2 performed better than the linear rank tests based on sign and signed rank statistics. Note that under H_A, '$\boldsymbol{\xi}\boldsymbol{\xi}$'-type and '$\boldsymbol{\eta}\boldsymbol{\eta}$'-type distances are expected to be smaller than '$\boldsymbol{\xi}\boldsymbol{\eta}$'-type distances. So, our path finding algorithm is supposed to start with either an '$\boldsymbol{\xi}\boldsymbol{\xi}$'-type edge or an '$\boldsymbol{\eta}\boldsymbol{\eta}$'-type edge. Also, if it starts with an '$\boldsymbol{\xi}\boldsymbol{\xi}$'-type edge, in the subsequent steps, it is supposed to choose

'$\boldsymbol{\xi\xi}$'-type edges with high probability. But, if an '$\boldsymbol{\xi\eta}$'-type edge is chosen in the middle, there is a high probability of choosing '$\boldsymbol{\eta\eta}$'-type edges in the subsequent steps. As a result, even under H_A, sometimes the values of T_0^* do not become large enough to reject H_0. We observed it in our experiments with sign and signed rank statistics. However, the tests based on T_1 and T_2 did not get much affected by this phenomenon. Therefore, here we report the results only for T_1 and T_2 based on the Euclidean distance.

First we considered some examples, where d is smaller than n. These examples involve multivariate normal, $t_{(2)}$ (t with 2 degrees of freedom) and Cauchy distributions. These distributions were chosen for varying degrees of heaviness of their tails. In each case, we generated 50 observations from a distribution with the location parameter $\Delta\mathbf{1}_d = (\Delta, \ldots, \Delta)^T$ and the scatter matrix $\mathbf{I}_d$ to test $H_0 : \Delta = 0$ against $H_A : \Delta \neq 0$. We considered two choices of d (30 and 40) and four choices of Δ (0, 0.1, 0.2 and 0.3) to study the level and the power properties of different tests. Each experiment was repeated 500 times, and the powers (sizes in the case of $\Delta = 0$) of different tests were estimated by the proportion of times they rejected H_0.

Table 3.2 shows that in the examples involving normal distributions, all tests had sizes close to 0.05, but in cases of $t_{(2)}$ and Cauchy distributions, the SR test had sizes much below the nominal level. The Hotelling's T^2 test and the PA test also had sizes below 0.05 in the case of Cauchy distribution. All other tests rejected the true $H_0 : \Delta = 0$ in nearly 5% of the cases.

In the examples involving normal distributions, CQ, PA, and SR tests had much higher powers than their competitors, though all other tests performed quite well. Among them, the test based on T_1 had the best performance for $d = 40$. In the examples involving 30-dimensional $t_{(2)}$ distributions, the Sp-rank test had the best performance closely followed by PS-rank, Hotelling's T^2, and Sp-sign tests. The PS-sign test and our proposed run tests also had competitive performance. However, in the case of $d = 40$, these run tests outperformed all other tests considered here. We observed similar results in the examples with Cauchy distributions as well. For $d = 30$, Hotelling's T^2, Sp-rank, PS-rank and these two run tests had comparable performance, but for $d = 40$, the run tests outperformed them. CQ, PA, and SR tests had poor performance in these examples.

Table 3.2: Observed levels and powers (in %) of one-sample tests.

	d	δ	Hot. T^2	Sp-sign	Sp-rank	PS-sign	PS-rank	CQ	PA	SR	Run (T_1)	L.Run (T_2)
Normal	30	0.0	4.8	4.8	4.8	4.4	5.2	4.4	4.4	4.4	6.2	5.6
		0.1	24.4	26.4	24.8	12.6	21.2	52.0	48.8	47.4	17.2	13.6
		0.2	85.4	88.6	86.8	63.6	84.4	99.8	99.8	99.8	78.4	48.2
		0.3	99.8	100.0	99.8	97.8	99.6	100.0	100.0	100.0	100.0	92.8
	40	0.0	5.0	4.8	5.2	5.2	5.2	4.8	5.0	3.8	4.8	4.8
		0.1	17.4	17.6	17.0	10.8	17.8	58.6	55.6	52.6	17.6	14.6
		0.2	64.6	68.2	66.4	39.6	65.2	100.0	100.0	100.0	85.2	57.4
		0.3	96.4	98.0	97.2	88.0	95.4	100.0	100.0	100.0	100.0	97.2
$t_{(2)}$	30	0.0	3.2	4.4	4.4	4.8	5.8	4.6	3.6	0.4	6.4	5.2
		0.1	16.2	18.6	20.6	10.2	15.8	13.8	9.0	2.0	13.4	11.2
		0.2	66.6	65.4	73.2	37.8	69.6	47.6	34.8	23.6	56.2	41.6
		0.3	95.6	94.6	97.2	82.0	95.4	77.4	64.0	59.8	94.2	87.2
	40	0.0	5.0	3.8	5.4	4.0	5.2	5.4	5.0	0.4	4.2	4.8
		0.1	11.6	12.6	14.0	9.6	27.8	16.0	11.6	2.6	13.2	10.6
		0.2	54.0	42.4	53.4	25.2	53.0	54.6	41.0	23.8	62.8	53.6
		0.3	88.8	73.4	88.8	54.6	84.0	82.2	70.2	65.0	97.4	95.4
Cauchy	30	0.0	1.6	3.0	3.4	3.6	3.6	4.0	2.8	0.0	5.8	4.4
		0.1	10.6	15.0	16.4	10.0	13.8	4.8	3.6	0.0	14.2	11.6
		0.2	47.2	45.8	57.0	30.8	54.0	8.6	6.0	0.6	43.4	47.4
		0.3	84.0	79.6	89.0	58.8	86.4	14.2	9.6	2.4	84.6	87.6
	40	0.0	3.4	4.8	4.6	5.6	5.4	4.6	2.4	0.0	3.4	4.8
		0.1	9.8	11.6	14.8	7.8	14.2	5.6	2.4	0.0	14.2	11.6
		0.2	45.6	29.2	46.6	17.4	45.6	10.4	6.0	0.0	57.6	56.0
		0.3	83.6	49.4	78.2	35.0	77.8	19.2	11.0	2.6	90.4	93.0
Mixture	30	0.0	3.6	4.6	3.2	13.8	3.4	8.6	6.8	8.6	6.8	6.4
		0.1	3.6	5.0	3.4	14.0	4.2	9.8	8.6	9.8	21.8	11.6
		0.2	3.8	5.4	3.8	10.2	4.2	12.6	12.0	12.6	92.6	52.4
		0.3	3.8	6.2	3.4	9.4	5.0	21.2	18.4	21.2	100.0	95.8
	40	0.0	6.2	6.2	4.2	19.6	4.4	8.6	7.4	8.6	5.2	3.4
		0.1	6.2	6.0	4.6	17.4	4.4	9.4	8.6	9.4	26.2	10.2
		0.2	6.0	6.0	4.8	16.2	5.6	14.0	12.8	14.0	98.0	52.8
		0.3	6.2	6.6	4.0	11.2	5.0	20.8	18.0	20.8	100.0	98.4

We considered another example involving an equal mixture of four normal distributions all having the same scatter matrix $\frac{1}{2}\mathbf{I}_d$. The locations of these normal distributions were $(-3+\Delta)\mathbf{1}_d$, $(-1+\Delta)\mathbf{1}_d$, $(1+\Delta)\mathbf{1}_d$ and $(3+\Delta)\mathbf{1}_d$. We carried out our experiment for two choices of d and four choices of Δ as before. In this example, CQ and SR tests had sizes higher than 0.05. Because of near singularity of the estimated dispersion matrix of coordinate-wise signs, the PS-sign test failed to maintain the level property. All other tests had sizes close to the nominal level (see Table 3.2). The powers of the two run tests were substantially higher than those of all other tests considered here. In the case of $\Delta = 0.3$, while the test based on T_1 rejected H_0 on all occasions, and that based on T_2 had power more than 0.95, all other tests had powers less than 0.25.

Next we considered some examples with normal, $t_{(2)}$ and Cauchy distributions, when d was much larger than n. In each case, we generated 20 observations from a distribution having the location parameter $(0.15, \ldots, 0.15)^T$ and the scatter matrix $\mathbf{I}_d$. The powers of different tests were computed based on 500 trials as before. We repeated the experiment for values of d ranging from 3 to 3000, and the results are reported in Fig. 3.9(a)-3.9(c).

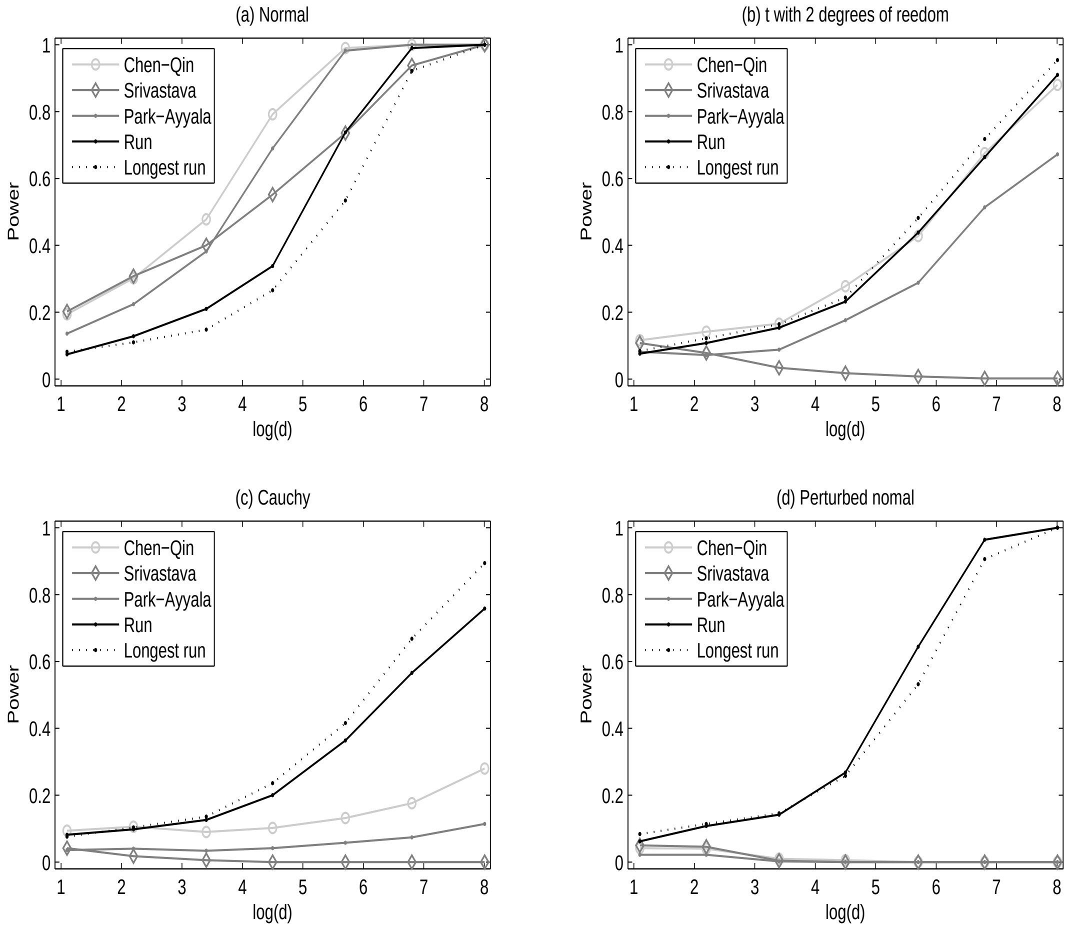


Figure 3.9: Powers of different one-sample tests for varying choices of d.

In these examples, since the location of each variable differs from the origin, one would expect the powers of these tests to tend to 1 as d increases. We observed this phenomenon in most of the cases. In the case of normal distribution, the CQ test had the best performance followed by the PA test. Though the SR test had the highest power for small values of d, in high dimensions, it was outperformed by the CQ test, the PA test, and the test based on T_1. In the case of $t_{(2)}$-distribution, the CQ test and two run tests performed better than PA and SR tests, while the test based on T_2 had an edge in high dimensions. The SR test performed poorly; its power dropped down to zero as d increased. In the examples involving Cauchy distributions, our run tests substantially outperformed all other tests considered here. This is consistent with what we observed in Table 3.2. Again, for large d, the power of the SR test was close to zero.

These examples show the robustness of our tests against heavy-tailed distributions. In cases of multivariate $t_{(2)}$ and Cauchy distributions, especially in the latter case, they had excellent performance when the other tests failed. However, in the examples with normal distributions, CQ and PA tests outperformed them. Even in that case, the situation gets completely changed in the presence of contaminations. We carried out one such experiment, where we generated 20 observations from the normal distribution as before, but perturbed one out of these 20 observations by subtracting 2 from each coordinate. This contamination heavily affected the performance of CQ, PA, and SR tests. All of them had zero power for almost all values of d (see Fig. 3.9(d)). However, the tests based on T_1 and T_2 did not get much affected. The powers of these two tests converged to 1 as before as the dimension increased.

3.7.4 Analysis of PEMS-SF data

We also analyzed the PEMS-SF data available at the UCI machine learning repository. This data set describes the occupancy rate, between 0 and 1, of different car lanes of San Francisco bay area freeways between Jan. 01, 2008 and Mar. 30, 2009. For each day, there is a time series of dimension 963 (the number of sensors) and length $6 \times 24 = 144$ (measurements are sampled every 10 minutes). This data set has separate training and test sets. For our analysis, we used the 126 observations in the test set after removing Saturdays and Sundays. Figure 3.10(a) shows average occupancy rates for different time points of a day computed over 126 days and 963 locations. In this figure, one can observe two modes at 8:30 A.M. and 5:30 P.M., half an hour before and after the office hours. Corresponding to these two time points, we have two distributions of dimension 963. Here, we subtracted one vector (corresponding to 5:30 P.M.) from the other (corresponding to 8:30 A.M.) and carried out our experiment to test whether the location of the difference differs from the origin. The distributions of the difference for different working days of the week are given in Fig. 3.10(b)-3.10(f). Clearly, for some of the sensors, the location differs the origin. So, one would expect the null hypothesis of no difference to be rejected.

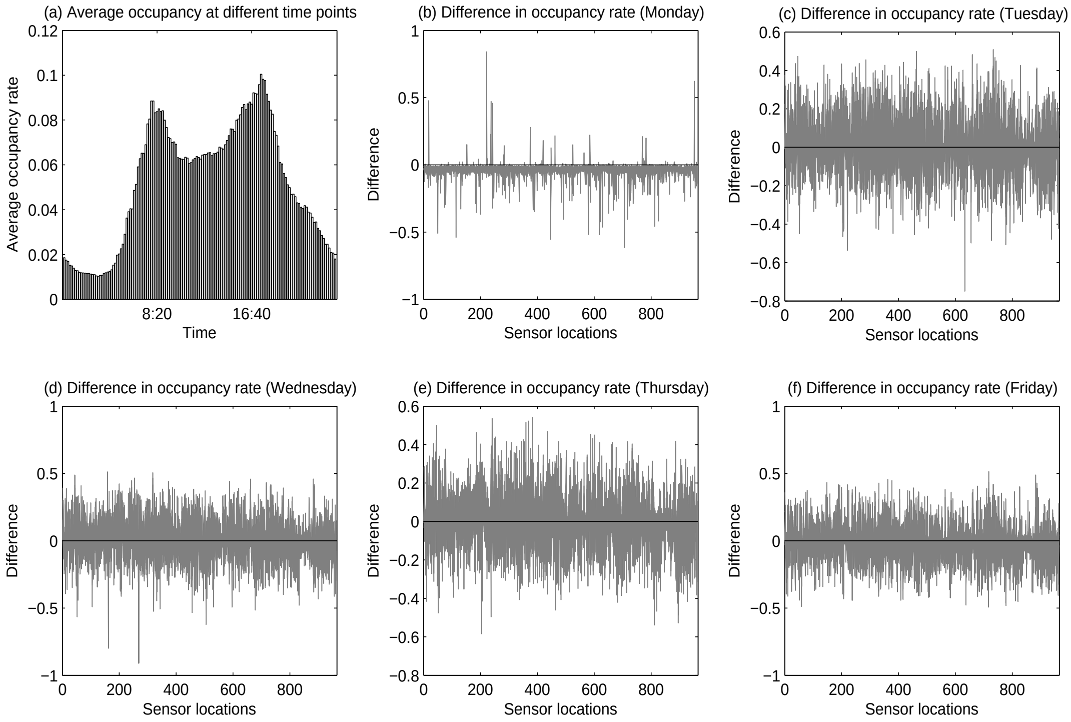


Figure 3.10: Occupancy rates of car lanes of San Francisco bay area freeways.

When we used 126 observations for testing, all five tests (CQ, PA, SR, and two run tests) rejected H_0. Based on that single experiment, it was not possible to compare among different tests. So, we carried out our experiment using random subsets of size 5 and 10. Each experiment was repeated 500 times to estimate the powers of different tests. CQ and SR tests had the highest power 1 both for $n = 5$ and $n = 10$. The tests based on T_1 and T_2 also had power 1 for $n = 10$, but for $n = 5$, they had powers 0.812 and 0.806, respectively. The PA test had power 0.976 for $n = 10$, but in the case of $n = 5$, it could not reject H_0 even on a single occasion. To study the level properties of different tests, along with these 126 observations, we added their negatives to have a data cloud consisting of 252 observations, which is symmetric about the origin. We chose random samples of size 5 and 10 from this cloud to perform these tests, and each experiment was repeated 500 times as before. Both for $n = 5$ and $n = 10$, the tests based on T_1 (0.054 and 0.040, respectively) and T_2 (0.056 and 0.044, respectively) had sizes close to 0.05, but for the PA test, they were much below the nominal level (0.000 and 0.008, respectively). In the case of $n = 10$, CQ and SR tests also had sizes close to 0.05 (0.058 and 0.062, respectively) but in the case of $n = 5$, they failed to maintain the level property and rejected H_0 in 13.6% and 15.8% cases, respectively. This bias towards the alternative hypothesis could be the reason for their high powers for $n = 5$.

3.8 Proofs and mathematical details

PROOF OF THEOREM 3.1: Recall that $T^{SHP}_{n_1,n_2}$ has the same null distribution as the univariate run statistic, and hence $P_{H_0}(T^{SHP}_{n_1,n_2} \leq 3) = n_1!\ n_2!/(n_1 + n_2 - 1)!$. Since $P_{H_0}(T^{SHP}_{n_1,n_2} \leq 3) \leq \alpha$, both $T^{SHP}_{n_1,n_2} = 2$ and $T^{SHP}_{n_1,n_2} = 3$ lead to the rejection of H_0. So, it is enough to prove that $P_{H_A}(T^{SHP}_{n_1,n_2} > 3) \to 0$ as $d \to \infty$.

Define $\upsilon_1 = \sigma_1\sqrt{2}$, $\upsilon_2 = \sigma_2\sqrt{2}$ and $\upsilon_3 = (\sigma_1^2 + \sigma_2^2 + \nu^2)^{1/2}$. As d tends to infinity, $\|\mathbf{x}_i - \mathbf{x}_j\|/\sqrt{d}$ converges in probability to υ_1 for $1 \leq i < j \leq n_1$, $\|\mathbf{y}_i - \mathbf{y}_j\|/\sqrt{d}$ converges in probability to υ_2 for $1 \leq i < j \leq n_2$, and $\|\mathbf{x}_i - \mathbf{y}_j\|/\sqrt{d}$ converges in probability to υ_3 for $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$. Clearly $2\upsilon_3 \geq \upsilon_1 + \upsilon_2$, where the equality holds if and only if $\nu^2 = 0$ and $\sigma_1^2 = \sigma_2^2$. Let $\mathcal{H}$ be a Hamiltonian path in the graph on $n_1 + n_2$ vertices. Now, $\mathcal{H}$ can either (i) start and end with observations from same distribution or (ii) start with an observation from one distribution and end with an observation from the other distribution. Let us consider these two cases separately.

In case (i), $T^{SHP}_{n_1,n_2}$ can take only odd values, i.e., $T^{SHP}_{n_1,n_2} = 2r{+}1$ for some integer $r > 0$. Now, if $\mathcal{H}$ starts and ends with observations from F, $\mathcal{H}$ contains $n_1 - r - 1$ '**XX**'-type edges, $n_2 - r$ '**YY**'-type edges and $2r$ '**XY**'-type edges. So, the total cost of $\mathcal{H}$ converges (in probability) to $(n_1 - r - 1)\upsilon_1 + (n_2 - r)\upsilon_2 + 2r\upsilon_3 = (n_1 - 1)\upsilon_1 + n_2\upsilon_2 + r(2\upsilon_3 - \upsilon_1 - \upsilon_2)$. Similarly, if $\mathcal{H}$ starts and ends with observations from G, the total cost of $\mathcal{H}$ converges to $(n_1 - r)\upsilon_1 + (n_2 - r - 1)\upsilon_2 + 2r\upsilon_3 = n_1\upsilon_1 + (n_2 - 1)\upsilon_2 + r(2\upsilon_3 - \upsilon_1 - \upsilon_2)$. Now, under the condition $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$, we have $2\upsilon_3 > \upsilon_1 + \upsilon_2$. So, irrespective of whether $\mathcal{H}$ starts (and ends) with F or G, the cost of $\mathcal{H}$ is minimum when $r = 1$. Therefore, $\mathcal{H}^*$, the shortest Hamiltonian path cannot have more that three runs, or in other words $P_{H_A}(T^{SHP}_{n_1,n_2} > 3 \mid T^{SHP}_{n_1,n_2} \text{ is odd }) \to 0$ as $d \to \infty$.

