



UNIVERSITY *of the*
WESTERN CAPE

POWER STUDIES OF MULTIVARIATE TWO-SAMPLE TESTS OF COMPARISON



IAN JOHN SILUYELE
UNIVERSITY *of the*
WESTERN CAPE

A thesis submitted in partial fulfillment of the requirements for the degree of
Magister Scientiae in Statistics in the Faculty of Natural Sciences at the University
of the Western Cape.

SUPERVISOR : PROFESSOR CHRIS KOEN

July 6, 2007

KEYWORDS

Multivariate two-sample problem

Non-parametric tests

Multivariate two-sample test

Permutation method

Data depth

Multivariate empirical distribution function

Euclidean distance

Interpoint distance distributions

Nearest neighbour tests

Power



ABSTRACT

POWER STUDIES OF MULTIVARIATE TWO-SAMPLE TESTS OF COMPARISON

IAN JOHN SILUYELE

MSc Statistics Thesis, Department of Statistics, University of the Western Cape.

The multivariate two-sample tests provide a means to test the match between two multivariate distributions. Although many tests exist in the literature, relatively little is known about the relative power of these procedures. The studies reported in the thesis contrasts the effectiveness, in terms of power, of seven such tests with a Monte Carlo study. The relative power of the tests was investigated against location, scale, and correlation alternatives. Samples were drawn from bivariate exponential, normal and uniform populations. Results from the power studies show that there is no single test which is the most powerful in all situations. The use of particular test statistics is recommended for specific alternatives.

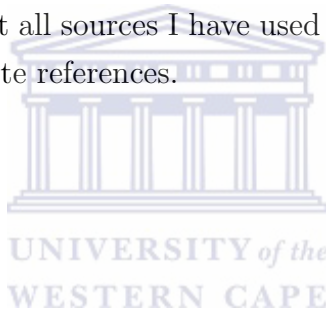
A possible supplementary non-parametric graphical procedure, such as the Depth-Depth plot, can be recommended for diagnosing possible differences between the multivariate samples, if the null hypothesis is rejected.

As an example of the utility of the procedures for real data, the multivariate two-sample tests were applied to photometric data of twenty galactic globular clusters. The results from the analyses support the recommendations associated with specific test statistics.

July 6, 2007

DECLARATION

I declare that *Power Studies of Multivariate Two-Sample Tests of Comparison* is my work, that it has not been submitted before for any degree or examination in any other university, and that all sources I have used or quoted have been indicated and acknowledged by complete references.



IAN JOHN SILUYELE

July 6, 2007

SIGNED

Table of Contents

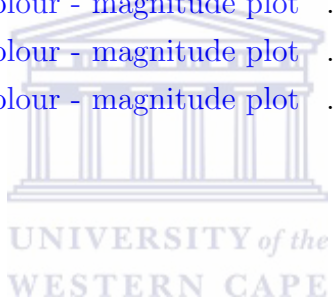
List of Figures	v
List of Tables	vi
List of Acronyms	vii
Acknowledgments	viii
1 General Introduction and Objectives	1
1.1 Introduction	1
1.2 The Research Problem	2
1.3 Objectives	4
1.4 The Thesis Structure	4
2 An Introduction to Two-Sample Testing	6
2.1 Basic Definitions	6
2.2 The Two-Sample Problem	7
2.2.1 The General Null Hypothesis	8
2.2.2 The Alternative Hypothesis	8
2.3 Significance Testing	9
2.4 Estimation of Sampling Distribution	9
2.4.1 The Permutation Method	10
3 Literature Review	12
3.1 The Graphical Approach To Multivariate Two-Sample Testing	12
3.2 Tests Based on the Empirical Distribution Function (EDF)	14