In case (ii), we have $T^{SHP}_{n_1,n_2} = 2r$ for some integer $r > 0$. In this case, there are $n_1 - r$ '**XX**'-type edges, $n_2 - r$ '**YY**'-type edges and $2r - 1$ '**XY**'-type edges in $\mathcal{H}$. So, the total cost of $\mathcal{H}$ converges (in probability) to $(n_1 - r)\upsilon_1 + (n_2 - r)\upsilon_2 + (2r - 1)\upsilon_3 =$

$(n_1-1)\upsilon_1+(n_2-1)\upsilon_2+\upsilon_3+(r-1)(2\upsilon_3-\upsilon_1-\upsilon_2)$, which is minimum when $r=1$. Therefore, $P_{H_A}(T^{SHP}_{n_1,n_2}>2 \mid T^{SHP}_{n_1,n_2}$ is even $)\to 0$ as $d\to\infty$. □

PROOF OF THEOREM 3.2: (i) Under the condition $\nu^2>|\sigma_1^2-\sigma_2^2|$, $T^{MST}_{m,n}$ converges in probability to 2 as $d\to\infty$ (see Fig. 3.3(a) and our discussion in Section 3.4). So, there is a subtree $\mathcal{T}_1$ on n_1 vertices correspond to n_1 observations from F and another subtree $\mathcal{T}_2$ on n_2 vertices correspond to n_2 observations from G. These two subtrees are connected by an edge $e=\{uv\}$, where u and v correspond to two vertices of $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively (see Fig. 3.11). Now, let us compute $P(T^{MST}_{n_1,n_2}=2)$ under the permutation distribution. First note that if $\mathcal{T}_1$ and $\mathcal{T}_2$ both contain some vertices labeled as F and some labeled as G, $T^{MST}_{n_1,n_2}$ cannot be 2. So, if $n_1=n_2$, there are only two possibilities. Either all vertices of $\mathcal{T}_1$ or all vertices of $\mathcal{T}_2$ should be labeled as F (see Fig. 3.11(a)). Therefore, in that case, $P(T^{MST}_{n_1,n_2}=2)$ turns out to be $2/\binom{n_1+n_2}{n_1}$. Now, without loss of generality, let us assume $n_1>n_2$. First note that in this case, all vertices of $\mathcal{T}_2$ should have the same label. If all of them are labeled as G, all vertices of $\mathcal{T}_1$ will get label F (see Fig. 3.11(b)). If all vertices of $\mathcal{T}_2$ are labeled as F, to count the number of favourable cases, first note that u must have label F. Also, at most one of its neighbors (vertices that share an edge with u) can have label G. Suppose w ($w\neq v$) is the neighbor having label G. Consider the collection C_w of all vertices in $\mathcal{T}_1$ that connect to u through w. All vertices in this collection (that includes w itself) should have label G, and no other vertices in $\mathcal{T}_1$ can have label G. So, the cardinality of C_w must be n_2. Similarly, the other neighbors of u can have label G only if the corresponding collection has cardinality n_2. So, if the collection corresponding to each of the k neighbors (including v) of u has cardinality n_2, the vertex w can be chosen in $k-1$ different ways, and the total number of favourable cases turns out to be k (including the one, where all vertices of $\mathcal{T}_2$ has label G). If u does not have any neighbor labeled as G, instead of u, the same argument can be used on each of the neighbors of u barring v. Note that in order to have these k favorable cases, we need if $kn_2+1\le n$ or $(n-1)/n_2\ge k$. So, we cannot have more than $\lfloor(n-1)/n_2\rfloor$ favourable cases. Similarly, if $n_2>n_1$, the number of favourable cases cannot exceed $\lfloor(n-1)/n_1\rfloor$. Recall that if $n/n_1=n/n_2=2$ (i.e., $n_1=n_2$), the number of favourable cases is 2. So, combining all these results, under the permutation distribution, we get $P(T^{MST}_{n_1,n_2} = 2) \leq k/\binom{n}{n_1}$, where $k = \max\{\lfloor n/n_1 \rfloor, \lfloor n/n_2 \rfloor\}$. If this upper bound is smaller than α, the power of the MST run test of level α converges to 1 as d tends to infinity.

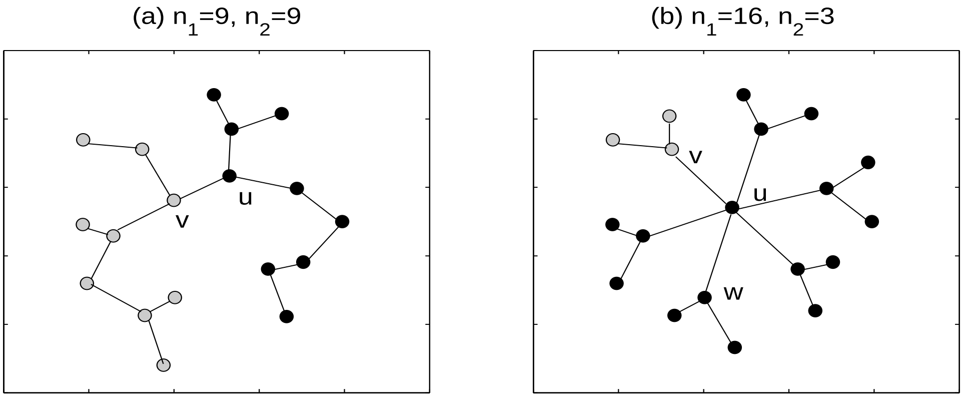


Figure 3.11: Minimal spanning trees with $T^{MST}_{n_1,n_2} = 2$.

(ii) Under the given condition $T = T^{MST}_{n_1,n_2} - 1$ converges to n_1 in probability (see Fig. 3.3(c) and our discussion in Section 3.4). Note that T is a non-negative random variable, and $E(T \mid \mathcal{Z})$, the conditional expectation of the permutation distribution of T given the data $\mathcal{Z}$=$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}, \mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_{n_2}\}$ does not depend on $\mathcal{Z}$ (see e.g., Friedman and Rafsky (1979)), and $E(T \mid \mathcal{Z}) = 2n_1n_2/n \ \forall \ \mathcal{Z}$, where $n = n_1 + n_2$. Therefore, using the Markov inequality, we have $P(T \geq n_1 \mid \mathcal{Z}) \leq 2n_2/n \Rightarrow P(T < n_1 \mid \mathcal{Z}) \geq (n_1 - n_2)/n$. Now, $n_1/n_2 > (1+\alpha)/(1-\alpha)$ implies $(n_1 - n_2)/n > \alpha$ and that completes the proof.□

PROOF OF THEOREM 3.3: Let $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ be two independent copies of $\boldsymbol{\xi}$ and define $\boldsymbol{\eta}_i = -\boldsymbol{\xi}_i$ for $i = 1, 2$. One can check that under (B1°) and (B2°), the weak law of large numbers holds for the sequence $\{\psi(|\xi_1^{(q)} - V^{(q)}|);\ q \geq 1\}$ (the proof is straight forward, and hence it is omitted), where $\mathbf{V} = \boldsymbol{\xi}_2$ or $\boldsymbol{\eta}_2$. Therefore, $\left| d^{-1} \sum_{q=1}^d \psi(|\boldsymbol{\xi}_1^{(q)} - \boldsymbol{\eta}_2^{(q)}|) - d^{-1} \sum_{q=1}^d \psi(|\boldsymbol{\xi}_1^{(q)} - \boldsymbol{\xi}_2^{(q)}|) - \tau_d(\boldsymbol{\theta}) \right| \xrightarrow{P} 0$ as $d \to \infty$. So, if we have n independent copies $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ of $\boldsymbol{\xi}$ and $\tau > 0$, for all $i \neq j$, $P\left[\sum_{q=1}^d \psi(|\boldsymbol{\xi}_i^{(q)} - \boldsymbol{\eta}_j^{(q)}|) > \sum_{q=1}^d \psi(|\boldsymbol{\xi}_i^{(q)} - \boldsymbol{\xi}_j^{(q)}|)\right] \to 1$ as $d \to \infty$. Since h is monotonically increasing and n is finite, as $d \to \infty$, all '$\boldsymbol{\xi}\boldsymbol{\xi}$'-type and '$\boldsymbol{\eta}\boldsymbol{\eta}$'-type distances become smaller than all '$\boldsymbol{\xi}\boldsymbol{\eta}$'-type distances with probability tending to one. So, the shortest covering path $\mathcal{P}_0$ will contain $n-1$ edges connecting either all $\boldsymbol{\xi}_i$s or all $\boldsymbol{\eta}_i$s. As a result, T_0 will take either its minimum value 0 or its maximum value $\sum_{i=1}^n a(i)$. Now, under

H_0, it takes each of these extreme values with probability $1/2^n < \alpha/2$. Therefore, the tests based on T_0 will reject H_0 with probability tending to 1. Since the path $\mathcal{P}_0$ tends to cover either all $\boldsymbol{\xi}_i$s or all $\boldsymbol{\eta}_i$s, T_1 converges in probability to its minimum value 1 and T_2 converges to its maximum value n. Since $P_{H_0}(T_1 \leq 1) = P_{H_0}(T_2 \geq n) = 1/2^{n-1} < \alpha$, the powers of these two tests also converge to unity as d grows to infinity. □

Chapter 4

Tests based on nearest neighbor type coincidences

In the previous chapters, we have used the NN test for the comparison purpose. Schilling (1986a) and Henze (1988) proposed this multivariate two-sample test based on nearest neighbors and proved its consistency in classical asymptotic regime. Under the general alternative $H_A : F \neq G$, the power of this test converges to unity when the dimension is fixed and the sample size tends to infinity (see e.g., Schilling (1986a); Henze (1988)). Though this test can be used even when the dimension exceeds the sample size, it often fails to have the consistency in HDLSS asymptotic regime. In fact, we will see that even when the separation between the two population increases with the dimension, its power may converge to zero as the dimension increases. In order to overcome this limitation of the NN test, in this chapter, we propose and investigate some alternative two-sample tests based on nearest neighbors. These proposed tests also have the large sample consistency under the general alternative $H_A : F \neq G$, but more importantly, unlike NN test, they have the high dimensional consistency under fairly general conditions.

Before we proceed, let us first recall the NN test. It rejects the null hypothesis $H_0 : F = G$ for large values of the statistic $T_{NN,k} = \frac{1}{kn}[\sum_{i=1}^{n_1} \sum_{r=1}^{k} I_{\mathbf{x}_i}(r) + \sum_{i=1}^{n_2} \sum_{r=1}^{k} I_{\mathbf{y}_i}(r)]$, where $I_{\mathbf{z}}(r)$ denotes the indicator variable that takes the value 1 if and only if $\mathbf{z}$ and its r-th ($r \leq k$) nearest neighbor come from the same distribution. For finding the neighbor of $\mathbf{z}$, here we use the leave-one-out method, where $\mathbf{z}$ itself is not considered as its neighbor. However, note that here we use this leave-one-out method only to compute the test statistic for a given value of k, not to choose the value of k based on the data. Since the NN test with $k = 3$ has been reported to perform well in the literature (see e.g., Schilling (1986a)), throughout this thesis, we report all numerical results for $k = 3$. Now let us consider a simple example involving two normal distributions and see how the NN test performs in high dimension. Like Chapter 3, here also we consider an example, where the components of F and G are i.i.d. normal variates. In Chapter 3, we considered examples, where F and G differed either in their locations or in their scales. Here we consider an example, where F and G differ both in locations and scales. While the components of F are i.i.d. $N(0, 1)$, those of G are i.i.d $N(0.2, 1.2)$. We generated 20 observations from each distribution to form the sample and used the NN test to check whether the two distributions differ significantly. We carried out this experiment for different values of d ranging between 2 and 1024, and for each value of d, the experiment was repeated 500 times. Figure 4.1 shows the estimated power of the NN test (i.e., proportion of times it rejected H_0) for various choices of d.

In this example, since each and every component variable provides some evidence against H_0, one would expect the power of any reasonable test to increase to 1 as d increases. Surprisingly, that was not the case for the NN test. Initially its power increased with d, but then it dropped down to zero (see Figure 4.1). Our proposed tests (described in the next section) could overcome this limitation of the NN test. Their powers converged to unity as the dimension increased (see the power curves for tests based on $T_{NN1,k}$ and $T_{NN2,k}$ in Figure 4.1). In the next section, we first investigate the reasons behind the failure of the NN test in the above example, and then we develop our proposed tests based on nearest neighbor type coincidences.

4.1 Construction of new tests based on nearest neighbors

Let $\mathbf{X}_1$, $\mathbf{X}_2$ be two independent observations from F, where the component variables are i.i.d. $N(\mu_1, \sigma_1^2)$, and $\mathbf{Y}_1, \mathbf{Y}_2$ be two independent observations from G, where the component variables are i.i.d. $N(\mu_2, \sigma_2^2)$. Clearly, $\|\mathbf{X}_1 - \mathbf{X}_2\|^2/2\sigma_1^2$ and $\|\mathbf{Y}_1 - \mathbf{Y}_2\|^2/2\sigma_2^2$ both follow chi-square distribution with d degrees of freedom, while $\|\mathbf{X}_1 - \mathbf{Y}_1\|^2/(\sigma_1^2 +$

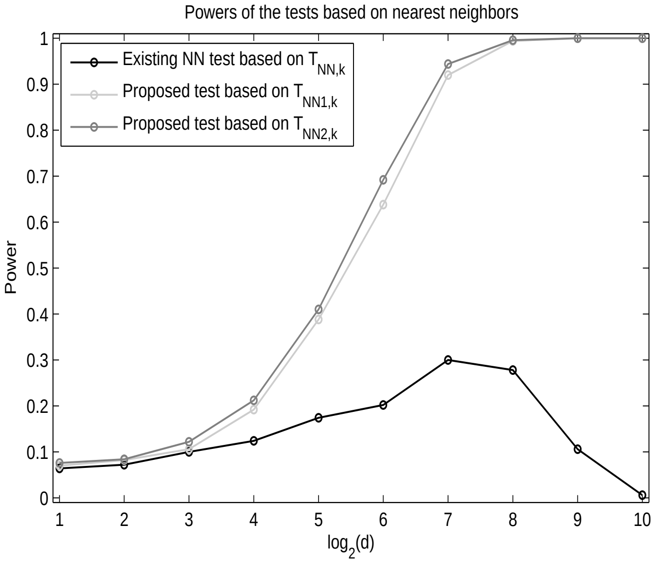


Figure 4.1: Powers of nearest neighbor tests for varying choices of data dimension.

σ_2^2) follows non-central chi-square distribution with d degrees of freedom and the non-centrality parameter $(\mu_2-\mu_1)^2/(\sigma_1^2+\sigma_2^2)$. It is easy to check that as $d\to\infty$, $d^{-1}\|\mathbf{X}_1-\mathbf{X}_2\|^2 \xrightarrow{p} 2\sigma_1^2$, $d^{-1}\|\mathbf{Y}_1-\mathbf{Y}_2\|^2 \xrightarrow{P} 2\sigma_2^2$ and $d^{-1}\|\mathbf{X}_1-\mathbf{Y}_1\|^2 \xrightarrow{P} \sigma_1^2+\sigma_2^2+(\mu_2-\mu_1)^2$. In fact, these above convergence results hold as long as the components of F and G are i.i.d. with finite second moments (follows from weak law of large numbers (WLLN)) or even when they are dependent and non-identically distributed, but satisfy assumptions (A1)-(A3) mentioned in Chapter 2.

In our example, we had $\mu_1=0$, $\mu_2=0.2$, $\sigma_1^2=1$ and $\sigma_2^2=1.2$ leading to $2\sigma_1^2 < \sigma_1^2+\sigma_2^2+(\mu_2-\mu_1)^2 < 2\sigma_2^2$. Therefore, in high dimension, each observation from F had its all $k=3$ neighbors from F, but no observation from G had any of its neighbors from G. As a result, $T_{NN,k}$ attained the value 1/2, which was close to its expected value under H_0. Consequently, the NN test could not reject H_0 even on a single occasion. Now, let us define $T_{1,k}=\frac{1}{n_1k}\sum_{i=1}^{n_1}\sum_{r=1}^{k} I_{\mathbf{X}_i}(r)$, the proportion of neighbors of $\mathbf{X}$-observations that come from F and $T_{2,k}=\frac{1}{n_2k}\sum_{i=1}^{n_2}\sum_{r=1}^{k} I_{\mathbf{Y}_i}(r)$, the proportion of neighbors of $\mathbf{Y}$-observations coming from G. Under H_0, $T_{1,k}$ and $T_{2,k}$ are expected to be close to the proportions $(n_1-1)/(n-1)$ and $(n_2-1)/(n-1)$, which are their respective expected values under H_0. But under the alternative, the deviations $T_{1,k}-E_{H_0}(T_{1,k})$ and $T_{2,k}-E_{H_0}(T_{2,k})$ are expected to be large. Note that the NN test statistic $T_{NN,k}$ is given by $T_{NN,k} = (n_1 T_{1,k} + n_2 T_{2,k})/n$, and hence $T_{NN,k} - E_{H_0}(T_{NN,k}) = \{n_1(T_{1,k} - E_{H_0}(T_{1,k})) + n_2(T_{2,k} - E_{H_0}(T_{2,k}))\}/n$. In our example, $T_{1,k}$ converges to 1 and $T_{2,k}$ converges to 0. So, in this type of examples, while $T_{1,k} - E_{H_0}(T_{1,k})$ turns out to be positive, $T_{2,k} - E_{H_0}(T_{2,k})$ becomes negative. Because of this cancellation of positive and negative terms, depending on the values of n_1 and n_2, the magnitude of $T_{NN,k} - E_{H_0}(T_{NN,k})$ may become very small. As a consequence, the NN test often fails to reject H_0. We also observed this in the scale problem discussed in Section 3.3.

We can easily overcome this problem if we slightly modify $T_{NN,k}$ to avoid this cancellation and use either

$$\begin{aligned} T_{NN1,k} &= \{n_1|T_{1,k} - E_{H_0}(T_{1,k})| + n_2|T_{2,k} - E_{H_0}(T_{2,k})|\}/n \quad \text{or} \\ T_{NN2,k} &= \{n_1[T_{1,k} - E_{H_0}(T_{1,k})]^2 + n_2[T_{2,k} - E_{H_0}(T_{2,k})]^2\}/n \end{aligned}$$

as test the statistic. A similar idea was also used by Liu et al. (2010) in a slightly different context. Like the NN test, here also we reject H_0 for large values of the test statistics, where the cut offs are determined using the permutation principle. Note that if we define $\vartheta_i = T_{i,k} - E_{H_0}(T_{i,k})$ and $w_i = n_i/n$ for $i = 1, 2$, we have $T^{\circ}_{NN,k} = T_{NN,k} - E_{H_0}(T_{NN,k}) = w_1\vartheta_1 + w_2\vartheta_2$, while $T_{NN1,k}$ and $T_{NN2,k}$ can be expressed as

$$\begin{aligned} T_{NN1,k} = &\quad w_1|\vartheta_1| + w_2|\vartheta_2| = T^{\circ}_{NN,k} + w_1(|\vartheta_1| - \vartheta_1) + w_2(|\vartheta_2| - \vartheta_2), \\ T_{NN2,k} = &\quad w_1\vartheta_1^2 + w_2\vartheta_2^2 = (T^{\circ}_{NN,k})^2 + w_1w_2(\vartheta_1 - \vartheta_2)^2. \end{aligned}$$

Therefore, under the alterative H_A, if both ϑ_1 and ϑ_2 are positive with very high probability, $T_{NN1,k}$ and $T^{\circ}_{NN,k}$ often take same values. But, since $T_{NN1,k}$ is stochastically larger than $T^{\circ}_{NN,k}$, the cut-off obtained from the permutation distribution of $T_{NN1,k}$ is expected to be larger than that for the test based on $T^{\circ}_{NN,k}$. So, in such cases, the test based on $T_{NN,k}$ or $T^{\circ}_{NN,k}$ can outperform the test based on $T_{NN1,k}$. For instance, if two distributions F and G differ only in their locations, the NN test is expected to yield better performance than the test based on $T_{NN1,k}$. But if either ϑ_1 or ϑ_2 takes negative values with high probability, as it was in our example, the test based on $T_{NN1,k}$ is expected to outperform the NN test, and we have observed it in Figure 4.1.

From the expression of $T_{NN2,k}$, it is clear that $T_{NN2,k}$ is stochastically larger than $(T^{\circ}_{NN,k})^2$. Therefore, under H_A, if the difference between ϑ_1 and ϑ_2 is small with high probability, because of the same reason as described above, the test based on $T^{\circ}_{NN,k}$ or $(T^{\circ}_{NN,k})^2$ may perform better than that based on $T_{NN2,k}$. For instance, in a location problem with $n_1 = n_2$, the test based on $T_{NN2,k}$ may have lower power than the NN test. But if F and G differ also in their scatters and/or shapes, this additional term $T_{NN2,k} - (T^{\circ}_{NN,k})^2$ may play a significant role to improve the performance of the resulting test. We have observed this in Figure 4.1, and we will see it again in subsequent sections.

4.2 Behavior of proposed tests for HDLSS data

Now, we carry out a theoretical investigation to study the power properties of the NN test and the proposed tests based on $T_{NN1,k}$ and $T_{NN2,k}$ when the sample size remains fixed and the dimension grows to infinity. For this investigation, we consider $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})^T$ and $\mathbf{Y} = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(d)})^T$ to be independent, and they satisfy the assumptions (A1)-(A3) stated in Chapter 2.

We have seen that if $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}$ are n_1 independent observations from F and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$ are n_2 independent observations from G, under (A1)-(A3) as $d \to \infty$,

(a) $d^{-\frac{1}{2}} \|\mathbf{x}_i - \mathbf{x}_j\| \xrightarrow{P} \sigma_1\sqrt{2}$ for $1 \leq i < j \leq n_1$.

(b) $d^{-\frac{1}{2}} \|\mathbf{y}_i - \mathbf{y}_j\| \xrightarrow{P} \sigma_2\sqrt{2}$ for $1 \leq i < j \leq n_2$.

(c) $d^{-\frac{1}{2}} \|\mathbf{x}_i - \mathbf{y}_j\| \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2}$ for $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$,

where σ_1^2, σ_2^2 and ν^2 are the limiting values (as $d \to \infty$) of $d^{-1}\sum_{q=1}^{d} VarX^{(q)}$, $d^{-1}\sum_{q=1}^{d} VarY^{(q)}$ and $d^{-1}\sum_{q=1}^{d}[E(X^{(q)} - Y^{(q)})]^2$, respectively (as in (A3)). Hall et al. (2005) showed that if $\nu^2 < |\sigma_1^2 - \sigma_2^2|$, the nearest neighbor classifier (see e.g., Cover and Hart (1967), Fix and Hodges (1989)) fails in high dimension, where it tends to classify all observations to a single class. The NN test fails under the same condition. The following theorem shows that in such cases, depending on the values of n_1 and n_2, its power may even converge to zero.

THEOREM 4.1: *Suppose that F and G satisfy (A1)-(A3) and $\nu^2 < \sigma_1^2 - \sigma_2^2$ (interchange F and G, if required, and also interchange n_1 and n_2 accordingly). If $n_2/(n-1) < (1-\alpha)/2$ and $k < \min\{n_1, n_2\}$, the power of a level α test based on $T_{NN,k}$ converges to zero as $d \to \infty$.*

The proof of the theorem is given in Section 4.6. Note that, Theorem 4.1 gives only a sufficient condition under which the NN test fails. This test may fail in many other situations. For instance, in the example in Figure 4.1, we had $n_2/(n-1) > 1/2 > (1-\alpha)/2$, but the power of the NN test converged to 0 as d increased.

Now, let us look at the power properties of the tests based on $T_{NN1,k}$ and $T_{NN2,k}$. If F and G satisfy (A1)-(A3) and $\nu^2 \neq |\sigma_1^2 - \sigma_2^2|$, depending on the values of σ_1^2, σ_2^2 and ν, $(T_{1,k}, T_{2,k})$ converges to $(1,0), (0,1)$ or $(1,1)$ in probability. Therefore, $T_{NN1,k}$ (respectively, $T_{NN2,k}$) converges to a value not smaller than $\mathbb{K}^{(1)}_{n_1,n_2} = n_0/n$ (respectively, $\mathbb{K}^{(2)}_{n_1,n_2} = (n_0-1)^2/(n-1)^2$), where $n_0 = \min\{n_1, n_2\}$. So, if we can show that under the permutation distribution $E(T^2_{NN1,k})/\mathbb{K}^{(1)2}_{n_1,n_2} < \alpha$ (respectively, $E(T_{NN2,k})/\mathbb{K}^{(2)}_{n_1,n_2} < \alpha$), the consistency of the test based on $T_{NN1,k}$ (respectively, $T_{NN2,k}$) follows from the Markov inequality. If $\nu^2 = |\sigma_1^2 - \sigma_2^2| > 0$, we can comment on the convergence of either $T_{1,k}$ or $T_{2,k}$ but not on both. In such cases, for the consistency of these tests, it is enough to show that $E(T^2_{NN1,k})/\mathbb{C}^{(1)2}_{n_1,n_2} < \alpha$ and $E(T_{NN2,k})/\mathbb{C}^{(2)}_{n_1,n_2} < \alpha$, where $\mathbb{C}^{(i)}_{n_1,n_2} = n_0\ \mathbb{K}^{(i)}_{n_1,n_2}/n$ for $i = 1,2$. Now, note that $E(T_{NN2,k}) = w_1 Var(T_{1,k}) + w_2 Var(T_{2,k})$ and $E(T^2_{NN1,k}) \leq 2(w_1 Var(T_{1,k}) + w_2 Var(T_{2,k}))$. So, if we can make $w_1 Var(T_{1,k}) + w_2 Var(T_{2,k})$ sufficiently small for some choices of (n_1, n_2), for that sample size, the powers of these tests will converge to 1 as d increases.

It can be shown that (see Section 4.6.1) under the permutation distribution, $w_1 Var(T_{1,k}) + w_2 Var(T_{2,k}) \leq \psi_1(n_1,n_2,k)\sum_{j=1}^n \delta_j^2 + \psi_2(n_1,n_2,k)$, where $\psi_1(n_2,n_2,k)$ is of the order $O(1/(n^2k^2))$, $\psi_2(n_1,n_2,k)$ is of the order $O(1/n)$ and δ_j $(j = 1,2,\ldots,n)$ is the number of observations, which have $\mathbf{z}_j$ (here $\mathbf{z}_j = \mathbf{x}_j$ for $j = 1,2,\ldots,n_1$ and $\mathbf{z}_{n_1+j} = \mathbf{y}_j$ for $j = 1,2,\ldots,n_2$) as one of its k neighbors. For any fixed d, the δ_js are bounded, but this bound increases with d. Given that $\sum_{j=1}^n \delta_j = nk$ and $0 \leq \delta_j \leq n-1$ for all $j = 1,2,\ldots,n$, $\sum_{j=1}^n \delta_j^2$ can be as large as $n(n-1)k$ (note that $\sum_{j=1}^n \delta_j^2 \leq (n-1)\sum_{j=1}^n \delta_j \leq n(n-1)k$). Therefore, to make $\psi_1(n_1,n_2,k)\sum_{j=1}^n \delta_j^2$ sufficiently small, one can consider k to be an appropriate increasing function of n (e.g., $k = \sqrt{n}$), and n to be reasonably large.

THEOREM 4.2: *Suppose that F and G satisfy (A1)-(A3), where $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$. If $\phi_1(n_1,n_2,k) = \psi_1(n_1,n_2,k)n(n-1)k + \psi_2(n_1,n_2,k) < \alpha\mathbb{C}^{(2)}_{n_1,n_2}$, the power of a level*

α ($0 < \alpha < 1$) test based on $T_{NN2,k}$ converges to unity as the dimension increases. If $\phi_1(n_1, n_2, k) < \alpha \mathbb{C}^{(1)2}_{n_1,n_2}/2$, we also have this convergence of power for a level α test based on $T_{NN1,k}$.

The proof of the theorem follows immediately from our above discussion. However, Hall et al. (2005) rightly pointed out that in HDLSS setting, where we deal with high dimensional data and have very limited number of observations at our disposal, it is not a great idea to use larger values of k. We also observed the same during our analysis of simulated and real data sets. In most of the cases, larger values of k did not lead to any substantial improvement over the results obtained using $k = 3$, and in some cases, they made the thing worse. So, we do not recommend this method. In order to make our tests consistent in HDLSS set up, we adopt a different strategy. We put a bound on the δ_js, the in-degrees of the $\mathbf{z}_j$s. Note that if $\delta_j \leq t$ ($k \leq t \leq n-1$), under the condition $\sum_{i=1}^{n} \delta_j = kn$, $\sum_{j=1}^{n} \delta_j^2$ cannot exceed nkt (note that $\sum_{j=1}^{n} \delta_j^2 \leq t \sum_{j=1}^{n} \delta_j \leq nkt$). Therefore, if n is not too small, using a small value of t, we can make $\psi_1(n_1, n_2, k) \sum_{j=1}^{n} \delta_j^2$ sufficiently small and hence make the tests based on $T_{NN1,k}$ and $T_{NN2,k}$ consistent in HDLSS situations.

THEOREM 4.3: *Suppose that F and G satisfy (A1)-(A3), where $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$. If the in-degrees of the observations are bounded by t and $\phi_2(n_1, n_2, k, t) = \psi_1(n_1, n_2, k)nkt + \psi_2(n_1, n_2, k) \leq \alpha \mathbb{C}^{(2)}_{n_1,n_2}$, the power of a level α ($0 < \alpha < 1$) test based on $T_{NN2,k}$ converges to unity as d increases. If $\phi_2(n_1, n_2, k, t) < \alpha \mathbb{C}^{(1)2}_{n_1,n_2}/2$, we also have this convergence for a level α test based on $T_{NN1,k}$.*

The proof of the theorem also follows from our above discussion. To construct a nearest neighbor graph with in-degrees bounded by t, we start with the smallest pairwise distance. If the distance between $\mathbf{z}_i$ and $\mathbf{z}_j$ is the smallest, we consider $\mathbf{z}_i$ as a neighbor of $\mathbf{z}_j$ and vice versa. In that case, in-degrees and out-degrees of both $\mathbf{z}_i$ and $\mathbf{z}_j$ are increased by 1. Next, we consider other pairwise distances one by one in increasing order. Suppose that the pairwise distance between $\mathbf{z}_r$ and $\mathbf{z}_s$ is chosen at a stage. Now, if the out-degree of $\mathbf{z}_r$ (respectively, $\mathbf{z}_s$) is smaller than k and the in-degree of $\mathbf{z}_s$ (respectively, $\mathbf{z}_r$) is smaller than t, we consider $\mathbf{z}_s$ (respectively, $\mathbf{z}_r$) as a neighbor of $\mathbf{z}_r$ (respectively, $\mathbf{z}_s$) and modify the out-degree of $\mathbf{z}_r$ (respectively, $\mathbf{z}_s$) and the in-degree of $\mathbf{z}_s$ (respectively, $\mathbf{z}_r$), accordingly. We stop when the in-degrees and the out-degrees of all $\mathbf{z}_i$s $(i = 1, 2, \ldots, n)$ reach the respective upper bounds t and k.

4.3 Results from the analysis of simulated data sets

We analyzed six simulated data sets to compare the performance of our proposed tests with the NN test (see e.g., Schilling (1986a); Henze (1988)) and some other existing two sample tests that can be used in HDLSS situations. In particular, we used the MST run test (see Friedman and Rafsky (1979)), Hall and Tajvidi (2002)'s test (HT test) based on nearest neighbors and the Cramer test (see Baringhaus and Franz (2004)). In all these cases, we used conditional tests based on 500 permutations. In each of these examples, we used $n_1 = n_2 = 20$. Each experiment was repeated 500 times to estimate the powers of different tests, and they are reported in Table 4.1 for three choices of d.

In most of these examples, the test based on $T_{NN2,k}$ performed slightly better than that based on $T_{NN1,k}$. Therefore, in Table 4.1, we have reported the results for the proposed test based on $T_{NN2,k}$ only. Here, we have used both versions of the test; the usual one and the one where we put an upper bound on the in-degrees, as discussed in Section 4.2. For the implementation of the bounded version, we chose the largest possible upper bound t that ensures $w_1 Var(T_{1,k}) + w_2 Var(T_{2,k}) < \alpha \mathbb{C}^{(2)}_{n_1,n_2}$. In Table 4.1, the usual version is referred to as $T_{NN2,k}$, while the bounded version is referred to as $T^{Bound}_{NN2,k}$.

In Example 1, two normal distributions F and G had the same scatter matrix $\mathbf{\Sigma} = ((\sigma_{ij}))$ with $\sigma_{ij} = (-0.5)^{|i-j|}$, but they differed in their locations. While F was symmetric about the origin, G had the center at $(0.3, 0.3, \ldots, 0.3)^T$. In this example, the Cramer test had the best performance followed by the NN test. As expected (see our discussion in Section 4.1), our proposed tests could not compete with the NN test in this location problem. They could only beat the HT test.

We observed a diametrically opposite picture in Example 2, where two normal distributions had the same mean vector $(0, 0, \ldots, 0)^T$ but different scatter matrices $\mathbf{\Sigma}$ (as in Example 1) and $1.3\,\mathbf{\Sigma}$. In this example, the HT test had the best performance closely followed by the proposed tests. In view of Theorem 4.1, the NN test was expected to have poor performance in this example. The MST run test and the Cramer test also failed to perform well. The reason for the failure of the MST run test in the scale problem has already been discussed in Chapter 3. We will discuss about the reason for such performance of the Cramer test in the next chapter.

Table 4.1: Observed powers (in %) of two-sample tests in simulated data sets

	Example-1						Example-2					
d	NN	HT	MST	Cramer	$T_{NN,2}$	$T_{NN,2}^{Bound}$	NN	HT	MST	Cramer	$T_{NN,2}$	$T_{NN,2}^{Bound}$
25	49.8	9.6	30.4	75.6	22.2	23.0	6.4	47.8	3.8	7.0	44.8	44.4
50	68.2	17.8	44.6	97.8	26.6	27.2	5.4	85.8	3.0	10.6	76.8	76.0
100	89.8	36.8	67.2	100.0	46.6	47.8	4.8	99.2	2.4	13.6	95.8	96.6
	Example-3						Example-4					
d	NN	HT	MST	Cramer	$T_{NN,2}$	$T_{NN,2}^{Bound}$	NN	HT	MST	Cramer	$T_{NN,2}$	$T_{NN,2}^{Bound}$
25	95.6	19.4	90.2	12.0	70.8	73.0	9.2	17.2	3.6	4.8	35.4	36.0
50	97.4	29.8	84.6	11.2	71.6	78.6	9.4	29.6	4.8	6.6	64.4	64.8
100	96.2	45.4	82.4	12.6	77.4	86.0	6.6	58.8	2.6	5.4	97.0	97.4
	Example-5						Example-6					
d	NN	HT	MST	Cramer	$T_{NN,2}$	$T_{NN,2}^{Bound}$	NN	HT	MST	Cramer	$T_{NN,2}$	$T_{NN,2}^{Bound}$
25	3.8	98.6	0.2	44.0	99.6	99.4	10.8	51.4	7.6	3.2	55.2	55.2
50	0.0	99.2	0.0	57.4	100.0	100.0	20.0	73.2	8.4	3.2	85.6	87.4
100	0.0	99.6	0.0	76.2	100.0	100.0	37.4	93.0	14.4	4.2	98.6	99.0

Next, we considered some examples (Examples 3-6) with $\nu = 0$ and $\sigma_1^2 = \sigma_2^2$, where ν, σ_1^2 and σ_2^2 have the same meaning as in (A3). We used these examples to investigate how the proposed tests perform when the assumptions of Theorems 4.2 and 4.3 do not hold. Example 3 dealt with two multivariate normal distributions, where F and G differed only in their correlation structures. While the scatter matrix of F had the (i, j)-th entry $0.5^{|i-j|}$, that of G had the (i, j)-th entry $(-0.5)^{|i-j|}$. The next three examples (Examples 4-6) were taken from Chapter 3 (see Examples 4-6 in Section 3.5).

In Example 3, the NN test had the highest power. The MST run test and our proposed tests also performed well, and they had higher powers than the other two tests considered here. The power of the Cramer test was not satisfactory at all. In Example 4, along with the Cramer test, the NN test and the MST run test also had miserable performance. Their powers were close to or even lower than the nominal level. In this example, our proposed tests outperformed all of their competitors. The HT test had somewhat better performance than NN, MST run and Cramer tests, but their powers were not at all comparable to those of our proposed tests. Our proposed tests had excellent performance in Example 5 as well. While the MST run test and the NN test both failed to reject H_0 even on a single occasion, they rejected H_0 in all cases. The performance of the HT test was also comparable to the proposed tests, but the Cramer test had relatively low power. Our proposed test outperformed its all competitors also in Example 6. The MST run test and the Cramer test had miserable performance in this example. Only powers of the HT test were somewhat comparable.

4.4 Results from the analysis of benchmark data sets

We analyzed five benchmark data sets for further evaluation of the proposed methods. Descriptions of Arcene data, Sonar data and Trace data have been given in the previous chapters. The other two data sets, ECG data and Gun-point data were obtained from the University of California, Riverside time series classification/clustering page http://www.cs.ucr.edu/∼eamonn/time_series_data/. For the Trace data set, which contains observations from four classes, we considered two two-sample problems as before, one between class-1 and class-2 and the other between class-3 amd class-4. We refer to them as Trace data-1 and Trace data-2, respectively. For each of these data sets, we repeated the experiment several times based on different random subsets of the same size chosen from the whole data set maintaining (as close as possible) the proportions of observations from the two distributions. In the case of Arcene data, we considered 100 random subsets to estimate the powers of different tests. In all other cases, we used 500 subsets. The results for different subset sizes are shown in Figure 4.2.

In Section 4.3, we have seen that barring Example-3, in all other cases, $T_{NN2,k}$ and $T_{NN2,k}^{Bound}$ led to similar results (see Table 4.1). So, instead of reporting the results for both of them, here we report the result for the test based on $T_{NN2,k}$ only. This is chosen because of its computational advantage. Here also, we use the NN test, the MST run test, the Cramer test and the HT test for comparison.

In the Sonar data set, barring the HT test, all other methods performed well. Among them, the NN test had the highest power for all choices of the sample size. The MST run test and the proposed test had almost similar performance. They outperformed the Cramer test for sample size larger than 40.

The ECG data contain measurements of cardiac electrical activity as recorded from electrodes at various locations on the body. Each observation contains 96 measurements recorded by one electrode during heartbeat. Observations were analyzed by domain experts, who tagged 133 observations as normal and the rest 67 as abnormal. In this data set, the proposed test and the NN test outperformed their competitors, while the latter had an edge. The Cramer test had good performance when the sample size was small, but its power did not increase appreciably when larger samples were used.

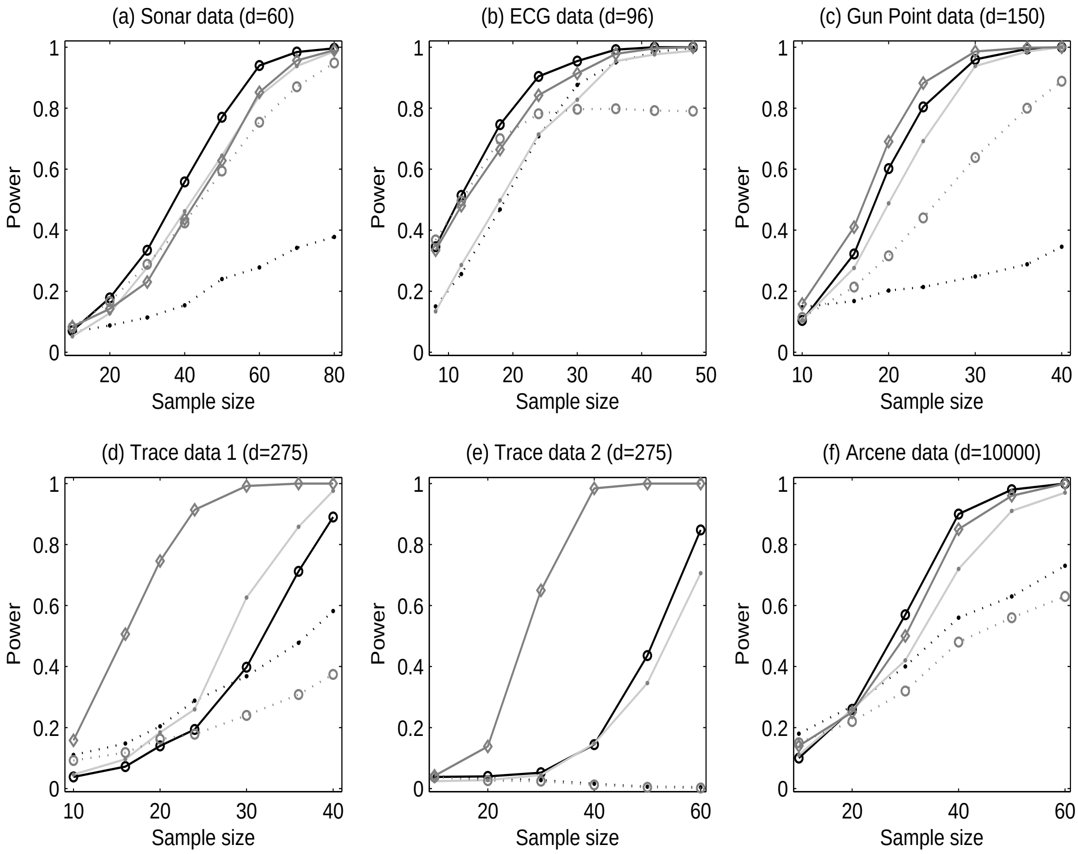


Figure 4.2: Powers of different two-sample tests in benchmark data sets (NN test= black solid curve, HT test= black dotted curve, MST run test=light grey solid curve, Cramer test= dark grey dotted, Proposed test= dark grey solid curve).

The Gun Point data set comes from the video surveillance domain. It contains 100 observations for each of the two classes, 'Gun draw' and 'Point'. At the beginning, the actors have their hands by their sides. In the first case, actors draw a replicate gun from a hip-mounted hostler, point it at a target for one second and then return it to the hostler. In the second case, the actors point their index fingers to a target for one second and return their hand to their sides. In both cases, the centroid of the hand was tracked 150 times during the process. In this data set, the proposed test had the best performance. The performance of the NN test and the MST run test were also comparable, but the Cramer test had relatively low power. The HT failed to yield satisfactory performance in this data set.

In both examples with the Trace data, i.e., Trace data-1 and Trace data-2, the proposed test outperformed all other tests considered here. In Trace data-2, the Cramer test and the HT test had miserable performance. They had poor performance in Trace data-1 as well.

Finally, we analyzed the Arcene data set. In this data set, the NN test and the proposed test had comparable performance, and they outperformed their competitors for sample size larger than 20. The HT test had the highest power for sample size smaller than 20, but this test and the Cramer test failed to compete with other test procedures when larger samples were used.

4.5 Large sample behavior of proposed tests

Now, we study the asymptotic behavior of our proposed tests when the dimension remains fixed and the sample size grows to infinity. First note that if the sample size is large, instead of conditional tests based on the permutation principle, one can use the tests based on the asymptotic null distributions of $T_{NN1,k}$ and $T_{NN2,k}$, which are given by the following theorem.

THEOREM 4.4: *Suppose that n grows to infinity in such a way that $n_1/n \to \lambda$ for some $\lambda \in (0,1)$. Then, for any fixed dimension d and any fixed k,*

(a) $\sqrt{n}T_{NN1,k}$ *is asymptotically distributed as a sum of correlated half normals.*

(b) $nT_{NN2,k}$ *is asymptotically distributed as a weighted sum of independent chi squares.*

From the proof of Theorem 4.4 (given in Section 4.6), it is clear that in order to implement the tests based on the large sample distributions of $T_{NN1,k}$ and $T_{NN2,k}$, one needs to find a consistent estimate for $\mathbf{\Sigma}_w$, the asymptotic dispersion matrix of

$\mathcal{W} = \sqrt{n}\Big[T_{1,k} - E_{H_0}(T_{1,k}) \quad T_{2,k} - E_{H_0}(T_{2,k})]^T$ under H_0. A brief description of this estimation procedure is given in Section 4.6.2. Interestingly, the elements of $\mathbf{\Sigma}_w$ do not depend on the common underlying distribution. So, these tests based on $T_{NN1,k}$ and $T_{NN,2}$ are asymptotically distribution-free. Note that when d is fixed, in-degrees of the observations are automatically bounded. So, here we do not need to consider the bounded versions of these tests separately. The following theorem shows the large sample consistency of these proposed tests, and the proof is given in Setion 4.6.

THEOREM 4.5: *Suppose that F and G have continuous densities f and g, respectively. If n grows to infinity in such a way that $n_1/n \to \lambda$ for some $\lambda \in (0,1)$, the powers of the proposed large sample tests based on $T_{NN1,k}$ and $T_{NN2,k}$ converge to unity as n increases.*

4.6 Proofs and mathematical details

PROOF OF THEOREM 4.1: Under the given condition, each and every observation, irrespective of whether it is from F or G, have all of its neighbors from G with probability tending to one (see our discussion in Section 4.2) as d increases. So, $T_{NN,k} \xrightarrow{P} n_2/n$ as $d \to \infty$. Hence, it is enough to show that the cut-off under the permutation distribution of $T_{NN,k}$ is bigger than n_2/n. Now, given a sample, let $\mathbb{E}$ denote the expectation under the permutation distribution $\mathbb{P}$. Following Schilling (1986a), one can check that this expectation is independent of the sample, and $\mathbb{E}(T_{NN,k}) = \frac{n_1(n_1-1)+n_2(n_2-1)}{n(n-1)}$. Now, consider the non-negative random variable $Y_{NN,k} = 1 - T_{NN,k}$. Since $\mathbb{E}(Y_{NN,k}) = 2n_1n_2/n(n-1)$, using Markov inequality, we get $\mathbb{P}(Y_{NN,k} \geq n_1/n) \leq 2n_2/(n-1)$. So, $n_2/(n-1) < (1-\alpha)/2$ implies $\mathbb{P}(T_{NN,k} \leq n_2/n) = \mathbb{P}(Y_{NN,k} \geq n_1/n) < 1-\alpha$. □

PROOF OF THEOREM 4.4: Let us define $\mathbf{z}_i = \mathbf{x}_i$ for $i = 1, 2, \ldots, n_1$ and $\mathbf{z}_{n_1+i} = \mathbf{y}_i$ for $i = 1, 2, \ldots, n_2$. Also define $\Omega_1^0 = \{1, 2, \ldots, n_1\}$ and $\Omega_2^0 = \{n_1+1, n_1+2, \ldots, n_1+n_2\}$. For $r = 1, 2$ and $j = 1, 2, \ldots, k$, let $\mathbb{S}_{r,j}$ denote the number of observations $\mathbf{z}_i$ ($i \in \Omega_r^0$) which have exactly j of its k nearest neighbors from Ω_r^0. Rogers (1976) showed that the vector of $\mathbb{S}_{r,j}$ values, when appropriately centered and scaled, asymptotically follows a multivariate normal distribution under H_0 with limiting covariance structure independent of $F = G$. Since $T_{1,k} = \frac{1}{n_1k}\sum_{j=1}^{k} j\mathbb{S}_{1,j}$ and $T_{2,k} = \frac{1}{n_2k}\sum_{j=1}^{k} j\mathbb{S}_{2,j}$ are finite linear combinations of the $\mathbb{S}_{r,j}$s, under the null hypothesis H_0, $\mathcal{W} = \Big[\mathbb{W}_1 \quad \mathbb{W}_2\Big]^T = \sqrt{n}\Big[T_{1,k} - E_{H_0}(T_{1,k}) \quad T_{2,k} - E_{H_0}(T_{2,k})\Big]^T$ is asymptotically bivariate normal with its centre at the origin.