3.2.1	The Simplified Kolmogorov-Smirnov Test	16
3.3	Statistics Based on Interpoint Distance	17
3.3.1	The Henze Nearest Neighbour Statistic	18
3.3.2	The Hall-Tajvidi Statistics	20
3.3.3	The Friedman-Rafsky Statistic	22
3.3.4	An Interpoint Distance Distribution Test	24
3.3.5	The Baringhaus-Franz Statistic	28
3.3.6	The Weiss Statistic	29
3.3.7	The Cross-Match Test	30
3.3.8	Notes	32
4	Power Studies	34
4.1	Sampled Populations	35
4.1.1	Bivariate Normal Population	36
4.1.2	Bivariate Uniform Population	36
4.1.3	Bivariate Exponential Population	37
4.2	Estimation of Power for Finite Samples	37
4.3	Computational Details	38
4.4	Power Comparisons	39
4.4.1	Estimates of Type I Error Rates	39
4.4.2	The Parameter K in the Henze statistic $T_{H(K)}$	41
4.4.3	Bivariate Normal Distribution	45
4.4.4	Bivariate Uniform Distribution	50
4.4.5	Bivariate Exponential Distribution	54
4.4.6	General Discussion and Recommendations	57
4.5	The Depth-Depth Plots	58
5	Analysis of Real Data	62
6	Conclusion	78
	Bibliography	81

List of Figures

4.1	$T_{H(K)}$ power functions for BVN location alternatives	41
4.2	$T_{H(K)}$ power functions for BVN scale alternatives	42
4.3	$T_{H(K)}$ power functions for BVN correlation alternatives	42
4.4	$T_{H(K)}$ power functions for BVU location alternatives	43
4.5	$T_{H(K)}$ power functions for BVU scale alternatives	43
4.6	$T_{H(K)}$ power functions for BVU correlation alternatives	44
4.7	$T_{H(K)}$ power functions for BVE scale/location alternatives	44
4.8	$T_{H(K)}$ power functions for BVE correlation alternatives	45
4.9	Power functions for BVN location alternatives	46
4.10	Power functions for BVN scale alternatives	47
4.11	Power functions for location-adjusted BVN scale alternatives	48
4.12	Power functions for BVN correlation alternatives	49
4.13	Power functions for BVU location alternatives	50
4.14	Power functions for BVU scale alternatives	51
4.15	Power functions for BVU correlation alternatives	53
4.16	Power functions for BVE scale/location alternatives	54
4.17	Power functions for BVE correlation alternatives	56
4.18	Identical populations Depth-Depth plot	59
4.19	Location difference Depth-Depth plot	60
4.20	Scale difference Depth-Depth plot	61
5.1	Positions of chips for NGC 4833	64
5.2	Portions of chips for NGC 4833 analyzed	64
5.3	NGC 4833 cluster colour - magnitude plot	68
5.4	IC 1257 cluster colour - magnitude plot	68
5.5	IC 4499 cluster colour - magnitude plot	69
5.6	NGC 3201 cluster colour - magnitude plot	69

5.7	NGC 4147 cluster colour - magnitude plot	70
5.8	NGC 4372 cluster colour - magnitude plot	70
5.9	NGC 4590 cluster colour - magnitude plot	71
5.10	NGC 5634 cluster colour - magnitude plot	71
5.11	NGC 6171 cluster colour - magnitude plot	72
5.12	NGC 6218 cluster colour - magnitude plot	72
5.13	NGC 6235 cluster colour - magnitude plot	73
5.14	NGC 6256 cluster colour - magnitude plot	73
5.15	NGC 6287 cluster colour - magnitude plot	74
5.16	NGC 6325 cluster colour - magnitude plot	74
5.17	NGC 6342 cluster colour - magnitude plot	75
5.18	NGC 6355 cluster colour - magnitude plot	75
5.19	NGC 6362 cluster colour - magnitude plot	76
5.20	NGC 6380 cluster colour - magnitude plot	76
5.21	NGC 6401 cluster colour - magnitude plot	77
5.22	NGC 6838 cluster colour - magnitude plot	77



List of Tables

4.1	Type I error rates	40
5.1	NGC 4833 Cluster partial data file	65
5.2	Results from step (i) analysis	66
5.3	Results from step (ii) analysis	66
5.4	Summary results of the analysis	67
6.1	Recommended statistics	78



List of Acronyms

BVE	Bivariate exponential
BVN	Bivariate normal
BVU	Bivariate uniform
CDF	Cumulative distribution function
DD-plot	Depth-Depth plot
EDF	Empirical distribution function
IPDD	Interpoint distance distribution
MST	Minimum spanning tree
NNDD	Nearest neighbour distance distribution
ONM	Optimal non-bipartite matching
PC	Planetary camera
SKS	Simplified Kolmogorov-Smirnov
WFC	Wide field camera
WFPC2	Wide field and planetary camera 2