(a) Let $\mathbf{\Sigma}_w$ be the dispersion matrix of the asymptotic null distribution of $\mathcal{W}$ and define $\mathcal{V} = [\mathbb{V}_1 \quad \mathbb{V}_2]^T = \mathbf{\Sigma}_w^{-1/2}\mathcal{W}$. Clearly, $\mathcal{V}$ is asymptotically distributed as a standard normal vector (i.e., $\mathbb{V}_1$ and $\mathbb{V}_2$ are asymptotically $N(0,1)$ variables and they are asymptotically independent). Let c_{ij} $(i,j = 1,2)$ be the (i,j)-th element of $\mathbf{\Sigma}_w^{1/2}$. Note that $\sqrt{n}T_{NN1,k}$ can be expressed as $\sqrt{n}T_{NN1,k} = \frac{n_1}{n}|\mathbb{W}_1| + \frac{n_2}{n}|\mathbb{W}_2|$, where $n_1/n \to \lambda$ and $n_2/n \to 1-\lambda$ as $n \to \infty$. So, $\sqrt{n}T_{NN1,k}$ is asymptotically distributed as $|\mathbb{U}_1|+|\mathbb{U}_2|$, where $\mathbb{U}_1 = \lambda(c_{11}\Psi_1 + c_{12}\Psi_2)$, $\mathbb{U}_2 = (1-\lambda)(c_{21}\Psi_1 + c_{22}\Psi_2)$ for Ψ_1 and Ψ_2 being two independent standard normal variates. Clearly, both $\mathbb{U}_1$ and $\mathbb{U}_2$ are zero mean normal variables (hence $|\mathbb{U}_1|$ and $|\mathbb{U}_2|$ are half-normals) but they are correlated, where the correlation depends only on the elements of $\mathbf{\Sigma}_w$ and λ.

(b) Define $\mathcal{W}_0 = [\sqrt{n_1/n}\ \mathbb{W}_1 \quad \sqrt{n_2/n}\ \mathbb{W}_2]^T$. Since $\sqrt{n_1/n} \to \lambda$ $(0 < \lambda < 1)$ as $n \to \infty$, under H_0, $\mathcal{W}_0$ asymptotically follows a bivariate normal distribution symmetric about $\mathbf{0}$. Clearly, the elements of $\mathbf{\Sigma}_w^0$, the scatter matrix of the asymptotic distribution of $\mathcal{W}_0$, can be expressed in terms of the elements of $\mathbf{\Sigma}_w$ and λ. If λ_1^* and λ_2^* $(\lambda_1^* \geq \lambda_2^* > 0)$ are two eigenvalues of $\mathbf{\Sigma}_w^0$, it can be expressed as $\mathbf{\Sigma}_w^0 = \mathbf{H}\mathbf{\Lambda}^*\mathbf{H}^T$, where $\mathbf{H}$ is an orthogonal matrix and $\mathbf{\Lambda}^* = \text{Diag}(\lambda_1^*, \lambda_2^*)$. Define $\mathbf{L} = (L_1 \ L_2)^T = \mathbf{H}^T\mathcal{W}_0$. Clearly, $\mathbf{L}$ is asymptotically normal with the location $\mathbf{0}$ and the scatter $\mathbf{\Lambda}^*$. So, L_1^2/λ_1^* and L_2^2/λ_2^* are asymptotically independent chi-square variables with one degree of freedom. Now, the proof follows from the fact that $nT_{NN2,k} = \mathcal{W}_0^{'}\mathcal{W}_0 = \mathbf{L}^{'}\mathbf{L} = L_1^2 + L_2^2 = \lambda_1^*(L_1^2/\lambda_1^*) + \lambda_2^*(L_2^2/\lambda_2^*)$. □

PROOF OF THEOREM 4.5: Recall the definitions of $T_{1,k} = n_1^{-1}k^{-1}\sum_{i=1}^{n_1}\sum_{r=1}^{k} I_{\mathbf{X}_i}(r)$ and $T_{2,k} = n_2^{-1}k^{-1}\sum_{i=1}^{n_2}\sum_{r=1}^{k} I_{\mathbf{Y}_i}(r)$ given in Section 4.1. Note that

$$\begin{aligned} Var(T_{1,k}) = \;& n_1^{-2}k^{-2}\textstyle\sum_{i=1}^{n_1}\sum_{r=1}^{k} Var(I_{\mathbf{X}_i}(r)) \\ & +n_1^{-2}k^{-2}\textstyle\sum_{i=1}^{n_1}\sum_{r=1}^{k}\sum_{s=1 s\neq r}^{k} Cov(I_{\mathbf{X}_i}(r), I_{\mathbf{X}_i}(s)) \\ & +n_1^{-2}k^{-2}\textstyle\sum_{i=1}^{n_1}\sum_{j=1,j\neq i}^{n_1}\sum_{r=1}^{k}\sum_{s=1}^{k} Cov(I_{\mathbf{X}_i}(r), I_{\mathbf{X}_j}(s)). \end{aligned}$$

Now, the first two terms on the right side are of the orders $O(n_1^{-1}k^{-1})$ and $O(n_1^{-1})$, respectively. So, both of them converge to zero as n_1 tends to infinity. Henze (1988) showed that (see Lemma 4.2 in p. 779) $\lim_{n\to\infty} E(I_{\mathbf{X}_1}(r) \mid \mathbf{x}_1 = \mathbf{x}) = \lambda f(\mathbf{x})/[\lambda f(\mathbf{x}) + (1-\lambda)g(\mathbf{x})]$. For any two distinct points $\mathbf{x}$ and $\mathbf{x}^{'}$, Henze (1984) (see p. 270) also showed that $\lim_{n\to\infty} E(I_{\mathbf{X}_1}(r)I_{\mathbf{X}_2}(s) \mid \mathbf{x}_1 = \mathbf{x}, \mathbf{x}_2 = \mathbf{x}^{'}) = \lambda^2 f(\mathbf{x})f(\mathbf{x}^{'})/\Big\{[\lambda f(\mathbf{x}) + (1-\lambda)g(\mathbf{x})]$ $[\lambda f(\mathbf{x}^{'}) + (1-\lambda)g(\mathbf{x}^{'})]\Big\}$ (though Henze (1984, 1988) formally proved these results for $k = 1$ and $r = s = 1$, the results for other choices of r, s and k follow from the arguments given in these articles). These two results and a simple application of the dominated convergence theorem imply that for all r, s, and $i \neq j$, $Cov(I_{\mathbf{X}_i}(r), I_{\mathbf{X}_j}(s)) \to 0$ as $n \to \infty$. Therefore, the third term and hence $Var(T_{1,k})$ also converge to 0 as n increases. Similarly, one can show that $Var(T_{2,k})$ also converges to 0 as n increases. So, as $n \to \infty$, $|T_{i,k} - E(T_{i,k})| \overset{P}{\to} 0$ for $i = 1, 2$. So, under H_0, both $T_{NN1,k}$ and $T_{NN2,k}$ converge to 0 in probability. Hence, the critical values, i.e., the $100(1-\alpha)$-th percentiles of the null distributions of $T_{NN1,k}$ and $T_{NN2,k}$, also converge to zero. Therefore, for the large sample consistency of the proposed tests, it is enough to show that under H_A both $T_{NN1,k}$ and $T_{NN2,k}$ converge to positive constants.

Note that $E_{H_0}(T_{1,k}) = (n_1 - 1)/(n-1) \to \lambda$ and $E_{H_0}(T_{2,k}) = (n_2 - 1)/(n-1) \to 1-\lambda$ as $n \to \infty$, while it follows from the dominated convergence theorem that $E_{H_A}(T_{1,k}) \to \displaystyle\int \frac{\lambda f(\mathbf{z})}{\lambda f(\mathbf{z}) + (1-\lambda)g(\mathbf{z})} f(\mathbf{z})d\mathbf{z}$ and $E_{H_A}(T_{2,k}) \to \displaystyle\int \frac{(1-\lambda)g(\mathbf{z})}{\lambda f(\mathbf{z}) + (1-\lambda)g(\mathbf{z})} g(\mathbf{z})d\mathbf{z}$ as $n \to \infty$. Therefore, from the continuous mapping theorem, we have

$$T_{NN1,k} \overset{P}{\to} \lambda \left|\int \frac{\lambda(1-\lambda)[f(\mathbf{z}) - g(\mathbf{z})]}{\lambda f(\mathbf{z}) + (1-\lambda)g(\mathbf{z})} f(\mathbf{z})d\mathbf{z}\right| + (1-\lambda)\left|\int \frac{\lambda(1-\lambda)[g(\mathbf{z}) - f(\mathbf{z})]}{\lambda f(\mathbf{z}) + (1-\lambda)g(\mathbf{z})} g(\mathbf{z})d\mathbf{z}\right|.$$

Similarly, we have

$$T_{NN2,k} \overset{P}{\to} \lambda \left[\int \frac{\lambda(1-\lambda)[f(\mathbf{z}) - g(\mathbf{z})]}{\lambda f(\mathbf{z}) + (1-\lambda)g(\mathbf{z})} f(\mathbf{z})d\mathbf{z}\right]^2 + (1-\lambda)\left[\int \frac{\lambda(1-\lambda)[g(\mathbf{z}) - f(\mathbf{z})]}{\lambda f(\mathbf{z}) + (1-\lambda)g(\mathbf{z})} g(\mathbf{z})d\mathbf{z}\right]^2.$$

These limiting values of $T_{NN1,k}$ and $T_{NN2,k}$ are 0 only if $\displaystyle\int \frac{f(\mathbf{z}) - g(\mathbf{z})}{\lambda f(\mathbf{z}) + (1-\lambda)g(\mathbf{z})} f(\mathbf{z})d\mathbf{z} = \int \frac{g(\mathbf{z}) - f(\mathbf{z})}{\lambda f(\mathbf{z}) + (1-\lambda)g(\mathbf{z})} g(\mathbf{z})d\mathbf{z} = 0 \Rightarrow \int \frac{[f(\mathbf{z}) - g(\mathbf{z})]^2}{\lambda f(\mathbf{z}) + (1-\lambda)g(\mathbf{z})} d\mathbf{z} = 0 \Rightarrow f = g$ almost everywhere. □

4.6.1 Upper bound of $w_1 Var(T_{1,k}) + w_2 Var(T_{2,k})$ under permutation

Let Ω_1 be the collection of all indices i such that $\mathbf{z}_i$ is labeled as an observation from F under permutation and $\Omega_2 = \{1, 2, \ldots, n\} - \Omega_1$. For $1 \leq i \neq j \leq n$, now define

$$a_{ij} = \begin{cases} 1 & \text{if } \mathbf{z}_j \text{ is one of the } k \text{ neighbors of } \mathbf{z}_i \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad b_{ij} = \begin{cases} 1 & \text{if } \ i, j \in \Omega_1 \\ 0 & \text{otherwise.} \end{cases}$$

So, for a given set of observations $\mathbf{z}_1, \ldots, \mathbf{z}_n$, the a_{ij}s are fixed, but the b_{ij}s are random. Under the permutation distribution, $E(b_{ij}) = P(b_{ij} = 1) = \varphi_1 = n_1(n_1 - 1)/n(n-1)$. Similarly for $i \neq j \neq p \neq q$, we have $E(b_{ij}b_{jp}) = (n_1 - 2)\varphi_1/(n-2)$ and $E(b_{ij}b_{pq}) = (n_1 - 2)(n_1 - 3)\varphi_1/(n-2)(n-3)$. If we take $a_{ii} = b_{ii} = 0$ for $i = 1, 2, \ldots, n$, we have $n_1 k T_{1,k} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} b_{ij}$. Therefore,

$$\begin{aligned} Var[n_1 k T_{1,k}] &= \textstyle\sum_{i=1}^n \sum_{j=1}^n (a_{ij} + a_{ij}a_{ji}) Var(b_{ij}) \\ &+ \textstyle\sum_{i=1}^n \sum_{j=1}^n \sum_{p=1, p\neq i}^n a_{ij} a_{pj} Cov(b_{ij}, b_{pj}) \\ &+ \textstyle\sum_{i=1}^n \sum_{j=1, j\neq i}^n \sum_{p=1}^n a_{pi} a_{pj} Cov(b_{pi}, b_{pj}) \\ &+ \textstyle\sum_{i=1}^n \sum_{j=1, j\neq i}^n \sum_{p=1}^n a_{ip} a_{pj} Cov(b_{ip}, b_{pj}) \\ &+ \textstyle\sum_{i=1}^n \sum_{j=1, j\neq i}^n \sum_{p=1}^n a_{pi} a_{jp} Cov(b_{pi}, b_{jp}) \\ &+ \textstyle\sum_{i=1}^n \sum_{j=1}^n \sum_{p=1, p\neq i,j}^n \sum_{q=1, q\neq i,j}^n a_{ij} a_{pq} Cov(b_{ij}, b_{pq}). \end{aligned}$$

For $j = 1, \ldots, n$, define $\delta_j = \sum_{i=1}^n a_{ij}$, the number of observations whose one of the k neighbors is $\mathbf{z}_j$. Now, note that $\sum_{i=1}^n \sum_{j=1}^n a_{ij} = nk$, $\sum_{i=1}^n \sum_{j=1}^n a_{ij} a_{ji} \leq nk$, $\sum_{i=1}^n \sum_{j=1}^n \sum_{p=1, p\neq i} a_{ij} a_{pj} = \sum_{j=1}^n \delta_j(\delta_j - 1)$, $\sum_{i=1}^n \sum_{j=1, j\neq i}^n \sum_{p=1}^n a_{pi} a_{pj} = nk(k-1)$, $\sum_{i=1}^n \sum_{j=1, j\neq i}^n \sum_{p=1}^n a_{ip} a_{pj} = \sum_{i=1}^n \sum_{j=1, j\neq i}^n \sum_{p=1}^n a_{pi} a_{jp} \leq k \sum_{i=1}^n \sum_{p=1}^n a_{ip} = nk^2$ and $\sum_{i=1}^n \sum_{j=1}^n \sum_{p=1, p\neq i,j}^n \sum_{q=1, q\neq i,j}^n a_{ij} a_{pq} \geq nk(n-2)(k-2)$. Also, observe that $Cov(b_{ij}, b_{pq}) < 0$ while the other covariances are positive. Therefore,

$$\begin{aligned} Var[n_1 k T_{1,k}] &\leq 2nk(\varphi_1 - \varphi_1^2) + \left[\textstyle\sum_{j=1}^n \delta_j^2 - 2nk + 3nk^2\right] \left[\tfrac{(n_1-2)\varphi_1}{n-2} - \varphi_1^2\right] \\ &+ \left[\tfrac{(n_1-2)(n_1-3)\varphi_1}{(n-2)(n-3)} - \varphi_1^2\right] [nk(n-2)(k-2)]. \end{aligned}$$

Replacing n_1 by n_2 and φ_1 by $\varphi_2 = n_2(n_2 - 1)/n(n-1)$, we get an upper bound for $Var(T_{2,k})$. Combining both, we get $w_1 Var(T_{1,k} + w_2 Var(T_{2,k}) \leq \psi_1(n_1, n_2, k) \sum_{j=1}^n \delta_j^2 + \psi_2(n_1, n_2, k)$, where $\psi_1(n_1, n_2, k) = \frac{1}{nk^2} \sum_{i=1}^2 \frac{1}{n_i} \left[\frac{(n_1-2)\varphi_1}{n-2} - \varphi_1^2\right]$ and $\psi_2(n_1, n_2, k) =$

$\sum_{i=1}^{2}\left\{\frac{2}{n_ik}(\varphi_i-\varphi_i^2)+\frac{3k-2}{n_ik}\left[\frac{(n_i-2)\varphi_i}{n-2}-\varphi_i^2\right]+\left[\frac{(n_i-2)(n_i-3)\varphi_i}{(n-2)(n-3)}-\varphi_i^2\right]\frac{(n-2)(k-2)}{n_ik}\right\}$. If we assume that n_1 and n_2 are of the same order as n, it can be shown that $\psi_1(n_1,n_2,k)=O(1/(n^2k^2))$ and $\psi_2(n_1,n_2,k)=O(1/n)$.

4.6.2 Estimation of Σ_w

For $r=1,2,\ldots,k$ and $i=1,2,\ldots,n$, let the r-th nearest neighbor of $\mathbf{z}_i$ be denoted by $NN_i(r)$. Now, for $i\neq j$, let us consider the following five mutually exclusive and exhaustive probabilities:-

$p_1(r,s)=P_{H_0}(NN_i(r)=\mathbf{z}_j,NN_j(s)=\mathbf{z}_i)$, $p_2(r,s)=P_{H_0}(NN_i(r)=NN_j(s))$, $p_3(r,s)=P_{H_0}(NN_i(r)=\mathbf{z}_j,NN_j(s)\neq\mathbf{z}_i)$, $p_4(r,s)=P_{H_0}(NN_i(r)\neq\mathbf{z}_j,NN_j(s)=\mathbf{z}_i)$ and $p_5(r,s)=P_{H_0}(NN_i(r)\neq\mathbf{z}_j,NN_j(s)\neq\mathbf{z}_i,NN_i(r)\neq NN_j(s))$.

Define $I_i(r)$ as the indicator variable that takes the value 1 if $\mathbf{z}_i$ and its r-th nearest neighbor belong to the same sample. Now,

$$\begin{aligned}Var_{H_0}(n_1kT_{1,k})&=Var_{H_0}(\textstyle\sum_{i=1}^{n_1}\sum_{r=1}^{k}I_i(r))\\&=E_{H_0}[(\textstyle\sum_{i=1}^{n_1}\sum_{r=1}^{k}I_i(r))^2]-[E_{H_0}(\sum_{i=1}^{n_1}\sum_{r=1}^{k}I_i(r))]^2\\&=\textstyle\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{r=1}^{k}\sum_{s=1}^{k}P_{H_0}(I_i(r)=I_j(s)=1)-(\frac{kn_1(n_1-1)}{(n-1)})^2\\&=\textstyle\sum_{i=1}^{n_1}\sum_{r=1}^{k}P_{H_0}(I_i(r)=1)+\sum_{i=1}^{n_1}\sum_{r=1}^{k}\sum_{s=1,s\neq r}^{k}P_{H_0}(I_i(r)=I_i(s)=1)\\&\quad+\textstyle\sum_{i=1}^{n_1}\sum_{j=1,j\neq i}^{n_1}\sum_{r=1}^{k}\sum_{s=1}^{k}P_{H_0}(I_i(r)=I_j(s)=1)-(\frac{kn_1(n_1-1)}{(n-1)})^2.\end{aligned}$$

Note that $\sum_{i=1}^{n_1}\sum_{r=1}^{k}P_{H_0}(I_i(r)=1)=kn_1(n_1-1)/(n-1)$,
$\sum_{i=1}^{n_1}\sum_{r=1}^{k}\sum_{s=1,s\neq r}^{k}P_{H_0}(I_i(r)=I_i(s)=1)=2n_1\binom{k}{2}(n_1-1)(n_1-2)/(n-1)(n-2)$ and $\sum_{i=1}^{n_1}\sum_{j=1,j\neq i}^{n_1}\sum_{r=1}^{k}\sum_{s=1}^{k}P_{H_0}(I_i(r)=I_j(s)=1)=n_1(n_1-1)\sum_{r=1}^{k}\sum_{s=1}^{k}p_0(r,s)$, where $p_0(r,s)=p_1(r,s)+\frac{n_1-2}{n-2}[p_2(r,s)+p_3(r,s)+p_4(r,s)]+\frac{(n_1-2)(n_1-3)}{(n-2)(n-3)}p_5(r,s)$.

We obtain $Var_{H_0}[n_2kT_{2,k}]$ by replacing n_1 by n_2 in the above expression. Also,

$$\begin{aligned}Cov_{H_0}(n_1kT_{1,k},n_2kT_{2,k})&=Cov_{H_0}(\textstyle\sum_{i=1}^{n_1}\sum_{r=1}^{k}I_i(r),\sum_{j=1}^{n_2}\sum_{s=1}^{k}I_j(s))\\&=\textstyle\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{r=1}^{k}\sum_{s=1}^{k}P_{H_0}(I_i(r)=I_j(s)=1)-(E_{H_0}(n_1kT_1))^2\\&=n_1n_2\textstyle\frac{(n_1-1)(n_2-1)}{(n-2)(n-3)}\sum_{r=1}^{k}\sum_{s=1}^{k}p_5(r,s)-(k\frac{n_1(n_1-1)}{(n-1)})(k\frac{n_2(n_2-1)}{(n-1)}).\end{aligned}$$

Now, using $p_1(r,s)=(n-1)^{-1}P_{H_0}(NN_j(s)=\mathbf{z}_i\mid NN_i(r)=\mathbf{z}_j)$, we easily obtain $p_3(r,s)=p_4(r,s)=\frac{1}{(n-1)}-p_1(r,s)$ and $p_5(r,s)=\frac{(n-3)}{(n-1)}+p_1(r,s)-p_2(r,s)$.

So, $p_i(r,s), i = 1(1)5$ can be represented in terms of $p_1(r,s)$and $p_2(r,s)$ only (see also Schilling (1986a)). Therefore, the estimation problem now reduces to estimating the limiting values (as $n \to \infty$) of $np_1(r,s)$ and $np_2(r,s)$.

For any $x_0 \in \mathbb{R}^d$ and $\rho{>}0$, define $SP(x_0, r) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq r\}$, where $\|\cdot\|$ represent the Euclidean norm. Define $SP_1 = SP(X_1, \|X_2 - X_1\|)$ and $SP_2 = SP(X_2,\ \|X_2 - X_1\|)$. Schilling(1986b) showed that (see Theorem 2.1 in p. 391) the limiting value of $np_1(r,s)$ is given by

$$np_1(r,s) \sim (1 - C_d) \sum_{l=0}^{\min(r',s')} \binom{r' + s' - l}{l, r' - l, s' - l} (1 - 2C_d)^l C_d^{r'+s'-2l},$$

where, $r' = r - 1$, $s' = s - 1$, and C_d is the proportion of volume of $SP_1 \cup SP_2$ that belongs to SP_1 only. It was also shown in Schilling (1986b) (see p. 394) that

$$np_2(r,s) \sim K_d \sum_{i,j=0}^{1} \sum_{l=0}^{\bar{l}} \tbinom{l+\epsilon_1+\epsilon_2+1}{l,\epsilon_1,\epsilon_2,1} \int\limits_{E^*_{i,j}} Vol^l\{S_1^* \cap S_2^*\} Vol^{\epsilon_1}\{S_1^* - S_2^*\}$$

$$Vol^{\epsilon_2}\{S_2^* - S_1^*\} Vol^{-(l+\epsilon_1+\epsilon_2+2)}\{S_1^* \cup S_2^*\} d\mathbf{u},$$

where K_d= Volume of d dimensional unit sphere, $\bar{l} = min(r + i - 2, s + j - 2)$, $\epsilon_1 = r - l + i - 2$, $\epsilon_2 = s - l + j - 2$, $Vol(E)$ = Volume of E, $S_1^* = SP(\mathbf{0}, \|\mathbf{u}\|)$ is the sphere around $\mathbf{0}$, $S_2^* = SP(\mathbf{e}, \|\mathbf{u} - \mathbf{e}\|)$ is a sphere around a unit vector $\mathbf{e}$, $E^*_{i,j} = (S_1^*)^i \cap (S_2^*)^j$, $(S_1^*)^0 = (S_1^*)^c$ and $(S_2^*)^0 = (S_2^*)^c$. These two quantities can be estimated by Monte Carlo simulation.

Chapter 5

A test based on averages of inter-point distances

Inter-point distances play an important role for constructing nonparametric methods for the multivariate two-sample problem. For $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d.}{\sim} F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{i.i.d.}{\sim} G$, let D_{FF}, D_{GG} and D_{FG} denote the distributions of the inter-point distances $\|\mathbf{X}_1 - \mathbf{X}_2\|$, $\|\mathbf{Y}_1 - \mathbf{Y}_2\|$ and $\|\mathbf{X}_1 - \mathbf{Y}_2\|$, respectively. Under mild conditions, Maa et al. (1996) proved that D_{FF}, D_{GG} and D_{FG} are identical if and only if $F = G$. Therefore, under the alternative $H_A : F \neq G$, the differences in the distributions of these inter-point distances contain useful information about the separation between F and G. Note that irrespective of the dimension of the data, D_{FF}, D_{GG} and D_{FG} are one-dimensional distributions, and we try to extract separation information contained in these univariate distributions to get evidence against H_0. Several nonparametric two-sample tests based on inter-point distances have been proposed in the literature. The MST run test (Friedman and Rafsky (1979)), the NN test (Schilling (1986a); Henze (1988)), the Cramer test (Baringhaus and Franz (2004)), the Adjacency test (Rosenbaum (2005)) and the HT test (Hall and Tajvidi (2002)) are all based on inter-point distances. We also used inter-point distances to construct our tests in Chapters 3 and 4. Since these tests are based on inter-point distances, they are invariant under location change, rotation and homogeneous scale transformation of the data, and they can be conveniently used for HDLSS data or even for functional data taking values in an infinite dimensional Banach space. Recall that in Chapter 3, we constructed a run test based on SHP that overcomes the limitations of the MST run test in high dimensions. In Chapter 4, we proposed a test based on nearest neighbors which has better consistency properties in high dimensions compared to the NN test of Schilling (1986a) and Henze (1988). In this chapter, we propose a test based on averages of three types of inter-point distances, which can be viewed as a modification over the Cramer test (Baringhaus and Franz (2004)). The description of this test is given in the following section.

5.1 Description of the proposed test

If $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d.}{\sim} F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{i.i.d.}{\sim} G$, from Maa et al. (1996), we know that $F = G \Leftrightarrow D_{FF} = D_{GG} = D_{FG}$. Now, consider two bivariate distributions. Let the distribution of $(\|\mathbf{X}_1 - \mathbf{X}_2\|, \|\mathbf{X}_1 - \mathbf{Y}_1\|)^T$ be denoted by D_F, and that of $(\|\mathbf{Y}_1 - \mathbf{X}_1\|, \|\mathbf{Y}_1 - \mathbf{Y}_2\|)^T$ be denoted by D_G. Clearly, D_F has marginals D_{FF} and D_{FG}, while D_G has marginals D_{FG} and D_{GG}. One can check that when F and G differ, D_F and D_G differ as well, and vice versa. Further, if D_F and D_G have finite means $\boldsymbol{\mu}_{D_F}$ and $\boldsymbol{\mu}_{D_G}$, we also have the following result.

LEMMA 5.1: *Suppose that* $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d}{\sim} F$ *and* $\mathbf{Y}_1, \mathbf{Y}_2 \overset{i.i.d}{\sim} G$. *Also assume that* $\mu_{FF} = E(\|\mathbf{X}_1 - \mathbf{X}_2\|)$, $\mu_{GG} = E(\|\mathbf{Y}_1 - \mathbf{Y}_2\|)$ *and* $\mu_{FG} = E(\|\mathbf{X}_1 - \mathbf{Y}_1\|)$ *exist. Then,* $\boldsymbol{\mu}_{D_F} = (\mu_{FF},\ \mu_{FG})^T$ *and* $\boldsymbol{\mu}_{D_G} = (\mu_{FG},\ \mu_{GG})^T$ *are equal if and only if* $F = G$.

The proof of the lemma is given in Section 5.6. From this lemma, it is clear that instead of testing $H_0 : F = G$ against $H_A : F \neq G$, one can equivalently test the null hypothesis $H_0^{''} : \boldsymbol{\mu}_{D_F} = \boldsymbol{\mu}_{D_G}$ against the alternative $H_A^{''} : \boldsymbol{\mu}_{D_F} \neq \boldsymbol{\mu}_{D_G}$. If we have n_1 independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ from F and n_2 independent observations $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ from G, we calculate the estimates of μ_{FF}, μ_{FG} and μ_{GG} given by $\hat{\mu}_{FF} = \binom{n_1}{2}^{-1} \sum_{i=1}^{n_1} \sum_{j=i+1}^{n_1} \|\mathbf{x}_i - \mathbf{x}_j\|$, $\hat{\mu}_{FG} = (n_1 n_2)^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|\mathbf{x}_i - \mathbf{y}_j\|$ and $\hat{\mu}_{GG} = \binom{n_2}{2}^{-1} \sum_{i=1}^{n_2} \sum_{j=i+1}^{n_2} \|\mathbf{y}_i - \mathbf{y}_j\|$, respectively. So, $\boldsymbol{\mu}_{D_F}$ and $\boldsymbol{\mu}_{D_G}$ are estimated by $\widehat{\boldsymbol{\mu}}_{D_F} = (\hat{\mu}_{FF}, \hat{\mu}_{FG})^T$ and $\widehat{\boldsymbol{\mu}}_{D_G} = (\hat{\mu}_{FG}, \hat{\mu}_{GG})^T$, respectively. While $\widehat{\boldsymbol{\mu}}_{D_F}$ and $\widehat{\boldsymbol{\mu}}_{D_G}$ are expected to be close under H_0, the distance between $\widehat{\boldsymbol{\mu}}_{D_F}$ and $\widehat{\boldsymbol{\mu}}_{D_G}$ is expected to be large under the alternative. So, we reject H_0 for higher values of the test statistic $T_{n_1,n_2} = \|\widehat{\boldsymbol{\mu}}_{D_F} - \widehat{\boldsymbol{\mu}}_{D_G}\|^2 = (\hat{\mu}_{FF} - \hat{\mu}_{FG})^2 + (\hat{\mu}_{FG} - \hat{\mu}_{GG})^2$. When n_1 and n_2 are small, we use the permutation method to calculate the cut-off. When they are large, this cut-off is chosen using the large sample distribution of T_{n_1,n_2} given in Section 5.5.

To investigate the performance of this proposed test, we considered three examples involving normal distributions discussed in Chapters 3 and 4, where the components of F were assumed to be i.i.d. $N(0,1)$ while those of G were assumed to be $N(\mu,\sigma^2)$. We considered three choices of for (μ,σ), (i) $\mu = 0.3$, $\sigma = 1$ (ii) $\mu = 0$, $\sigma = 1.3$ and (iii) $\mu = 0.2$, $\sigma = 1.2$. Note that the first two examples were used in Chapter 3, whereas the third one was used in Chapter 4. In all these cases, we used $n_1 = n_2 = 20$ and different values of d ranging between 2 and 500. Each experiment was repeated 500 times as before. Figure 5.1 shows the powers of the proposed test in these three examples along with those of the MST run test, the NN test and the Cramer test.

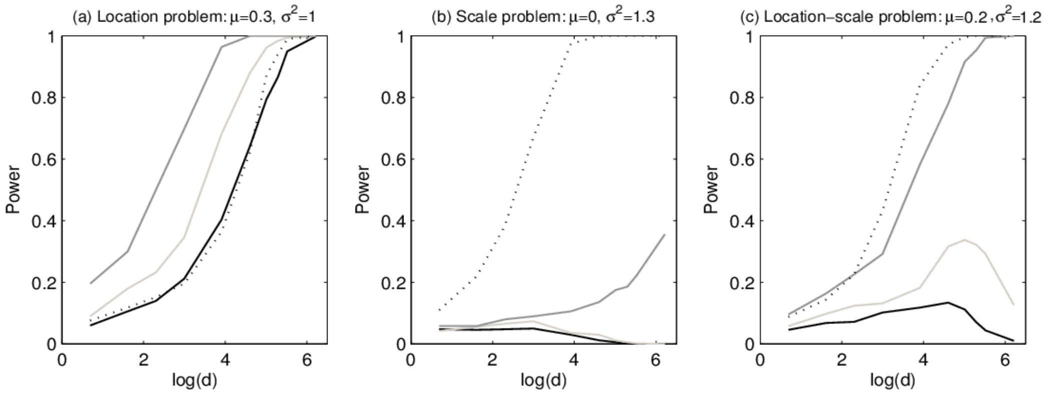


Figure 5.1: Powers of different two-sample tests for varying choices of d. (MST run test (black solid), NN test (light grey), Cramer test (dark grey) and proposed test (black dotted))

We have seen that in each of these examples, the separation between F and G increases with d. So, one should expect the powers of these tests to tend to unity as d increases. We observed that in the case of location problem (see Figure 5.1(a)), but not in other two cases. In the location-scale problem, although the power of the Cramer test increased with d, those of the MST run and the NN tests dropped down to zero as d increased (see Figure 5.1(c)). In the case of scale problem, all of these three methods yielded poor performance (see Figure 5.1(b)). But in all these three cases, the power of our proposed test converged to 1 as the dimension increased. In the scale problem and the location-scale problem, it outperformed all the three competing tests considered here. Only in the case of location problem, the Cramer test had the best performance. The reasons behind such performance of MST run test, the NN test have already been discussed in previous chapters. In this chapter, we investigate the behavior of the Cramer test (see Baringhaus and Franz (2004)) and the proposed test for high dimensional data.

5.2 Behavior of the proposed test in high dimensions

In this section, we investigate the behavior of the proposed test when n_1 and n_2 are fixed, and the dimension d diverges to infinity. For this investigation, we assume the regularity conditions (A1)-(A3) mentioned in Chapter 2. The following theorem shows the behavior of the power function of the proposed test under these regularity conditions.

THEOREM 5.1: *Suppose that we have n_0 independent observations from each of F and G (i.e., $n_1 = n_2 = n_0$), which satisfy (A1)-(A3). Also assume that either $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$. Then, unless n_0 is very small (i.e., $\binom{2n_0}{n_0} \leq 2/\alpha$), the power of the proposed test of level α converges to* 1 *as d tends to infinity.*

In the proof of Theorem 5.1 (see Section 5.6), one can see that for all choices of n_1 and n_2, as $d \to \infty$, under the assumptions (A1)-(A3), we have $T_{n_1,n_2}/d \xrightarrow{P} \upsilon_0^*$, where $\upsilon_0^* = (\sigma_1\sqrt{2} - \sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2})^2 + (\sigma_2\sqrt{2} - \sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2})^2$. Also, it is clear from the proof that for $n_1 = n_2 = n_0$, the limiting p-value (as $d \to \infty$) of the permutation test, i.e., the limiting value of $P(T_{n_1,n_2}/d \geq \upsilon_0^*)$ under the permutation distribution is $2/\binom{2n_0}{n_0}$. So, for a test of level 0.05 (respectively, 0.01), it is enough to have four (respectively, five) observations from each class for the convergence of the power to unity. The case $n_1 \neq n_2$ calls for more complicated calculations, but for $n_1 \geq 4$ and $n_2 \geq 4$ (or $n_1 \geq 5$ and $n_2 \geq 5$), it can be viewed as the case $n_1 = n_2 = 4$ (or $n_1 = n_2 = 5$) with some additional information on at least one of the distributions. So, the resulting test is expected to have more power, and one can expect it to have the large dimensional consistency for all such choices of n_1 and n_2. Figure 5.2, which shows the limiting p-values for different choices of $n_1 \geq 4$ and $n_2 \geq 4$, justifies this claim. In this figure, one can notice that in the three examples discussed in Section 5.1, the limiting p-values were almost the same, and in each example, for any fixed choice of n_1 (respectively,

n_2), the limiting p-value was non-increasing in n_2 (respectively, n_1). We carried out our investigation for various other choices of the parameters ν, σ_1^2 and σ_2^2, but this basic pattern remained the same.

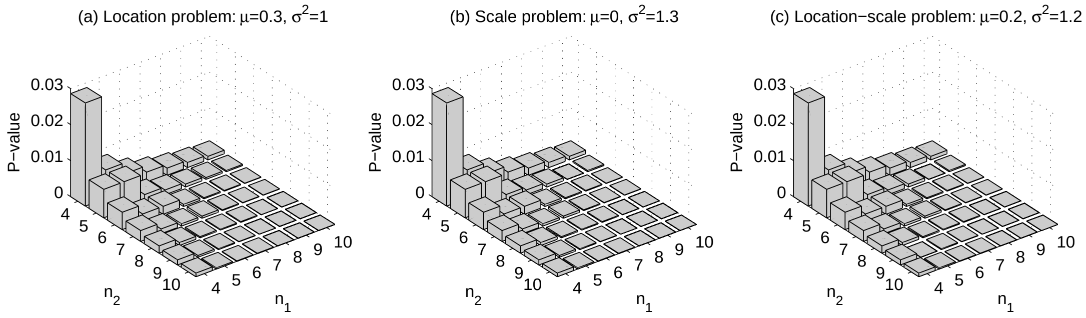


Figure 5.2: Limiting p-values for different choices of n_1 and n_2.

Note that the Cramer test rejects H_0 for large values of the test statistic $T^{CR}_{n_1,n_2} = 2\hat{\mu}_{FG} - \hat{\mu}_{FF} - \hat{\mu}_{GG}$. It is easy to see that when n_1 and n_2 are fixed and $d \to \infty$, under the assumptions (A1)-(A3), $\hat{\mu}_{FF}/\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2}$, $\hat{\mu}_{GG}/\sqrt{d} \xrightarrow{P} \sigma_2\sqrt{2}$, $\hat{\mu}_{FG}/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2+\sigma_2^2+\nu^2}$, and hence the scaled version of the test statistic, $T^{CR}_{n_1,n_2}/\sqrt{d}$, converges to $2\sqrt{\sigma_1^2+\sigma_2^2+\nu^2} - \sigma_1\sqrt{2} - \sigma_2\sqrt{2}$ $(= \upsilon^*, \text{say})$ in probability. Now, υ^* is positive unless $\sigma_1^2 = \sigma_2^2$ and $\nu^2 = 0$, and a consistency result similar to Theorem 5.1 can be proved for the Cramer test as well. But, in Section 5.1, we have seen that in the location-scale problem and the scale problem, especially in the latter case, it did not perform well. Note that in such cases, we had $\nu^2 < |\sigma_1^2 - \sigma_2^2|$. Now, $\nu^2 < |\sigma_1^2 - \sigma_2^2|$ implies that $\sqrt{\sigma_1^2+\sigma_2^2+\nu^2}$ lies between $\sigma_1\sqrt{2}$ and $\sigma_2\sqrt{2}$. So, even when both $(\hat{\mu}_{FG} - \hat{\mu}_{FF})$ and $(\hat{\mu}_{FG} - \hat{\mu}_{GG})$ are significantly different from zero, they are likely to be of different sign. As a result, when they are added up, $T^{CR}_{n_1,n_2} = (\hat{\mu}_{FG} - \hat{\mu}_{FF}) + (\hat{\mu}_{FG} - \hat{\mu}_{GG})$ may take a value close to zero, and consequently, $H_0 : F = G$ may get accepted. We observed this phenomenon several times in the location-scale problem and the scale problem in Section 5.1. In the case of scale problem, υ^* was also close to zero. So, even for $d = 500$, the Cramer test did not have satisfactory power. But, if we take the sum of $(\hat{\mu}_{FG} - \hat{\mu}_{FF})^2$ and $(\hat{\mu}_{FG} - \hat{\mu}_{GG})^2$, such cancellations are not possible, and H_0 is more likely to be rejected. That is why our test based on $T_{n_1,n_2} = (\hat{\mu}_{FG} - \hat{\mu}_{FF})^2 + (\hat{\mu}_{FG} - \hat{\mu}_{GG})^2$ had better performance in these two examples. Note that T_{n_1,n_2} can also be expressed as $T_{n_1,n_2} =$

$\frac{1}{2}\left[(2\hat{\mu}_{FG} - \hat{\mu}_{FF} - \hat{\mu}_{GG})^2 \ + \ (\hat{\mu}_{FF} - \hat{\mu}_{GG})^2\right]$, where the first part $(2\hat{\mu}_{FG} - \hat{\mu}_{FF} - \hat{\mu}_{GG})^2$ is the square of the Cramer test statistic. We have seen that the Cramer test works well when F and G differ in location, but it is not very sensitive against small changes in scale. The second part $(\hat{\mu}_{FF} - \hat{\mu}_{GG})^2$ compensates for that and makes the test sensitive against scale alternatives. However, in the case of location problem, the term $(\hat{\mu}_{FF} - \hat{\mu}_{GG})^2$ serves as noise. Therefore, in such cases, the proposed test is unlikely to outperform the Cramer test, and that is what we observed in our experiment.

Baringhaus and Franz (2010) proposed a class of rigid motion invariant two sample tests that includes the Cramer test. They considered a continuous function $\phi : [0, \infty) \rightarrow [0, \infty)$ and defined $\mu^{\phi}_{FF} = E\phi(\|\mathbf{X}_1 - \mathbf{X}_2\|^2)$, $\mu^{\phi}_{GG} = E\phi(\|\mathbf{Y}_1 - \mathbf{Y}_2\|^2)$ and $\mu^{\phi}_{FG} = E\phi(\|\mathbf{X}_1 - \mathbf{Y}_1\|^2)$. They proved that if ϕ is non decreasing and it satisfies some appropriate regularity conditions, the inequality $2\mu^{\phi}_{FG} - \mu^{\phi}_{FF} - \mu^{\phi}_{GG} \geq 0$ is satisfied, where the equality holds if and only if $F = G$. So, replacing μ^{ϕ}_{FF}, μ^{ϕ}_{GG} and μ^{ϕ}_{FG} by their empirical analogs, a class of test statistics $2\hat{\mu}^{\phi}_{FG} - \hat{\mu}^{\phi}_{FF} - \hat{\mu}^{\phi}_{GG}$ was constructed. Note that when H_0 fails to hold, due to monotonicity of ϕ, depending on the ordering of the three types of distances, here also $\hat{\mu}^{\phi}_{FG}$ can lie between $\hat{\mu}^{\phi}_{FF}$ and $\hat{\mu}^{\phi}_{GG}$. In such cases, due to cancellation of positive and negative terms, the test statistic may take small values leading to the acceptance of H_0. But such cancellations are not possible if we use $(\hat{\mu}^{\phi}_{FG} - \hat{\mu}^{\phi}_{FF})^2 + (\hat{\mu}^{\phi}_{FG} - \hat{\mu}^{\phi}_{GG})^2$ as the test statistic. As a consequence, the resulting test can have better power properties in such situations.

5.3 Results from the analysis of simulated data sets

We carried out further simulation studies to evaluate the performance of our proposed test in high dimensional data. For this study, we used some examples involving 500 dimensional normal and Laplace distributions as well as some examples involving autoregressive processes. In all these cases, we generated 20 observations from each of the two distributions, F and G, to constitute the sample and used it to test $H_0 : F = G$ against $H_A : F \neq G$. Each experiment was carried out 200 times, and the estimated power of the proposed test is reported in Table 5.1. To facilitate comparison, powers of MST run (Friedman and Rafsky (1979)), NN (Schilling (1986a); Henze (1988)), Cramer

(Baringhaus and Franz (2004), HT (Hall and Tajvidi (2002)) and Adjacency (Rosenbaum (2005)) tests are also reported. Recall that the Adjacency test is distribution-free. For all other methods, we used the conditional tests based on 500 permutations.

We began with some examples involving normal distribution. Recall that in Section 5.1, we used some examples with multivariate normal distributions, where the component variables $X^{(1)}, \ldots, X^{(d)}$ (and $Y^{(1)}, \ldots, Y^{(d)}$) were independent and identically distributed. So, here we considered some examples, where both in F and G, the component variables were positively correlated. While F had the mean vector $(0, 0, \ldots, 0)^T$ and the dispersion matrix $\mathbf{\Sigma}$, those for G are taken to be $(\mu, \mu, \ldots, \mu)^T$ and $\sigma^2\mathbf{\Sigma}$, respectively, where $\mathbf{\Sigma}$ had the (i, j)-th entry $(0.5)^{|i-j|}$ for $i, j = 1, 2, \ldots, d$. Here also, we considered three different choices of μ and σ^2 $\big[(\mu, \sigma^2)$=$(0.25, 1), (0, 1.25)$ and $(0.1, 1.1)\big]$ to have three different types of problems. Once again, in cases of scale problem and location-scale problem, the proposed test yielded the highest power among all two-sample tests considered here. Only in the case of location problem, the Cramer test and the NN test performed better than the proposed test. However, the proposed test and the HT test had comparable performance in this example as well, and they yielded much higher powers than those of the Adjacency test and the MST run test.

We obtained similar results when we carried out our experiment with Laplace distributions, where the component variables in F and G were assumed to be independent and identically distributed. We considered three different types of problems (location, scale and location-scale) as before, and in each case, the component variables in F and G had the same means and variances as in the corresponding examples with normal distributions. Again, in the location problem, the Cramer test and the NN test had the best performance, but in other two cases, the proposed test and the HT outperformed their competitors. In the case of location-scale problem, the proposed test performed better than the HT test, while in other two cases, they had nearly the same power.

Next, we considered an example, where the component variables in F were i.i.d. standard normal variates, while those in G were i.i.d. standard Laplace variates. In this example, while the proposed test, the HT test and the Cramer test rejected H_0 in all of the 200 cases, the NN test and the MST run test could not reject it even on a single occasion. The Adjacency test had power 0.885.