Acknowledgments

I would like to thank the following people variously: Billiard Lishiko, a history PhD candidate, for advice on how the thesis should be written and editorial issues during the initial stage of write-up; Peter Karanja, a coaching staff member from the PET (Postgraduate Enrolment and Throughput) program, for proofreading the thesis for grammatical errors, and general layout; Timothy K. K. Kamanu (The MATLAB guru) for helpful discussions regarding MATLAB programming; Professor Renette Blignaut for her concern, encouragement and facilitation of my Master of Science degree (MSc) programme at the University of the Western Cape; and all members of staff in the department of statistics who helped variously to make my research a possibility.

To my MSc supervisor Professor Chris Koen, thank you for your support, time, encouragement and enthusiasm for my work, the right advice at the right time and everything else. It was a pleasure working with you.

To my mother and all my brothers, sisters and friends: thank you for your support and encouragement when most of you had no slightest inkling what I was doing.

Funding for the research was jointly provided by the Faculty of Science office of the University of the Western Cape through the Department of Statistics, and the African Institute for Mathematical Sciences (AIMS) in Muizenburg, Cape Town, Republic of South Africa.

The thesis was typeset in L^AT_EX, an open source typesetting software.


They sought it with thimbles, they sought it with care...

Lewis Carroll

Chapter 1

General Introduction and Objectives

1.1 Introduction



A statistical problem, which is common in many areas of research, is the need to test whether two samples drawn independently have the same underlying distribution. Such statistical problems are known as two-sample problems. Often, researchers are interested in determining whether the two samples observed from some specific studies or phenomena are statistically different or not. Detailed explanation of the two-sample problem is given in Chapter 2. In particular, this thesis focuses on the multivariate two-sample problem. A statistical measure of the degree of compatibility of the two samples is the basis of the two-sample statistics (see Friedman and Rafsky, 1979; Baringhaus and Franz, 2001; Hall and Tajvidi, 2002; Maa, Pearl and Bartoszyński, 1996; Henze, 1988; Greenberg, 2006; Rosenbaum, 2005; Weiss, 1960). Primarily, the objective of two-sample tests of comparison treated here is to test the validity of the hypothesis that:

The two observed samples come from populations with identical distributions.

Generally, the form of the common distribution assumed under the null hypothesis is not known. For this reason, a parametric approach is ruled out, and non-parametric methods indicated.

Ideally, any statistical test for assessing the hypothesis above should satisfy the following properties:

1. it should be consistent and have good power against all alternatives;
2. the test statistic should be distribution-free and have a known null distribution.

The implication of property (1) is that as the number of observations in the samples increase, the test should be able to reject the hypothesis if the distributions of the parent populations of the two samples are different. As far as property (2) is concerned, it has been difficult to determine the exact null distributions of two-sample test statistics and, therefore, asymptotic approximations have been used instead. Asymptotic approximations depend on assumptions which may not always be met and, furthermore, asymptotic distributions are not available for all multivariate two-sample test statistics (for example the Baringhaus-Franz statistic (Baringhaus and Franz, 2001)). Therefore, because of the potential difficulties of satisfying property (2) stated above, the distribution of the test statistic under the null hypothesis can be approximated very accurately by the permutation method described in Section 2.4.1 of Chapter 2. The permutation approximation of the null distribution is also possible even if the assumptions required for an asymptotic distribution are satisfied.

1.2 The Research Problem

The validity of the null hypothesis is not difficult to assess when the two independent samples being investigated are univariate. In this case, there are several well-known two-sample tests which genuinely satisfy the aforementioned properties. Some of the most commonly used include the Mann-Whitney, two-sample Kolmogorov-Smirnov, Smirnov deviation, Wald-Wolfowitz runs, Cramér-von Mises, Anderson-Darling, and χ^2 tests. Descriptions of these tests can be found in, for example, (Fisz, 1963; Friedman and Rafsky, 1979; Gibbons, 1985; Rohatgi, 1984; Thas, 2001).