Table 5.1: Observed powers (in %) of two-sample tests in simulated data sets

	Normal			Laplace			Normal vs.	AR(1)	AR(2)
	location	scale	loc-scale	location	scale	loc-scale	Laplace	process	process
Cramer	100.0	20.0	37.0	100.0	28.0	48.5	100.0	12.5	8.5
NN	94.5	0.0	11.5	100.0	0.5	15.0	0.0	6.0	6.0
MST run	77.0	0.0	7.5	91.0	0.0	6.0	0.0	3.0	4.0
HT	83.5	100.0	92.5	70.0	100.0	79.0	100.0	88.5	96.5
Adjacency	67.5	9.0	11.0	86.5	10.0	11.5	88.5	10.5	6.0
Proposed	82.0	100.0	94.0	69.5	100.0	85.5	100.0	91.0	99.0

Finally, we used two examples involving auto-regressive (AR) processes of order 1 (AR(1)) and order 2 (AR(2)), respectively. In the first example, we generated the observations in F using the AR(1) model $X^{(t)} = 0.25+0.3X^{(t-1)}+U_t$ for $t = 1, \ldots, 500$, where $X^{(0)}, U_1, U_2, \ldots, U_{500} \overset{i.i.d.}{\sim} N(0,1)$. Observations in G were generated using another AR(1) model $Y^{(t)} = 0.25+0.4Y^{(t-1)}+V_t$, where $Y^{(0)}$, V_1, $V_2, \ldots, V_{500} \overset{i.i.d.}{\sim} N(0,1)$. Note that in this example, F and G have difference both in locations and scales. In the second example, F and G differ only in their scales. In this example, the observations in F were generated using the AR(2) model $X^{(t)} = 0.3X^{(t-1)}+0.2X^{(t-2)}+U_t$ for $t = 1, 2, \ldots, 500$, and those in G were generated using the model $Y^{(t)} = 0.35Y^{(t-1)} + 0.25Y^{(t-2)} + V_t$ for $t = 1, 2, \ldots, 500$, where $X^{(0)}, X^{(-1)}, Y^{(0)}, Y^{(-1)}, U_1, U_2, \ldots, U_{500}, V_1, V_2, \ldots, V_{500}$ are all i.i.d. standard normal variates. In these two examples, the proposed test had excellent performance, and it outperformed its all competitors. While NN, MST run, Cramer and Adjacency tests failed to yield satisfactory results (see Table 5.1) in these two examples, the proposed test had powers 0.91 and 0.99, respectively.

5.4 Results from the analysis of benchmark data sets

We analyzed three benchmark data sets, namely, ECG data, Arcene data and Synthetic Control Chart data, for further evaluation of the proposed test. The first two data sets have already been introduced in previous chapters. The Control chart data set and its description are available at the UCI machine learning repository (http://www.ics.uci.edu/ml/datasets). Though this data set contains observations from six classes, for our analysis, we considered only two classes labeled as 'Cyclic' and 'Normal'. For each of these data sets, we repeated the experiment 500 times based on 500 different subsets chosen from the data at random, and the results for different subset sizes are reported in Table 5.2.

In the ECG data set, we considered subsets of three different sizes. In the case of $n_1 = 20, n_2 = 10$, i.e., when the subset sizes were proportional to the number of observations from that class in the pooled sample, the Cramer test, the NN test and the proposed test performed better than other three tests. In the case of equal subset size $n_1 = n_2 = 10$, the NN test had the best performance, but the performance of the proposed test and that of Cramer and HT tests were also comparable. The MST run test and the Adjacency test had relatively low power. In the case of $n_1 = 10$ and $n_2 = 20$, all methods except the Adjacency test rejected H_0 in more than 92% of the cases, while the NN test had the best performance.

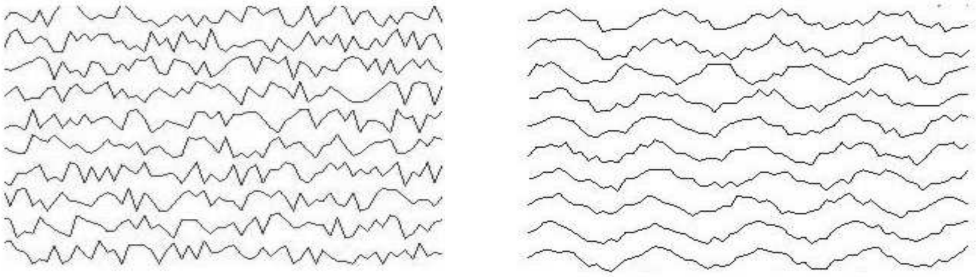


Figure 5.3: 'Normal' (on left) and 'Cyclic' (on right) classes in Control chart data.

The Control chart data is a synthetically generated time series data set, which contains 60-dimensional observations from each of 6 classes. However, we considered only two classes ('Normal' and 'Cyclic') for our experiment. The time series in the normal class are purely white noise, while those in cyclic class contain some cyclic pattern (see Figure 5.3). There are 100 observations from each class, but we used subsets of size 5 (i.e., $n_1 = n_2 = 5$). In this data set, the HT test and the proposed test rejected H_0 in all the 500 cases, while the Cramer test failed only once. The NN test had power 0.916, but the MST run test and the Adjacency test yielded poor performance.

In the case of Arcene data, we considered three choices of n_1 and n_2 (see Table 5.2). In all these cases, the NN test had the best performance closely followed by the MST run test. The proposed test also had reasonably high power, and it outperformed the Cramer test and the HT test on all these three occasions.

Table 5.2: Observed powers of two-sample tests (in %) in benchmark data sets.

Data sets	ECG			Control chart	Arcene		
(n_1,n_2)	(20,10)	(10,10)	(10,20)	(5,5)	(25,30)	(25,25)	(30,25)
Cramer	97.8	86.2	95.8	99.8	75.8	69.6	73.6
NN	98.0	90.2	98.8	91.6	99.4	98.8	99.2
MST run	88.0	75.2	92.2	29.8	97.2	94.4	96.6
HT	89.8	85.6	94.6	100.0	70.8	66.2	67.4
Adjacency	61.2	54.8	73.0	18.4	87.2	81.0	85.2
New	96.0	86.8	92.8	100.0	80.8	75.2	79.4

5.5 Large sample propoerties of the proposed test

So far, we have investigated the behavior of the proposed test in HDLSS situations. In this section, we study its large sample properties when the dimension of the data remains fixed. Here also, we use the test statistic T_{n_1,n_2} to test $H_0 : F = G$ against $H_A : F \neq G$ and reject H_0 for higher values of T_{n_1,n_2}. However, it is computationally expensive to use the permutation method when n_1 and n_2 are too large. So, here we construct the test based on the large sample distribution of T_{n_1,n_2}. This asymptotic distribution is given by the following theorem.

THEOREM 5.2: *Consider two sets of independent observations* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n_1}$ *and* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{n_2}$ *from* F*, which has finite second moments. Also assume that as* $n = (n_1 + n_2) \to \infty$*,* $n_1/n \to \lambda$ *for some* $\lambda \in (0, 1)$*. Then,* nT_{n_1,n_2} *is asymptotically distributed as* $\frac{2\varsigma^2}{\lambda(1-\lambda)}\chi_1^2$*, where* $\varsigma^2 = Var\{E(\|\mathbf{X}_1 - \mathbf{X}_2\| | \mathbf{X}_1)\}$*, and* χ_1^2 *denotes the chi-square distribution with* 1 *degree of freedom.*

To construct a test based on this large sample distribution, one needs to find consistent estimates for λ and ς^2. From the condition stated in the theorem, it is clear that $\hat{\lambda} = n_1/(n_1 + n_2)$ is consistent for λ. To find a consistent estimate for ς^2, first note that it can also be expressed as $\varsigma^2 = Cov(\|X_1 - X_2\|, \|X_1 - X_3\|) = E(\|X_1 - X_2\|\|X_1 - X_3\|) - E^2(\|X_1 - X_2\|)$. Now define

$$V_1^\circ = \left[\binom{n_1}{3}^{-1} \sum_{1 \leq i < j < l \leq n_1} \|\mathbf{x}_i - \mathbf{x}_j\|\|\mathbf{x}_i - \mathbf{x}_l\|\right] - \left[\binom{n_1}{2}^{-1} \sum_{1 \leq i < j \leq n_1} \|\mathbf{x}_i - \mathbf{x}_j\|\right]^2$$

$$\text{and } V_2^\circ = \left[\binom{n_2}{3}^{-1} \sum_{1 \leq i < j < l \leq n_2} \|\mathbf{y}_i - \mathbf{y}_j\|\|\mathbf{y}_i - \mathbf{y}_l\|\right] - \left[\binom{n_2}{2}^{-1} \sum_{1 \leq i < j \leq n_2} \|\mathbf{y}_i - \mathbf{y}_j\|\right]^2.$$

From the results on the probability convergence of U-Statistics (see e.g., Lee (1990)), one can check that V_1° and V_2° both are consistent for ς^2. Consequently, one can use $\hat{\varsigma}^2 = (n_1V_1^\circ + n_2V_2^\circ)/(n_1+n_2)$ as a consistent estimator ς^2 and show that under H_0, $T^*_{n_1,n_2} = (n_1+n_2)\hat{\lambda}(1-\hat{\lambda})T_{n_1,n_2}/2\hat{\varsigma}^2 \xrightarrow{d} \chi^2_1$. So, any test based on $T^*_{n_1,n_2}$ turns out to be asymptotically distribution-free. We compute $T^*_{n_1,n_2}$ from the data, and for a test of nominal level α, we reject H_0 if $T^*_{n_1,n_2}$ exceeds $\chi^2_{1,\alpha}$, where $P(\chi^2_1 > \chi^2_{1,\alpha}) = \alpha$. The following theorem shows the large sample consistency of the proposed test under the general alternative. Note that unlike the proposed test, the Cramer test and the HT test do not have the asymptotic distribution-free property. So, one has to use the bootstrap or the permutation method to find out the cut-off, which involves substantially higher computing cost.

THEOREM 5.3 : *Suppose that F and G both have finite second moments, and as $n_1, n_2 \to \infty$, $n_1/(n_1+n_2) \to \lambda$ for some $\lambda \in (0,1)$. Then, the power of the proposed test based on $T^*_{n_1,n_2}$ converges to* 1 *as n_1 and n_2 both tend to infinity.*

Figure 5.4 shows the power curve of the proposed test based on the large sample distribution of $T^*_{n_1,n_2}$ in normal location and scale problems. To facilitate comparison, it also shows the power curves of the other five tests considered in Section 5.3. For MST run, NN and Adjacency tests, we used the tests based on large sample distributions of the test statistics (see e.g., Schilling (1986a); Henze and Penrose (1999); Rosenbaum (2005)). In the case of Cramer test, we used the codes for the large sample test based on bootstrap approximation available at the R package 'cramer'. Since the large sample distribution of the HT test statistic is not known, we used its conditional version based on the permutation principle. Here F was considered to be a normal with the mean $(0,0,\ldots,0)^T$ and the scatter matrix $\mathbf{I}_d$, while G differed from F either in location $(\mu,\mu,\ldots,\mu)^T$ or in scatter $\sigma^2\mathbf{I}_d$. We used $d=5$ and $n_1 = n_2 = 100$ and each experiment was carried out 200 times to estimate the powers of different tests. In the location problem, the Cramer test had the best performance, but in the case of scale problem, once again, the proposed test outperformed all of its competitors. We observed the same phenomenon when we carried out our experiments with Laplace distribution. Therefore, we do not report it here. Clearly, these results are consistent with what we observed in

Section 5.3. We also considered another example, where all the five component variables in F were i.i.d. standard normal variates, and those in G were i.i.d. standard Laplace variates. In this example, the proposed method had an excellent performance. While Cramer, NN, MST run, HT and Adjacency tests had powers 0.605, 0.380, 0.340, 0.635 and 0.265, respectively, it rejected H_0 in 99% of the cases.

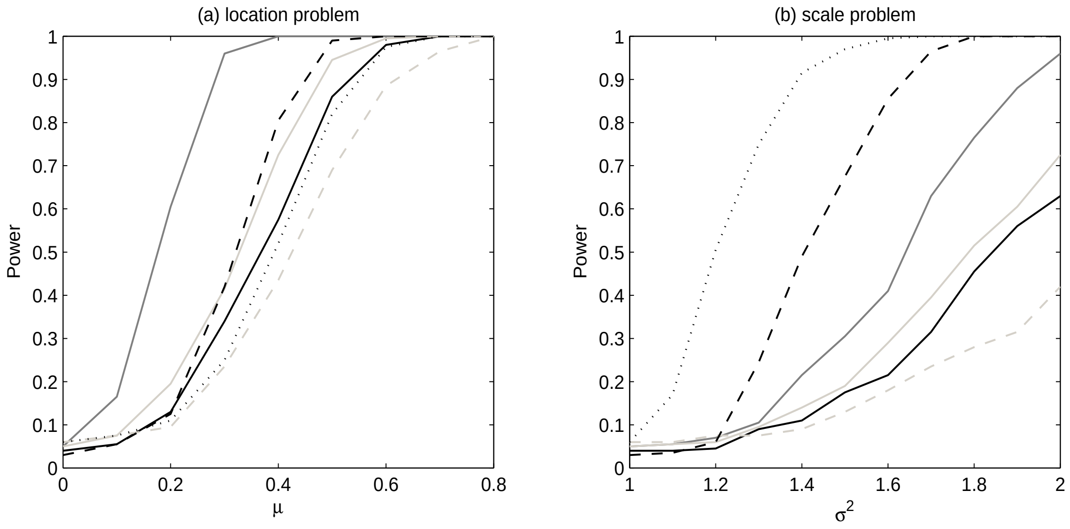


Figure 5.4: Power curves of two-sample tests in normal location and scale problems. (Cramer (dark grey solid), MST run (black solid), NN (light grey solid), Adjacency (light grey dashed), HT (black dashed), proposed (black dotted) tests)

5.6 Proofs and mathematical details

PROOF OF LEMMA 5.1: If $F = G$, there is nothing to prove. So, let us prove the 'only if' part. If $E(\|\mathbf{X}_1 - \mathbf{X}_2\|)$, $E(\|\mathbf{Y}_1 - \mathbf{Y}_2\|)$ and $E(\|\mathbf{X}_1 - \mathbf{Y}_1\|)$ are equal, we have $2E(\|\mathbf{X}_1 - \mathbf{Y}_1\|) - E(\|\mathbf{X}_1 - \mathbf{X}_2\|) - E(\|\mathbf{Y}_1 - \mathbf{Y}_2\|) = 0$. Now, from Baringhaus and Franz (2004), we know that for F and G with finite expected norm, $2E(\|\mathbf{X}_1 - \mathbf{Y}_1\|) - E(\|\mathbf{X}_1 - \mathbf{X}_2\|) - E(\|\mathbf{Y}_1 - \mathbf{Y}_2\|) = 0$ implies $F = G$. □

PROOF OF THEOREM 5.1: If F and G satisfy (A1)-(A3), using the results (a)-(c) stated in Section 4.2, for fixed n_1, n_2 and $d \to \infty$, we have $\hat{\mu}_{FF}/\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2}$, $\hat{\mu}_{GG}/\sqrt{d} \xrightarrow{P} \sigma_2\sqrt{2}$ and $\hat{\mu}_{FG}/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2}$. Let these three limiting values be denoted by υ_1, υ_2 and υ_3, respectively. So, a re-scaled version of our test statistic,

$T^d_{n_1,n_2}/d$ (instead of T_{n_1,n_2}, here we use $T^d_{n_1,n_2}$ to show its dependence on d) converges to $\upsilon_0^* = (\upsilon_1 - \upsilon_3)^2 + (\upsilon_2 - \upsilon_3)^2$ in probability.

Now, let us consider the permutation distribution of $T^d_{n_1,n_2}$ when $n_1 = n_2 = n_0$. If $n_0 - r$ observations ($r = 0, \ldots, n_0$) from F and r observations from G are assumed to come from one distribution and the rest from the other, as $d \to \infty$, the value of the test statistic converges to $\upsilon_r^* = (\upsilon_{1,r} - \upsilon_{3,r})^2 + (\upsilon_{2,r} - \upsilon_{3,r})^2$ in probability, where $\upsilon_{1,r} = \left[\binom{n_0-r}{2}\upsilon_1 + \binom{r}{2}\upsilon_2 + (n_0 - r)r\upsilon_3\right]/\binom{n_0}{2}$, $\upsilon_{2,r} = \left[\binom{r}{2}\ \upsilon_1 + \binom{n_0-r}{2}\upsilon_2 + (n_0 - r)r\upsilon_3\right]/\binom{n_0}{2}$ and $\upsilon_{3,r} = \left[(n_0 - r)r\upsilon_1 + r(n_0 - r)\upsilon_2 + \{(n_0 - r)^2 + r^2\}\upsilon_3\right]/n_0^2$. So, as $d \to \infty$, the permutation distribution tend to have $(n_0 + 1)$ mass points $\upsilon_0^*, \upsilon_1^*, \ldots, \upsilon_{n_0}^*$ with probabilities $\binom{n_0}{n_0}\binom{n_0}{0}/\binom{2n_0}{n_0}$, $\binom{n_0}{n_0-1}\binom{n_0}{1}/\binom{2n_0}{n_0}$, $\ldots$, $\binom{n_0}{0}\binom{n_0}{n_0}/\binom{2n_0}{n_0}$, respectively.

Now, we will show that $\upsilon_r^* \leq \upsilon_0^*$ for all choices of r, where the equality holds for $r = 0$ and $r = n_0$. First note that, under the given condition ($\sigma_1^2 \neq \sigma_2^2$ or $\nu^2 > 0$), we have $2\upsilon_3 - \upsilon_1 - \upsilon_2 > 0$. Also, note that $\upsilon_{1,r}$, $\upsilon_{2,r}$ and $\upsilon_{3,r}$ can be expressed as $\upsilon_{1,r} = \upsilon_1 + \binom{r}{2}(\upsilon_2 - \upsilon_1)/\binom{n_0}{2} + (n_0 - r)r(\upsilon_3 - \upsilon_1)/\binom{n_0}{2}$, $\upsilon_{2,r} = \upsilon_2 + \binom{r}{2}(\upsilon_1 - \upsilon_2)/\binom{n_0}{2} + (n_0 - r)r(\upsilon_3 - \upsilon_2)/\binom{n_0}{2}$ and $\upsilon_{3,r} = \upsilon_3 + (n_0 - r)r(\upsilon_1 + \upsilon_2 - 2\upsilon_3)/n_0^2$. So, we have $\upsilon_{2,r} - \upsilon_{1,r} = (\upsilon_2 - \upsilon_1)\left[\binom{n_0}{2} - r(n_0 - r) - 2\binom{r}{2}\right]/\binom{n_0}{2}$. Now it is easy to check that $-1 \leq \left[\binom{n_0}{2} - r(n_0 - r) - 2\binom{r}{2}\right]/\binom{n_0}{2} \leq 1$, which implies $|\upsilon_{2,r} - \upsilon_{1,r}| \leq |\upsilon_2 - \upsilon_1|$ or $(\upsilon_{2,r} - \upsilon_{1,r})^2 \leq (\upsilon_2 - \upsilon_1)^2$. Unless $\upsilon_1 = \upsilon_2$, here the equality holds only when $r = 0$ or $r = n_0$. Again, we have $2\upsilon_{3,r} - \upsilon_{1,r} - \upsilon_{2,r} = (2\upsilon_3 - \upsilon_1 - \upsilon_2)\left[1 - r(n_0 - r)/n_0^2 \ - \ r(n_0 - r)/\binom{n_0}{2}\right]$. Since $-1 \leq \left[1 - r(n_0 - r)/n_0^2 \ - \ r(n_0 - r)/\binom{n_0}{2}\right] \leq 1$, we have $(2\upsilon_{3,r} - \upsilon_{1,r} - \upsilon_{2,r})^2 \leq (2\upsilon_3 - \upsilon_1 - \upsilon_2)^2$, where the equality holds only when $r = 0$ or $r = n_0$. So, we have $\upsilon_r^* = \frac{1}{2}[(2\upsilon_{3,r} - \upsilon_{1,r} - \upsilon_{2,r})^2 + (\upsilon_{2,r} - \upsilon_{1,r})^2] \leq \frac{1}{2}[(2\upsilon_3 - \upsilon_1 - \upsilon_2)^2 + (\upsilon_2 - \upsilon_1)^2] = \upsilon_0^*$, where equality holds only for $r = 0$ and $r = n_0$.

Therefore, as $d \to \infty$, under the permutation distribution, the test statistic takes the value υ_0^* or higher with probability tending to $2/\binom{n_0}{2}$. So, for all n_0 with $2/\binom{n_0}{2} < \alpha$, the new test rejects H_0 with probability tending to 1 as d tends to infinity. □

PROOF OF THEOREM 5.2: Note that nT_{n_1,n_2} can be expressed as $nT_{n_1,n_2} = \frac{1}{2}\left[\{\sqrt{n}(\hat{\mu}_{FF} - \hat{\mu}_{GG})\}^2 + \ \{\sqrt{n}T^{CR}_{n_1,n_2}\}^2\right]$, where $T^{CR}_{n_1,n_2}$ is the Cramer test statistic. From Baringhaus and Franz (2004), under H_0, we have $nT^{CR}_{n_1,n_2} = O_p(1)$ or $\sqrt{n}T^{CR}_{n_1,n_2} = o_p(1)$. Again, under H_0, $\mu_{FF} = \mu_{GG}$, and hence $\sqrt{n}(\hat{\mu}_{FF} - \hat{\mu}_{GG}) = \sqrt{n}\{(\hat{\mu}_{FF} - \mu_{FF}) - (\hat{\mu}_{GG} -$

$\mu_{GG})\}$. Now, $\hat{\mu}_{FF} - \mu_{FF} = \binom{n_1}{2}^{-1} \sum_{1 \leq i < j \leq n_1} (\|\mathbf{x}_i - \mathbf{x}_j\| - \mu_{FF})$ is a U-statistic with a symmetric kernel function. Therefore, from standard results on U-statistic (see e.g., Lee, 1990), $\sqrt{n_1}(\hat{\mu}_{FF} - \mu_{FF}) \overset{d}{\to} N(0, 4\varsigma^2)$, where $\varsigma^2 = Var(E(\|X_1 - X_2\| | X_1))$. Similarly, $\sqrt{n_2}(\hat{\mu}_{GG} - \mu_{GG}) \overset{d}{\to} N(0, 4\varsigma^2)$, and they are independent. So, using the fact that $n_1/n \to \lambda$ as $n \to \infty$, we have $\sqrt{n}(\hat{\mu}_{FF} - \hat{\mu}_{GG}) = \sqrt{n/n_1}\ \left[\sqrt{n_1}(\hat{\mu}_{FF} - \mu_{FF})\right] - \sqrt{n/n_2}\ \left[\sqrt{n_2}(\hat{\mu}_{GG} - \mu_{GG})\right] \overset{d}{\to} N(0, (\frac{1}{\lambda} + \frac{1}{1-\lambda})4\varsigma^2)$. Therefore, $nT_{n_1,n_2} \overset{d}{\to} \frac{2\varsigma^2}{\lambda(1-\lambda)}\chi_1^2$ as n_1 and n_2 both tend to infinity. □

PROOF OF THEOREM 5.3: Here, $T^*_{n_1,n_2} = n\hat{\lambda}(1-\hat{\lambda})T_{n_1,n_2}/2\hat{\varsigma}^2 \overset{d}{\to} \chi_1^2$ as $n_1, n_2 \to \infty$, and we reject H_0 at level α if $T^*_{n_1,n_2} > \chi^2_{1,\alpha}$. So, the power of the test is given by $P_{H_A}(T^*_{n_1,n_2} > \chi^2_{1,\alpha}) = P_{H_A}\left[\hat{\lambda}(1-\hat{\lambda})T_{n_1,n_2}/2\hat{\varsigma}^2 > \chi^2_{1,\alpha}/n\right]$.

Now, from the results on probability convergence of U-statistic, we have $\hat{\mu}_{FF} \overset{P}{\to} \mu_{FF}$, $\hat{\mu}_{GG} \overset{P}{\to} \mu_{GG}$ and $\hat{\mu}_{FG} \overset{P}{\to} \mu_{FG}$ as $n_1, n_2 \to \infty$. This implies $\hat{\lambda}(1-\hat{\lambda})T_{n_1,n_2}/2\hat{\varsigma}^2 \overset{P}{\to} \lambda(1-\lambda)\{(\mu_{FF} - \mu_{FG})^2 + (\mu_{FG} - \mu_{GG})^2\}/2\varsigma^2$ as n_1 and n_2 tend to infinity. Since $(\mu_{FF} - \mu_{FG})^2 + (\mu_{FG} - \mu_{GG})^2 = 0 \Leftrightarrow \mu_{FF} = \mu_{FG} = \mu_{GG}$, from Lemma 5.1, we have $(\mu_{FF} - \mu_{FG})^2 + (\mu_{FG} - \mu_{GG})^2 = 0$ if and only if $F = G$. So, under H_A, as $n_1, n_2 \to \infty$, $\hat{\lambda}(1-\hat{\lambda})T_{n_1,n_2}/2\hat{\varsigma}^2$ converges (in probability) to a positive quantity, but $\chi^2_{1,\alpha}/n$ converges to 0. Therefore, the power of the test $P_{H_A}\left[\hat{\lambda}(1-\hat{\lambda})T_{n_1,n_2}/2\hat{\varsigma}^2 > \chi^2_{1,\alpha}/n\right]$ converges to 1 as n_1 and n_2 both tend to infinity. □

Chapter 6

Concluding Remarks

In this thesis, we have proposed and investigated four different types of tests for the multivariate two-sample problem. While two types of them (tests proposed in Chapter 2 and 3) have the exact distribution-free property, the other two types (tests proposed in Chapters 4 and 5) are conditionally as well as asymptotically distribution-free. All these tests are applicable to HDLSS data, where the dimension is much larger than the sample size, and good performances of these tests for such high dimensional data have been demonstrated using theoretical as well as numerical results.

The WMW test and the KS test based on linear classification (discussed in Chapter 2) provide good lower-dimensional views of separability between two distributions, and they are particularly useful when the two distributions differ in their locations. However, if the underlying distributions differ only in their scatters and/or shapes, these tests may not be sensitive enough to yield good power. In such cases, it is better to adopt any of the other three methods. Unlike the method based on linear classification, for the tests based on SHP (discussed in Chapter 3), we do not need to split the whole sample into two sub-samples to achieve the distribution-free property. Therefore, if the sample size is very small, the run test based on SHP may outperform the tests based on linear projection, where we need to sacrifice some observations for estimating the optimal discriminating surface. The test proposed in Chapter 5 is based on sample moments of three types of inter-point distances. If the underlying distributions have exponential tails and they differ in their locations and/or scales, this test often outperforms the tests based on SHP and nearest neighbor type coincidences (proposed in Chapters 3 and 4, respectively). However, if the underlying distributions have heavy tails, the tests based on SHP and nearest neighbors are preferred. These two types of tests are more robust against outliers and extreme values generated from heavy tailed distributions. Among these two types of tests, there is no clear winner. Depending on the nature of the difference between the two distributions, one of them outperforms the other. For instance, while the former one performs better in the location problem, in the scale problem, the latter one yields better performance. During our theoretical investigation, though we have assumed either a difference in location or a difference in scale to prove the high dimensional consistency of the proposed tests, our empirical results clearly show that most of these tests, particularly those based on SHP and nearest neighbors, can yield excellent performance even when the two populations have the same location and the same scale, and they differ only in their shapes.

Following the idea of corresponding two-sample tests, in this thesis, we have developed two general methods for multivariate generalizations of the univariate paired-sample tests as well. One of these methods is based on the idea of discriminating hyperplane (see Chapter 2), while the other is based on the idea of SCP (see Chapter 3). Both of these methods lead to tests having the exact distribution-free property in finite sample situations. In the first case, we need to split the whole sample into two sub-samples to achieve this distribution-free property, but no such splitting is required for the tests based on SCP. Therefore, one should prefer the latter one when the sample size is very small. Here it is needless to mention that all these tests for matched pair data can be used even when the dimension exceeds the sample size, and good power properties of these tests for HDLSS data have been established using both theoretical and numerical results. Using a similar type of technique based on differences of observations and their negatives (see Chapters 2 and 3), it is possible to develop paired sample versions of our two-sample tests based on nearest neighbors or that based on averages of inter-point distances. But note that paired sample tests are mainly constructed for the location problem, and we have seen in Chapters 4 and 5 that in the case of location problem, our proposed tests based on nearest neighbors and average inter-point distances are somewhat inferior to the NN test (Schilling (1986a); Henze (1984)) and the

Cramer test (Baringhaus and Franz (2004)), respectively. So, we do not recommend to construct any paired sample version of these proposed tests. Instead, using the ideas of the NN test and the Cramer test, multivariate paired sample tests can be constructed. But, we did not investigate them in this thesis.

The two-sample and paired sample tests proposed in this thesis are all invariant under location shift, rotation and homogeneous scale transformation of the data, but they are not affine invariant. So, when the component variables are not of comparable units and scales, sometimes it can be a better idea to standardize whole the data set (if n is large compared to d) or standardize the component variables (if d is larger than n), and use those standardized observations for testing. This standardization will make these tests scale invariant, but they will lose their rotational invariance property. Several researchers (see e.g., Srivastava et al. (2013); Park and Ayyala (2013)) have pointed out that the rotational invariant tests have drawback in power when variances are inhomogeneous. Standardization can be helpful in such situation.

In this thesis, we have proved the consistency of all proposed tests in HDLSS asymptotic regime. Under reasonable regularity conditions, the powers of these tests increase to unity when the sample size remains fixed and the dimension increases. This high dimensional consistency makes them useful for HDLSS data. However, these tests are consistent in classical asymptotic regime as well, where the dimension remains fixed and the sample size grows to infinity. Except for the tests based on SHP and SCP, we have proved this large sample consistency for all other proposed tests. Though the large sample consistency of the two-sample tests based on SHP is also apparent from our empirical study (see Table 3.1), a formal proof needs to be sketched. Similarly, a formal proof is yet to be framed for the paired sample test based on SCP. One may also be interested in investigating the power properties of our proposed tests when both the dimension and the sample size increase simultaneously. We have not carried out this theoretical investigation in this thesis. This could be an interesting area for future research.

Another interesting area for investigation would be the multisample extension of these proposed two-sample tests. Székely and Rizzo (2004) proposed a multisample version of the Cramer test (Baringhaus and Franz (2004)), where they considered the

Cramer test statistic for each pair of classes and then added them to come up with the final test statistic. Following a similar idea, we can easily develop a multisample test based on averages of inter-point distances, which is expected to outperform Székely and Rizzo's test in various situations including the cases when the distributions differ in their scales. Clearly, this idea can be used for other tests as well, but sometimes it makes the resulting test computationally very expensive, especially when the number of distributions is not very small. In such cases, one needs to find alternative methods. As we have mentioned before, following the idea of Friedman and Rafsky (1979), it is possible to construct a multisample run test based on SHP. The test will retain the distribution-free property, and the cut-off can be determined using the results in Mood (1940). However, one needs to investigate its theoretical and empirical performance. Schilling (1986a) pointed out that the NN test has a natural extension for more than two distributions. Our tests based on nearest neighbors have similar natural extensions as well, but the behavior of the resulting tests needs to thoroughly investigated. Also, it is not clear to us how to use the idea based on linear classification to develop a meaningful distribution-free tests when there are more than two-classes. This seems to be a challenging problem at this moment.

Appendix A

Some existing tests for the multivariate two-sample problem

In this thesis, we have proposed some nonparametric tests for the multivariate two-sample problem and compared their performance with some popular tests available in the literature. A brief description of those existing tests are given below.

A.1 Tests involving two independent samples

Suppose that we have n_1 independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ from F and n_2 independent observations $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ from G, and we want to test null hypothesis $H_0 : F = G$ against the alternative $H_A : F \neq G$. Several multivariate tests are available for this two-sample problem, and some of them have been used in this thesis. Before, we describe them, let us first define the combined sample $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ of size $n = n_1 + n_2$, where $\mathbf{z}_i = \mathbf{x}_i$ for $i = 1, \ldots, n_1$ and $\mathbf{z}_{n_1+i} = \mathbf{y}_i$ for $i = 1, \ldots, n_2$.

- **Hotelling T^2 test** (see e.g., Anderson (2003)): It assumes F and G to be normal with the same scatter and uses the test statistic

$$T^2 = \frac{n_1 n_2}{n}(\bar{\mathbf{x}} - \bar{\mathbf{y}})^T \tilde{S}^{-1}(\bar{\mathbf{x}} - \bar{\mathbf{y}}),$$

where $\bar{\mathbf{x}} = n_1^{-1}\sum_{i=1}^{n_1}\mathbf{x}_i$, $\bar{\mathbf{y}} = n_2^{-1}\sum_{i=1}^{n_2}\mathbf{y}_i$ and $\tilde{S} = [\sum_{i=1}^{n_1}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T + \sum_{i=1}^{n_2}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T]/(n_1 + n_2 - 2)$. The null hypothesis H_0 is rejected for large values of T^2. Under H_0, $\frac{n-d}{d(n-1)}T^2$ follows $F_{d,n-d}$ distribution (F distribution with d and $n-d$ degrees of freedom), and it is used to find the critical value.

- **Puri and Sen's coordinate-wise sign and rank tests (PS-sign and PS-rank tests)** (see Puri and Sen (1971)): Consider a score function $a : \{1, 2, \ldots, n\} \to \mathbb{R}$ and define $\mathbf{z}_i^* = [a(R_i^{(1)}), \ldots, a(R_i^{(d)})]^T$ for $i = 1, \ldots, n$, where $R_i^{(q)}$ $(q = 1, 2, \ldots, d)$ is the rank of $z_i^{(q)}$ in $\{z_1^{(q)}, z_2^{(q)}, \ldots, z_n^{(q)}\}$, and $z_i^{(q)}$ is the q-th coordinate of $\mathbf{z}_i$. Now define $\mathcal{L}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{z}_i^*$, $\mathcal{L}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n}\mathbf{z}_i^*$ and $\tilde{V} = \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(\mathbf{z}_i^* - \mathcal{L}_1)(\mathbf{z}_i^* - \mathcal{L}_1)^T + \sum_{i=n_1+1}^{n}(\mathbf{z}_i^* - \mathcal{L}_2)(\mathbf{z}_i^* - \mathcal{L}_2)^T\right]$. Clearly, a Hotelling T^2 type statistic $T_a = \frac{n_1 n_2}{n}(\mathcal{L}_1 - \mathcal{L}_2)^T\tilde{V}^{-1}(\mathcal{L}_1 - \mathcal{L}_2)$ can be used for testing the null hypothsis. The PS-sign test uses the score function $a(i) = 1$ (respectively, -1) if $i > n/2$ (respectively, $i \leq n/2$) and the PS-rank test uses the score function $a(i) = i$. Each of them rejects H_0 when the observed value of T_a exceeds the critical value, which is determined either using the permutation method or using the asymptotic null distribution of T_a.

- **Spatial sign and rank tests (Sp-sign and Sp-rank tests)** (see e.g., Oja (2010)): Note that the spatial sign function $\mathbb{S}\text{gn}(\cdot)$ is defined as

$$\mathbb{S}\text{gn}(\mathbf{z}) = \begin{cases} \frac{\mathbf{z}}{\|\mathbf{z}\|} & \text{if } \mathbf{z} \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Choose a $d \times d$ non-singular matrix $\mathbb{A}$ and a d-dimensional vector $\mathbf{b}$ such that $\frac{1}{n}\sum_{i=1}^{n}\mathbb{S}\text{gn}(\mathbb{A}(\mathbf{z}_i - \mathbf{b})) = 0$ and $\frac{d}{n}\sum_{i=1}^{n}\mathbb{S}\text{gn}(\mathbb{A}(\mathbf{z}_i - \mathbf{b}))[\mathbb{S}\text{gn}(\mathbb{A}(\mathbf{z}_i - \mathbf{b}))]^T = \mathbf{I}_d$. For instance, Tyler (1987)'s algorithm can be used to find $\mathbb{A}$ and $\mathbf{b}$. Spatial rank ($\mathbb{R}$ank) of $\mathbf{z}_i$ $(i = 1, 2, \ldots, n)$ is defined as $\mathbb{R}\text{ank}(\mathbf{z}_i) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{S}\text{gn}(\mathbb{A}(\mathbf{z}_i - \mathbf{z}_j))$.

The spatial sign (Sp-sign) test uses the test statistic

$$T_{SpS} = d\left[n_1 \left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{S}\text{gn}(\mathbb{A}(\mathbf{z}_i - \mathbf{b}))\right\|^2 + n_2 \left\|\frac{1}{n_2}\sum_{i=n_1+1}^{n}\mathbb{S}\text{gn}(\mathbb{A}(\mathbf{z}_i - \mathbf{b}))\right\|^2\right].$$

The spatial rank (Sp-rank) test uses the test statistic

$$T_{SpR} = \frac{d}{C_{\mathbf{Z}}^2}\left[n_1 \left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{R}\text{ank}(\mathbf{z}_i)\right\|^2 + n_2 \left\|\frac{1}{n_2}\sum_{i=n_1+1}^{n}\mathbb{R}\text{ank}(\mathbf{z}_i)\right\|^2\right],$$

where $C_{\mathbf{Z}}^2 = \frac{1}{n}\sum_{i=1}^{n}\|\mathbb{R}\text{ank}(\mathbf{z}_i)\|^2$.

The null hypothesis H_0 is rejected for higher values of test statistics. When the sample size is large, one can use the large sample distributions of T_{SpS} and T_{SpR} under H_0 to calculate the corresponding critical values. Otherwise, they can be computed using the permutation principle.

- **Multivariate run test based on minimal spanning tree (MST run test)** (see Friedman and Rafsky (1979)): Consider $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ as n vertices of an edge-weighted complete graph $\mathcal{G}$, where the edge joining $\mathbf{z}_i$ and $\mathbf{z}_j$ has the cost $\|\mathbf{z}_i - \mathbf{z}_j\|$. Let $\mathcal{M}$ be the MST of this graph $\mathcal{G}$. The MST run test uses the test statistic given by $T^{MST}_{n_1,n_2} = 1 + \sum_{i=1}^{n-1} \Lambda_i^{\mathcal{M}}$, where $\Lambda_i^{\mathcal{M}}$ denotes the indicator variable that takes the value 1 if and only if the i-th edge of $\mathcal{M}$ connects two observations from two different distributions. This test rejects H_0 for smaller values of $T^{MST}_{n_1,n_2}$, where the cut-off can be obtained either using the permutation method (if n is small) or the asymptotic null distribution of $T^{MST}_{n_1,n_2}$ (if n is large).

- **Test based on number of nearest neighbor type coincidences (NN test)** (see e.g., Schilling (1986a); Henze (1988)): For any fixed choice of k, the NN test statistic is given by

$$T_{NN,k} = \frac{1}{nk}\left[\sum_{i=1}^{n_1}\sum_{r=1}^{k} I_{\mathbf{x}_i}(r) + \sum_{i=1}^{n_2}\sum_{r=1}^{k} I_{\mathbf{y}_i}(r)\right],$$

where $I_{\mathbf{z}}(r)$ denotes the indicator variable that takes the value 1 if and only if $\mathbf{z}$ and its r-th ($r \leq k$) nearest neighbor come from the same distribution. The NN test rejects H_0 for large values of $T_{NN,k}$. Cut-off can be determined either using the permutation method or from the large sample distribution of $T_{NN,k}$ under H_0.

- **Hall and Tajvidi's test based on nearest neighbors (HT test)** (see Hall and Tajvidi (2002)): Define the indicator function $I^c_{\mathbf{z}}(r) = 1 - I_{\mathbf{z}}(r)$, where $I_{\mathbf{z}}(r)$ is as defined in the NN test. The HT test considers the test statistic

$$T_{HT} = \frac{1}{n_1}\sum_{i=1}^{n_1}\sum_{k=1}^{n_2}\left|\sum_{r=1}^{k} I^c_{\mathbf{x}_i}(r) - k\frac{n_2}{n-1}\right| + \frac{1}{n_2}\sum_{i=1}^{n_2}\sum_{k=1}^{n_1}\left|\sum_{r=1}^{k} I^c_{\mathbf{y}_i}(r) - k\frac{n_1}{n-1}\right|.$$

A weighted version of this statistic is also available. Hall and Tajvidi (2002) proposed to use the permutation method to determine the cut-off, and H_0 is rejected if the observed value of T_{HT} exceeds that cut-off.

- **Cramer test** (see Baringhaus and Franz (2004)): It uses the test statistic

$$T^{CR}_{n_1,n_2} = \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|\mathbf{x}_i - \mathbf{y}_j\| - \frac{1}{\binom{n_1}{2}}\sum_{i=1}^{n_1}\sum_{j=i+1}^{n_1}\|\mathbf{x}_i - \mathbf{x}_j\| - \frac{1}{\binom{n_2}{2}}\sum_{i=1}^{n_2}\sum_{j=i+1}^{n_2}\|\mathbf{y}_i - \mathbf{y}_j\|$$

  and rejects H_0 for large positive values of $T^{CR}_{n_1,n_2}$. The cut-off can be computed either using the permutation method or the method based on asymptotic bootstrapped distribution under H_0.

- **Adjacency test based on non-bipartite matching (Adjacency test)** (see Rosenbaum (2005)): First assume that $n = n_1 + n_2$ is even i.e., $n = 2r$ for some integer $r \geq 1$. For any non-bipartite matching (see e.g., Lu et al. (2011)), where r non-overlapping pairs $\{(\mathbf{z}_{i_1}, \mathbf{z}_{i_2}); i = 1, 2, \ldots, r\}$ are formed by taking two indices i_1 and i_2 from $\{1, 2, \ldots, n\}$ at a time, define $\sum_{i=1}^{r}\|\mathbf{z}_{i_1} - \mathbf{z}_{i_2}\|$ as the associated cost. Now, consider the matching that leads to the lowest cost and call it optimal non-bipartite matching. In this optimal matching, the number of pairs having both observations from the same distribution is considered as the test statistic T_{Adj}. If n is odd, a pseudo observation is included in the combined sample, whose distance from any one of the n data points is taken as zero. After obtaining the $(n+1)/2$ optimal pairs, the pair with the pseudo observation is discarded and the rest $(n-1)/2$ pairs are considered to compute T_{Adj}. This test statistic has the exact distribution-free property under H_0, and that distribution is used to determine the cut-off. Naturally, H_0 is rejected for large values of T_{Adj}. Instead of Euclidean distance, the distance between marginal rank vectors can also be used to define the cost, find the associated optimal matching and to compute T_{Adj}.

- **Chen and Qin's test (CQ test)** (see Chen and Qin (2010)): The CQ test statistic is given by

$$T_{CQ} = \frac{1}{\sqrt{v_n^{CQ}}}\left[\frac{\sum_{i\neq j}^{n_1}\mathbf{x}_i^T\mathbf{x}_j}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2}\mathbf{y}_i^T\mathbf{y}_j}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbf{x}_i^T\mathbf{y}_j}{n_1 n_2}\right],$$

where $v_n^{CQ} = \frac{2}{n_1^2(n_1-1)^2} tr\{\sum_{i\neq j}^{n_1}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i,j)})\mathbf{x}_i^T(\mathbf{x}_j - \bar{\mathbf{x}}_{(i,j)})\mathbf{x}_j^T\}$
$+\frac{2}{n_2^2(n_2-1)^2} tr\{\sum_{i\neq j}^{n_2}(\mathbf{y}_i - \bar{\mathbf{y}}_{(i,j)})\mathbf{y}_i^T(\mathbf{y}_j - \bar{\mathbf{y}}_{(i,j)})\mathbf{y}_j^T\}$
$+\frac{4}{n_1^2 n_2^2} tr\{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})\mathbf{x}_i^T(\mathbf{y}_j - \bar{\mathbf{y}}_{(j)})\mathbf{y}_j^T\}$,

$\bar{\mathbf{x}}_{(i)} = \frac{1}{n_1-1}\sum_{j\neq i}\mathbf{x}_j$, $\bar{\mathbf{y}}_{(i)} = \frac{1}{n_2-1}\sum_{j\neq i}\mathbf{y}_j$, $\bar{\mathbf{x}}_{(i,j)} = \frac{1}{n_1-2}\sum_{r\neq i,j}\mathbf{x}_r$, $\bar{\mathbf{y}}_{(i,j)} = \frac{1}{n_2-2}\sum_{r\neq i,j}\mathbf{y}_r$ and $tr(\mathbb{A})$ denotes the trace of the matrix $\mathbb{A}$. The test is performed based on the asymptotic null distribution of T_{CQ} when d also is assumed to increase with n. The null hypothesis is rejected for higher values of T_{CQ}.

- **Test of Srivastava, Katayama and Kano (SKK test)** (see Srivastava et al. (2013)): Define $S_{\mathbf{X}} = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$, $D_{\mathbf{X}}$= diagonal matrix of $S_{\mathbf{X}}$, $S_{\mathbf{Y}} = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T$, $D_{\mathbf{Y}}$= diagonal matrix of $S_{\mathbf{Y}}$, $D_0 = n_1^{-1}D_{\mathbf{X}} + n_2^{-1}D_{\mathbf{Y}}$ and $R = D_0^{-1/2}\left(\frac{S_{\mathbf{X}}}{n_1} + \frac{S_{\mathbf{Y}}}{n_2}\right)D_0^{-1/2}$. The SKK test statistic given by
$$T_{SKK} = \frac{(\bar{\mathbf{x}} - \bar{\mathbf{y}})^T D_0^{-1}(\bar{\mathbf{x}} - \bar{\mathbf{y}}) - d}{\sqrt{v_n^{SKK} c_{d,n}}},$$
where $v_n^{SKK} = 2tr(R^2) - \frac{2(tr(D_0^{-1}S_{\mathbf{X}}))^2}{n_1^2(n_1-1)} - \frac{2(tr(D_0^{-1}S_{\mathbf{Y}}))^2}{n_2^2(n_2-1)}$ and $c_{d,n} = 1 + p^{-3/2}tr(R^2)$. The test is performed based on the asymptotic null distribution of the test statistic when n and d both diverge to infinity. H_0 is rejected for higher values of T_{SKK}.