Conceptually, some of the univariate tests can be extended to multivariate settings albeit for large sample sizes. One such example is the χ^2 test. Although it can be applied to multidimensional cells, it requires binning and the choice of bin sizes is arbitrary. Many suggestions on binning procedures exist in literature (see Steele (2002) for references). However, in high dimensional space finite, samples are

sparse, a phenomenon referred to in the literature by the term *curse of dimensionality* (Annis, 2006). Therefore, tests based on binning are inefficient (have lower power), unless the sample sizes are very large. On the other hand, tests which are based on the ranks have no obvious nor unique extension to multivariate settings because there is essentially more than one way of ranking higher dimensional observations (Friedman and Rafsky, 1979; Liu, Parelius and Singh, 1999). When applied to marginal distributions, as some researchers have suggested, they neglect information embedded in the dependence structures of the data sets that may be essential in accounting for the degree of similarity between them. Consequently, research has been prompted in the area of new non-parametric two-sample procedures.

Various multivariate two-sample tests satisfying the two aforementioned requirements have been proposed independently. The number of these tests has increased because of recent theoretical developments, for example, Morgenstern's proof of Deuber's theory (Morgenstern, 2001; Baringhaus and Franz, 2001); the theoretical framework for dimension reduction by Maa, Pearl and Bartoszyński (1996); and the expanding capabilities of modern high speed computers which can cope with the heavy computational demand involved. The earliest studies date at least as far back as 1960 (Weiss, 1960). In some papers, theoretical properties of practical importance such as distribution-freeness, consistency against all alternative hypotheses, and relative power performance of the proposed tests have been studied and illustrated via Monte Carlo experiments (Friedman and Rafsky, 1979; Baringhaus and Franz, 2001; Hall and Tajvidi, 2002), while in others these properties have not been investigated (for example Maa, Pearl and Bartoszyński, 1996).

The lack of information about the comparative power properties of the available tests motivated this study. The power of a selection of multivariate two-sample statistics is investigated in the thesis for a variety of distributions. Bivariate normal (symmetric, mesokurtic and infinite support), bivariate exponential (highly skewed and infinite support) and bivariate uniform (symmetric, highly platykurtic and finite support) are used in the studies. The variety of distributions considered will enable users to make informed decisions concerning the test to use.

Power against differences in location (shift), scale (dispersion), and correlation are studied. Fixed sample sizes are used. The significance levels in the power studies

are fixed at a nominal standard value of 5%.

1.3 Objectives

In summary, the work presented in the thesis aims to:

1. provide a review of the literature on multivariate two-sample test statistics for continuous data;
2. conduct power studies of the multivariate two-sample test statistics for selected bivariate distributions;
3. compare the relative power of the selected test statistics.

1.4 The Thesis Structure

Chapter 2 states the two-sample problem, and introduces various terminology, notation and concepts which are used in the rest of the thesis.

Chapter 3 reviews the test statistics studied, including a few not used in the power study. The selection of the tests is primarily based on three criteria including applicability in arbitrary dimensional settings (although only bivariate examples are studied), appealing logic, and, most importantly, simplicity. The study of the literature is mainly limited to statistical tests but also includes an informal exploratory tool for assessing the equality of two multivariate distributions by graphical means. Although there are multivariate two-sample tests for both continuous and discrete data, this thesis concentrates on tests developed for continuous data.

A distinction is made between three broad classes of multivariate two-sample tests investigated in the thesis, namely:

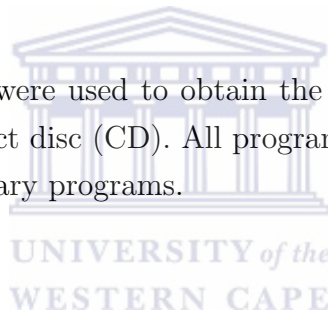
1. graphical exploratory techniques;
2. empirical distribution function type;
3. interpoint distance type.

The general methodology for computing the power of multivariate two-sample statistics, described in Chapter 3, is outlined in Chapter 4. The results from power studies performed through Monte Carlo simulations for a variety of distributions and parameter values, are presented and discussed.

Applications of the studied multivariate two-sample statistics to real data are reported in Chapter 5. More applications of the statistics are given by Koen and Siluyele (2007) (see the article in the directory **Accepted Paper**, on the accompanying CD).

The thesis concludes in Chapter 6 with recommendations and an outlook for possible extensions.

Computer programs, which were used to obtain the results reported, are included on the accompanying compact disc (CD). All programming was done in MATLAB. The CD also includes auxiliary programs.