- **Park and Ayyala's test (PA test)** (see Park and Ayyala (2013)): Define $\bar{\mathbf{x}}_{(i)}$, $\bar{\mathbf{y}}_{(i)}$, $\bar{\mathbf{x}}_{(i,j)}$, $\bar{\mathbf{y}}_{(i,j)}$ as in the CQ test and $S_{\mathbf{X}}$, $S_{\mathbf{Y}}$ as in the SKK test. Also define $S_{\mathbf{X}(i)} = \frac{1}{n_1-2}\sum_{j=1,\ j\neq i}^{n_1}(\mathbf{x}_j - \bar{\mathbf{x}}_{(i)})(\mathbf{x}_j - \bar{\mathbf{x}}_{(i)})^T$ and $S_{\mathbf{X}(i,j)} = \frac{1}{n_1-3}\sum_{r=1,\ r\neq i,j}^{n_1}(\mathbf{x}_r - \bar{\mathbf{x}}_{(i,j)})(\mathbf{x}_r - \bar{\mathbf{x}}_{(i,j)})^T$. The matrices $S_{\mathbf{Y}(i)}$ and $S_{\mathbf{Y}(i,j)}$ can be defined accordingly. Now, consider three other matrices
$$\begin{aligned} S_{1(i,j)} &= (n_1+n_2-4)^{-1}[(n_1-3)S_{\mathbf{X}(i,j)} + (n_2-1)S_{\mathbf{Y}}], \\ S_{2(i,j)} &= (n_1+n_2-4)^{-1}[(n_1-1)S_{\mathbf{X}} + (n_2-3)S_{\mathbf{Y}(i,j)}], \\ S_{12(i,j)} &= (n_1+n_2-4)^{-1}[(n_1-2)S_{\mathbf{X}(i)} + (n_2-2)S_{\mathbf{Y}(j)}], \end{aligned}$$
and their diagonal versions $D_{1(i,j)}$, $D_{2(i,j)}$ and $D_{12(i,j)}$, respectively. The PA test statistic is given by
$$\begin{aligned} T_{PA} = \quad & \frac{n_1+n_2-6}{(n_1+n_2-4)\sqrt{v_n^{PA}}}\left[\frac{1}{n_1(n_1-1)}\sum_{1\leq i\neq j\leq n_1}\mathbf{x}_i^T D_{1(i,j)}^{-1}\mathbf{x}_j\right. \\ & \left. + \frac{1}{n_2(n_2-1)}\sum_{1\leq i\neq j\leq n_2}\mathbf{y}_i^T D_{2(i,j)}^{-1}\mathbf{y}_j - \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbf{x}_i^T D_{12(i,j)}^{-1}\mathbf{y}_j\right], \end{aligned}$$

where $v_n^{PA} = (\frac{n_1+n_2-4}{n_1+n_2-6})^2 \Big\{ \frac{2}{n_1^2(n_1-1)^2} \sum_{i \neq j} \mathbf{x}_i^T D_{1(i,j)}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}_{(i,j)})\mathbf{x}_j^T D_{1(i,j)}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i,j)}) + \frac{2}{n_2^2(n_2-1)^2} \sum_{i \neq j} \mathbf{y}_i^T D_{2(i,j)}^{-1}(\mathbf{y}_j - \bar{\mathbf{y}}_{(i,j)})\mathbf{y}_j^T D_{2(i,j)}^{-1}(\mathbf{y}_i - \bar{\mathbf{y}}_{(i,j)}) + \frac{4}{n_1^2 n_2^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbf{x}_i^T D_{12(i,j)}^{-1} (\mathbf{y}_j - \bar{\mathbf{y}}_{(i,j)})\mathbf{y}_j^T D_{12(i,j)}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i,j)}) \Big\}$. The test is performed using the asymptotic null distribution of the test statistic when n and d both diverge to infinity. The null hypothesis H_0 is rejected for large values of T_{PA}.

A.2 Tests for matched pair data

Here we deal with n independent pairs of observations $\binom{\mathbf{x}_1}{\mathbf{y}_1}, \binom{\mathbf{x}_2}{\mathbf{y}_2}, \ldots, \binom{\mathbf{x}_n}{\mathbf{y}_n}$ from a $2d$-variate distribution with d-dimensional marginals F and G for $\mathbf{X}$ and $\mathbf{Y}$, respectively. We consider the location model, i.e., $F(\mathbf{x}) = G(\mathbf{x} - \boldsymbol{\theta})$ for all $\mathbf{x} \in \mathbb{R}^d$ and some $\boldsymbol{\theta} \in \mathbb{R}^d$, and we test $H_0 : \boldsymbol{\theta} = \mathbf{0}$ against the alternative $H_A : \boldsymbol{\theta} \neq 0$. In such cases, it is a common practice to consider $\{\boldsymbol{\xi}_i = \mathbf{x}_i - \mathbf{y}_i,\ i = 1, \ldots, n\}$ as sample observations and perform one sample tests. Some of the existing one sample tests that we have used in this thesis are briefly described below.

- **Hotelling T^2 test** (see e.g., Anderson (2003)): It uses the test statistic $T^2 = n\bar{\boldsymbol{\xi}}^T S_{\boldsymbol{\xi}}^{-1}\bar{\boldsymbol{\xi}}$, where $\bar{\boldsymbol{\xi}} = n^{-1}\sum_{i=1}^n \boldsymbol{\xi}_i$ and $S_{\boldsymbol{\xi}} = (n-1)^{-1}\sum_{i=1}^n (\boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}})(\boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}})^T$ are the sample mean and the sample covariance matrix, respectively. The null hypothesis H_0 is rejected at level α $(0 < \alpha < 1)$ if the observed value of T^2 exceeds $\frac{nd}{n-d}F_{d,n-d}(\alpha)$, where $F_{d,n-d}(\alpha)$ is the upper α point of the F-distribution with d and $n-d$ degrees of freedom.

- **Puri and Sen's coordinate-wise sign and rank tests (PS-sign and PS-rank tests)** (see Puri and Sen (1971)): Consider a score function $a : \{1, 2, \ldots, n\} \to \mathbb{R}$ and define $\boldsymbol{\xi}_i^\circ = (S_i^{(1)}a(R_i^{(1)}), S_i^{(2)}a(R_i^{(2)}), \ldots, S_i^{(d)}a(R_i^{(d)}))^T$ for $i = 1, \ldots, n$, where $S_i^{(q)} = sign(\xi_i^{(q)})$ and $R_i^{(q)}$ is the rank of $|\xi_i^{(q)}|$ in the set $\{|\xi_1^{(q)}|, |\xi_2^{(q)}|, \ldots, |\xi_n^{(q)}|\}$ (here $\xi_i^{(q)}$ denotes the q-th $(q = 1, 2, \ldots, d)$ coordinate of $\boldsymbol{\xi}_i$). Now consider a Hotelling T^2 type test statistic $T_a = n\bar{\boldsymbol{\xi}}^{\circ T} S_{\boldsymbol{\xi}^\circ}^{-1}\bar{\boldsymbol{\xi}}^\circ$ based on the score function a. T_a gives the PS-sign statistic when $a(i) = 1$ for all $i = 1, 2, \ldots, n$, and it gives the PS-rank statistic if $a(i) = i$ for $i = 1, 2, \ldots, n$. In each of these two cases, H_0 is rejected for higher values of the test statistic, where the cut-off is obtained either using the permutation method or from the large sample distribution of T_a under H_0.

- **Spatial sign and rank tests (Sp-sign and Sp-rank tests)** (see e.g., Oja (2010)): Find a $d \times d$ non-singular matrix $\mathbb{A}$ such that $\frac{d}{n}\sum_{i=1}^{n}\mathbb{S}\text{gn}(\mathbb{A}\boldsymbol{\xi}_i)[\mathbb{S}\text{gn}(\mathbb{A}\boldsymbol{\xi}_i)]^T$ $= \mathbf{I}_d$, where $\mathbb{S}$gn is the spatial sign function defined in Section A.1. One can use Tyler's shape matrix (see e.g., Tyler (1987)) for this purpose. The Sp-sign test uses the test statistic $T_{SpS} = nd\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{S}\text{gn}(\mathbb{A}\boldsymbol{\xi}_i)\|^2$.

  Now, define the spatial rank of $\boldsymbol{\xi}_i$ as $\mathbb{R}\text{ank}(\boldsymbol{\xi}_i) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{S}\text{gn}(\mathbb{A}(\boldsymbol{\xi}_i - \boldsymbol{\xi}_j))$. In the case of Sp-rank test, the matrix $\mathbb{A}$ is chosen such that $\frac{d}{n}\sum_{i=1}^{n}\mathbb{R}\text{ank}(\boldsymbol{\xi}_i)[\mathbb{R}\text{ank}(\boldsymbol{\xi})]^T =$ $C^2_{\boldsymbol{\xi}}\mathbf{I}_d$, where $C^2_{\boldsymbol{\xi}} = \frac{1}{n}\sum_{i=1}^{n}\|\mathbb{R}\text{ank}(\boldsymbol{\xi}_i)\|^2$. The Sp-rank test uses the test statistic given by $T_{SpR} = \frac{nd}{4C^2_{\boldsymbol{\xi}}}\|\frac{2}{n(n+1)}\sum_{i \leq j}\mathbb{S}\text{gn}(\mathbb{A}(\boldsymbol{\xi}_i + \boldsymbol{\xi}_j))\|^2$.

  In each of these cases, either the permutation method or the large sample distribution of the test statistic can be used to determine the critical value, and H_0 is rejected for higher values of T_{SpS} and T_{SpR}.

- **Chen and Qin's test (CQ test)** (see Chen and Qin (2010)): The one sample CQ test statistic is given by

$$T_{CQ} = \frac{\frac{1}{n(n-1)}\sum_{i \neq j}\boldsymbol{\xi}_i^T\boldsymbol{\xi}_j}{\sqrt{\frac{2}{n(n-1)}tr(\sum_{i \neq j}(\boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}}_{(i,j)})\boldsymbol{\xi}_i^T(\boldsymbol{\xi}_j - \bar{\boldsymbol{\xi}}_{(i,j)})\boldsymbol{\xi}_j^T)}},$$

  where $\bar{\boldsymbol{\xi}}_{(i,j)} = \sum_{k=1,k \neq i,j}^{n}\boldsymbol{\xi}_k$ is the sample mean computed excluding $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$. The null hypothesis is rejected if the observed value of T_{CQ} exceeds the cut-off determined by the asymptotic null distribution of the test statistic, when n and d both increase to infinity.

- **Srivastava's test (SR test)** (see Srivastava (2009)): The SR test statistic is given by

$$T_{SR} = \frac{n\,\bar{\boldsymbol{\xi}}^T D_{\boldsymbol{\xi}}^{-1}\bar{\boldsymbol{\xi}} - \frac{(n-1)d}{n-3}}{\sqrt{2\Big[tr(R^2_{\boldsymbol{\xi}}) - \frac{d^2}{n-1}\Big]}},$$

  where $\bar{\boldsymbol{\xi}} = n^{1-}\sum_{i=1}^{n}\boldsymbol{\xi}_i$, $D_{\boldsymbol{\xi}}$ is the diagonal matrix of the sample covariance matrix $S_{\boldsymbol{\xi}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}})(\boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}})^T$ and $R_{\boldsymbol{\xi}} = D_{\boldsymbol{\xi}}^{-1/2}S_{\boldsymbol{\xi}}D_{\boldsymbol{\xi}}^{-1/2}$. This test is based on the asymptotic null distribution of T_{SR}, when d increases with n. The null hypothesis is rejected when the observed value of T_{SR} is large.

- **Park and Ayyala's test (PA test)** (see Park and Ayyala (2013)): The PA test uses the test statistic

$$T_{PA} = \frac{n-5}{n(n-1)(n-3)} \sum_{i \neq j} \boldsymbol{\xi}_i^T D^{-1}_{\boldsymbol{\xi}(i,j)} \boldsymbol{\xi}_j,$$

where $D_{\boldsymbol{\xi}(i,j)}$ is the diagonal matrix of $S_{\boldsymbol{\xi}(i,j)} = \frac{1}{n-3} \sum_{r=1,\ r \neq i,j}^{n} (\boldsymbol{\xi}_r - \bar{\boldsymbol{\xi}}_{(i,j)})(\boldsymbol{\xi}_r - \bar{\boldsymbol{\xi}}_{(i,j)})^T$ and $\bar{\boldsymbol{\xi}}_{(i,j)} = \frac{1}{n-2} \sum_{r \neq i,j} \boldsymbol{\xi}_r$. This test rejects H_0 if the observed value of T_{PA} is higher than the cut-off obtained from the asymptotic distribution, which is derived under H_0 when both n and d diverge to infinity.

Bibliography

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Mack, D., and Leine, A. J. (1999). Broad pattern of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences, USA*, 96(12):6745–6750.

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences, USA*, 97(18):10101–10106.

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

Andrews, D. W. K. (1988). Laws of large numbers for dependent nonidentically distributed random variables. *Econometric Theory*, 4(3):458–467.

Aslan, B. and Zech, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *Journal of Statistical Computation and Simulation*, 75(2):109–119.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6(2):311–329.

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206.

Baringhaus, L. and Franz, C. (2010). Rigid motion invariant two-sample tests. *Statistica Sinica*, 20(4):1333–1361.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B.*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Bickel, P. J. (1965). On some asymptotically nonparametric competitors of Hotelling's T^2. *The Annals of Mathematical Statistics*, 36:160–173.

Bickel, P. J. (1969). A distribution-free version of the Smirnov two-sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23.

Bickel, P. J. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.

Biswas, M. and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, 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(4):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. *Statistica Sinica*, 25(4):1421–1435.

Blumen, I. (1958). A new bivariate sign test for location. *Journal of the American Statistical Association*, 53:448–456.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Chakraborty, B., Chaudhuri, P., and Oja, H. (1998). Operating transformation retransformation on spatial median and angle test. *Statistica Sinica*, 8(3):767–784.

Chaudhuri, P. and Sengupta, D. (1993). Sign tests in multidimension: inference based on the geometry of the data cloud. *Journal of the American Statistical Association*, 88(424):1363–1370.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835.

Choi, K. and Marden, J. (1997). An approach to multivariate rank tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 92(440):1581–1590.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Cuesta-Albertos, J. and Febrero-Bande, M. (2010). A simple multiway ANOVA for functional data. *Test*, 19(3):537–557.

Cuesta-Albertos, J. A., Fraiman, R., and Ransford, T. (2007). A sharp form of the Cramér-Wold theorem. *Journal of Theoretical Probability*, 20(2):201–209.

de Jong, R. M. (1995). Laws of large numbers for dependent heterogeneous processes. *Econometric Theory*, 11(2):347–358.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Eisen, M. B. and Brown, P. O. (1999). DNA arrays for analysis of gene expression. *Methods in Enzymology*, 303:179–204.

Ferger, D. (2000). Optimal tests for the general two sample problem. *Journal of Multivariate Analysis*, 74(1):1–35.

Fix, E. and Hodges, J. L. J. (1989). Discriminatory analysis. nonparametric discrimination: consistency properties. *International Statistical Review*, 57:238–247.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717.

Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs: a Markov chain approach. *Journal of the American Statistical Association*, 89(427):1050–1058.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-completeness.* W. H. Freeman and Co., San Francisco, California.

Ghosh, A. K. and Biswas, M. (2015). Distribution-free high dimensional two-sample tests based on discriminating hyperplanes. *Test*, To appear.

Ghosh, A. K. and Chaudhuri, P. (2005). On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, 11(1):1–27.

Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric Statistical Inference.* Marcel Dekker, New York.

Gordon, L., Schilling, M. F., and Waterman, M. S. (1986). An extreme value theory for long head runs. *Probability Theory and Related Fields*, 72(2):279–287.

Gorman, R. P. and Sejnowski, T. J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1(1):75–89.

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

Guyon, I., Li, J., Mader, T., Pletscher, P. A., Schneider, G., and Uhr, M. (2007). Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark. *Pattern Recognition Letters*, 28(12):1438–1444.

Hájek, J., Šidák, Z., and Sen, P. K. (1999). *Theory of Rank Tests.* Academic Press, San Diego, California.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society. Series B.*, 67(3):427–444.

Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374.

Hallin, M. and Paindaveine, D. (2002). Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *The Annals of Statistics*, 30(4):1103–1133.

Hallin, M. and Werker, B. J. (2003). Semi-parametric efficiency, distribution-freeness and invariance. *Bernoulli*, 9(1):137–165.

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415.

Henze, N. (1984). Über die Anzahl von Zufallspunkten mit typ-gleichem nächsten Nachbarn und einen multivariaten Zwei-Stichproben-Test. *Metrika*, 31(5):259–273.

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783.

Henze, N. and Penrose, M. D. (1999). On the multivariate runs test. *The Annals of Statistics*, 27(1):290–298.

Hettmansperger, T. P., Möttönen, J., and Oja, H. (1997). Affine-invariant multivariate one-sample signed-rank tests. *Journal of the American Statistical Association*, 92(440):1591–1600.

Hettmansperger, T. P., Möttönen, J., and Oja, H. (1998). Affine invariant multivariate rank tests for several samples. *Statistica Sinica*, 8(3):785–800.

Hettmansperger, T. P., Nyblom, J., and Oja, H. (1994). Affine invariant multivariate one-sample sign tests. *Journal of the Royal Statistical Society. Series B*, 56(1):221–234.

Hettmansperger, T. P. and Oja, H. (1994). Affine invariant multivariate multisample sign tests. *Journal of the Royal Statistical Society. Series B.*, 56(1):235–249.

Hodges, J. L. (1955). A bivariate sign test. *The Annals of Mathematical Statistics*, 26(3):523–527.

Hollander, M. and Wolfe, D. A. (1999). *Noparametric Statistical Methods*. Wiley, New York.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6):4104–4130.

Kolmogorov, A. N. and Rozanov, Y. A. (1960). On strong mixing conditions for stationary Gaussian processes. *Theory of Probability and its Applications*, 5(2):204–208.

Koo, J.-Y., Lee, Y., Kim, Y., and Park, C. (2008). A Bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, 9:1343–1368.

Kruskal, Jr., J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50.

Lawler, E. L., Lenstra, J. K., Kan, A. R., and Shmoys, D. B. (1985). *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley, New York.

Lee, A. J. (1990). *U-statistics: Theory and Practice*. Marcel Dekker, New York.

Lin, Y. (2002). A note on margin-based loss function in classification. *Statistics and Probability Letters*, 68:73–82.

Liu, A., Li, Q., Liu, C., Yu, K., and Yu, K. F. (2010). A rank-based test for comparison of multidimensional outcomes. *Journal of the American Statistical Association*, 105(490):578–587.

Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260.

Liu, Z. and Modarres, R. (2011). A triangle test for equality of distribution functions in high dimensions. *Journal of Nonparametric Statistics*, 23(3):605–615.

Loftsgaarden, D. O. and Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051.

Lopes, M., Jacob, L., and Wainwright, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. *Advances in Neural Information Processing Systems*, 24:1206–1214.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011). Optimal nonbipartite matching and its statistical applications. *The American Statistician*, 65(1):21–30.

Maa, J.-F., Pearl, D. K., Bartoszyński, R., et al. (1996). Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *The Annals of Statistics*, 24(3):1069–1074.

Marden, J. I. (1999). Multivariate rank tests. In *Multivariate Analysis, Design of Experiments, and Survey Sampling*, pages 401–432. Marcel Dekker, New York.

Mardia, K. V. (1967). A non-parametric test for the bivariate two-sample location problem. *Journal of Royal Statistical Society. Series B*, 29:320–342.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, London.

Marron, J. S., Todd, M. J., and Ahn, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271.

Mondal, P. K., Biswas, M., and Ghosh, A. K. (2015). On high dimensional two-sample tests based on nearest neighbors. *Journal of Multivariate Analysis*, 141:168–178.

Mood, A. M. (1940). The distribution theory of runs. *The Annals of Mathematical Statistics*, 11:367–392.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, 5(2):201–213.

Möttönen, J., Oja, H., and Tienari, J. (1997). On the efficieny of multivariate sign and rank tests. *The Annals of Mathematical Statistics*, 25:542–552.

Oja, H. (2010). *Multivariate Nonparametric Methods with R.* Springer, New York.

Oja, H. and Randles, R. H. (2004). Multivariate nonparametric tests. *Statistical Science*, 19(4):598–605.

Park, J. and Ayyala, D. N. (2013). A test for the mean vector in large dimension and small samples. *Journal of Statistical Planning and Inference*, 143(5):929–943.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36(6):1389–1401.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis.* Wiley, New York.

Qiao, X., Zhang, H. H., Liu, Y., Todd, M. J., and Marron, J. S. (2010). Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statistical Association*, 105(489):401–414.

Randles, R. H. and Peters, D. (1990). Multivariate rank tests for the two-sample location problem. *Communications in Statistics. Theory and Methods*, 19(11):4225–4238.

Rogers, W. H. (1976). Some convergence properties of k-nearest neighbor estimates. *Unpublished Ph. D. Thesis, Stanford University, Department of Statistics.*

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society. Series B*, 67(4):515–530.

Rousson, V. (2002). On distribution-free tests for the multivariate two-sample location-scale model. *Journal of Multivariate Analysis*, 80(1):43–57.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, 24:220–238.

Schilling, M. F. (1986a). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806.

Schilling, M. F. (1986b). Mutual and shared neighbor probabilities: finite- and infinite-dimensional results. *Advances in Applied Probability*, 18(2):388–405.

Schoonover, J. R., Marx, R., and Zhang, S. L. (2003). Multivariate curve resolution in the analysis of vibrational spectroscopy data files. *Applied Spectroscopy*, 57:483–490.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10:262–266.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis*, 100(3):518–532.

Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386–402.

Srivastava, M. S., Katayama, S., and Kano, Y. (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis*, 114:349–358.

Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5.

Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, 15:234–251.

Van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

Vapnik, V. N. (1998). *Statistical Learning Theory.* Wiley, New York.

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11:147–162.

Wei, S., Lee, C., Wichers, L., Li, G., and Marron, J. (2015). Direction-projection-permutation for high dimensional hypothesis tests. *Journal of Computational and Graphical Statistics*, To appear.

Yushkevich, P., Pizer, S. M., Joshi, S., and Marron, J. (2001). Intuitive, localized analysis of shape variability. In *Information Processing in Medical Imaging*, pages 402–408. Springer, New York.