Chapter 2

An Introduction to Two-Sample Testing

This chapter gives a formal definition of the two-sample problem. Some basic terminology and notation that are used in the thesis are also presented. Some of the terminology given here is drawn from probability theory, hence, for a detailed exposition see the probability literature (for example Bauer, 1972).

2.1 Basic Definitions

The notation $\mathbf{X}=\{\mathbf{X}_1, \dots\}$ and $\mathbf{Y}=\{\mathbf{Y}_1, \dots\}$ will be used to represent collections of d -dimensional random vectors ($d > 1$), defined on sample spaces $S_{\mathbf{X}}$ and $S_{\mathbf{Y}}$, respectively. Realizations of \mathbf{X} and \mathbf{Y} will be denoted by respectively \mathbf{x} and \mathbf{y} . The sets \mathbf{X} and \mathbf{Y} will be called continuous if all their elements \mathbf{X}_i , for all i , and \mathbf{Y}_j , for all j , are continuous random vectors. Throughout this thesis, \mathbf{X} and \mathbf{Y} are assumed to be continuous random vectors drawn from continuous multivariate distributions. For continuous \mathbf{X} and \mathbf{Y} , we assume that the cumulative distribution functions denoted respectively by $F(\mathbf{x})$ and $G(\mathbf{y})$ are differentiable. Hence, the multivariate probability density functions of both \mathbf{x} and \mathbf{y} exist and will be denoted by $f(\mathbf{X})$ and $g(\mathbf{Y})$. Finite sample sizes of \mathbf{X} and \mathbf{Y} will be represented by m and n respectively. The combined sample shall be denoted by

$$\mathbf{Z}_i = \begin{cases} \mathbf{X}_i & , \quad 1 \leq i \leq m, \\ \mathbf{Y}_{i-m} & , \quad m+1 \leq i \leq N, \end{cases}$$

where $N = m + n$. This notation is used for both random vectors \mathbf{X}_i and \mathbf{Y}_j , and their realizations \mathbf{x}_i and \mathbf{y}_j .

We digress slightly in order to clarify the meaning of the underlying distribution function of the observed sample. The cumulative distribution function is uniquely related to a specifically constructed probability law $P_{\mathbb{F}}$ and choice of an appropriate σ -algebra B on the sample space S . The three mathematical objects together form a probability space $(S, B, P_{\mathbb{F}})$ (Bauer, 1972). When an experiment is conducted, a point \mathbf{x} in the sample space S is randomly sampled according to the probability law $P_{\mathbb{F}}$. The process $P_{\mathbb{F}}$ is called the *underlying data generating mechanism*. Additionally, $P_{\mathbb{F}}$ is regarded as an abstract formulation of a statistical or probabilistic model of the mechanism that generates the sample points. The point \mathbf{x} which is chosen determines the outcome of the experiment. According to the definition of a probability space, the event \mathbf{x} is in the σ -algebra B , and the measure $P_{\mathbb{F}}(\mathbf{x})$ denotes the probability of observing an experimental outcome \mathbf{x} (Bauer, 1972). In this thesis, we will not make explicit reference to the probability space but, instead, assumptions about the uniquely constructed cumulative distribution function and, where necessary, the probability density function of the probability space will be made. The cumulative distribution function denoted by $F(\mathbf{x})$, is regarded as the statistical or probabilistic model that generates the sample which is observed in the experiment.

2.2 The Two-Sample Problem

A classical problem in statistical analysis is testing the equality of two distributions based on independent multivariate samples. Several proposals have been made in the literature (Baringhaus and Franz, 2001; Hall and Tajvidi, 2002; Friedman and Rafsky, 1979; Henze, 1988; Greenberg, 2006; Maa, Pearl and Bartoszyński, 1996). The question can be addressed by the application of one of the multivariate two-sample testing procedures outlined in Chapter 3. This kind of problem is generally referred to as the *two-sample problem*.

As in classical hypothesis testing, two hypotheses are constructed in the context of the two-sample problem: the general null hypothesis, the assertion of equality

of distributions; and the general alternative hypothesis, the negation of the null hypothesis.

2.2.1 The General Null Hypothesis

Generally, the hypothesis given on page 1 is symbolically stated as:

$$H_0 : F \equiv G, \quad (2.1)$$

where F and G are the true but unknown cumulative distribution functions of the random variables \mathbf{X}_i and \mathbf{Y}_j , respectively.

In practice, the assumptions, as aforementioned, are that the cumulative distribution functions, $F(\mathbf{x})$ and $G(\mathbf{y})$ with densities $f(\mathbf{x})$ and $g(\mathbf{y})$ are assumed to be continuous on their supports (sample spaces) and when the null hypothesis is true, the cumulative distribution functions have identical sample spaces. If the sample spaces were not identical, the potential differences in sample space might be used to test for differences between cumulative distribution functions (Hall and Tajvidi, 2002). No knowledge of F and G is proclaimed by the researcher under the hypothesis (2.1), only their equivalence.

2.2.2 The Alternative Hypothesis

In the general setting, when the null hypothesis is not true, we do not know in what sense the true distributions $F(\mathbf{x})$ and $G(\mathbf{y})$ of the two populations differ from each other. Therefore, the alternative hypothesis is taken to be the negation of the hypothesis at (2.1) represented symbolically by

$$H_1 : F(\mathbf{x}) \neq G(\mathbf{x}) \text{ for at least one } \mathbf{x}. \quad (2.2)$$

Two-sample tests constructed for this purpose, and which are sensitive to all types of deviations from the null hypothesis, are called *omnibus tests*. Unfortunately, tests of this nature possess very low power for some specific alternatives compared to those two-sample tests which are designed to detect very specific deviations from the null hypothesis in the direction of the alternatives.

2.3 Significance Testing

In the present context, significance tests indicate whether an observed measure of discrepancy between the distributions of two samples could reasonably occur just by chance in the selection processes of the random samples. Highly significant discrepancies imply that there are differences between the respective populations from which the samples were drawn. Generally, testing for significance involves the following procedures:

- a. choose the test statistic which measures possible differences;
- b. determine the sampling distribution which the statistic would have if the populations were the same, that is when the null hypothesis is true;
- c. locate the observed value of the statistic on the distribution in (b).

The statement that the discrepancy we test for is not present in the population implies the null hypothesis (2.1). The probability of the value of a statistic as extreme or more extreme than the observed, calculated taking the null hypothesis to be true, is the p -value. P -values smaller than the level of significance are evidence against the null hypothesis and in favour of a true discrepancy in the populations from which the samples were drawn.

2.4 Estimation of Sampling Distribution

The sampling distribution is the distribution of a statistic based on a random sample from the population. Statistical inference relies on the sampling distribution of the statistics. However, if the exact or asymptotic null distribution of the statistic is unknown, then it may still be possible to estimate the null distribution and the p -value of the statistic by either bootstrapping or permutation methods (Baringhaus and Franz, 2001). The latter method is used for two-sample problems considered in the thesis. Bootstrapping can also be applied (for details regarding this method see Baringhaus and Franz (2001)). In implementing the permutation method to estimate the sampling distribution of the statistic, the observed random sample is taken to be the “population”. Then, in the place of many random samples from the population, many resamples are created by repeatedly sampling without replacement from the original samples as is explained below.

2.4.1 The Permutation Method

Permutation distributions provide reliable substitutes for formula-based asymptotic distributions of statistics. The main step in the general procedure of permutation tests is to form permutation samples in a way that is consistent with the null hypothesis. Below is an outline of the permutation procedure for testing the compatibility between distributions of two multivariate samples.

Consider two multivariate samples \mathbf{X} and \mathbf{Y} of sizes m and n drawn independently. We merge the samples, since under the null hypothesis the underlying multivariate distributions of the parent populations are presumed to be the same. Thus the population under the null hypothesis is represented by the original pooled sample \mathbf{Z} . From this sample, we randomly choose a subsample of size m and assign it to sample $\mathbf{X}^{(1)}$. The remaining subsample of size n becomes sample $\mathbf{Y}^{(1)}$. The sample $\mathbf{X}^{(1)}$ is an ordinary simple random sample (SRS) drawn without replacement (sampling without replacement means that, after we randomly draw an observation from the pooled sample it cannot be drawn again). The statistic of interest is computed - in the context of this thesis it is a measure of discrepancy between the two observed multivariate samples. The resampling process and computation of the statistic are repeated for all $Q = \binom{N}{m}$ possible permutations of the two sample combinations from the pooled sample \mathbf{Z} , where $N = m + n$. The distribution formed by the statistics from the resamples estimates the sampling distribution of the statistic when the null hypothesis is true, and is called a "permutation distribution conditioned on the pooled samples" (Hall and Tajvidi, 2002). Obtain order statistics and then choose an integer V_0 from the set $\{1, \dots, V_0, \dots, Q\}$, such that $\alpha' = 1 - \frac{V_0}{Q}$ is as large as possible, without exceeding the nominal significance level α . Take as the critical point the V_0 th order statistic. Label this value $t_{\alpha'}$. Then α' will accurately approximate the exact level of the resulting test. The hypothesis at (2.1) is rejected if the observed value of the statistic is greater than $t_{\alpha'}$, for tests with upper tail rejection regions, for example, the Henze's nearest neighbour test (Henze, 1988). For tests with lower tail rejection region, for example, the Friedman-Rafsky statistic (Friedman and Rafsky, 1979), an analogous procedure is carried out.

For large N , the value of Q is very large making this procedure laborious and expensive in terms of computer power and time. Therefore, in circumstances where

α is given, choose integers V and B , which are such that $V < B$, $B < Q$ and $\alpha \approx 1 - \frac{V}{B+1}$, where V is the position of the V th order statistic of permutation statistics and B is the number of permutations, and proceed as outlined above (Hall and Tajvidi, 2002). In order to obtain accurately estimated p -values, the value of B must be sufficiently large because accuracy of estimation improves as B becomes larger. In the studies reported below, $B = 500$. The procedure was implemented in MATLAB using the routine `permutation_resamples.m` in the folder `Statistics` on the accompanying CD.



Chapter 3

Literature Review

In view of the large literature on multivariate two-sample tests of comparison, we cannot describe adequately in this work all the important developments on the subject. We confine the review to multivariate two-sample tests, which are investigated in the power study, and mention a few others. Various informal and formal procedures have been proposed in the literature to test the hypothesis (2.1). Generally, the tests studied in this work are grouped into three categories, namely, the graphical approach, empirical distribution function based tests, and those based on interpoint distances of observations in the samples.

3.1 The Graphical Approach To Multivariate Two-Sample Testing

This is an exploratory visual approach to comparing the underlying distributions of two multivariate samples. It involves computing the depth of each data point with respect to the centroids of each of the two samples, giving N pairs of depth values (the depth is a measure of "closeness" to the centroid). A plot of the N depth pairs constitutes a "depth-depth plot" or DD-plot. In their work, Liu, Parelius and Singh (1999) observed that different distributional characteristics of the data exhibit different patterns in DD-plots. Distributional differences studied by them included location and scale among others.

The procedure is presented in more detail as follows. Consider samples \mathbf{X} and \mathbf{Y} of sizes m and n respectively. Denote their population distributions by F and G

respectively. Let \mathbf{Z} be the pooled sample. Generally, data depth is a way of measuring how central a given observation $\mathbf{x} \in \mathbb{R}^d$ is with respect to a given distribution or, alternatively, a data cloud. Thus, given the two multivariate samples, the DD-plot is the plot of the depth values of each observation from the pooled sample \mathbf{Z} , relative to F (or sample \mathbf{X}) and relative to G (or sample \mathbf{Y}). If both samples are from the same population, we would expect to see points in the DD-plot cluster around a 45 degree line passing through the origin. Changes in the relation between the two samples will result in changes in the DD-plot.

Several methods of measuring data depth have been proposed (see Liu, Parelius and Singh (1999) for references). Some examples of data depth discussed in Liu, Parelius and Singh (1999) include Euclidean depth, Oja depth, simplicial depth, likelihood depth, and Mahalanobis depth. The usefulness of this method is demonstrated in this thesis via the Mahalanobis depth function.

The Mahalanobis depth $M_h D$ of $\mathbf{u} \in \mathbb{R}^d$ with respect to F is defined by

$$M_h D_F(\mathbf{u}) = \frac{1}{1 + (\mathbf{u} - \mu_F) \Sigma_F^{-1} (\mathbf{u} - \mu_F)'}, \quad (3.1)$$

where μ_F and Σ_F are the mean vector and covariance matrix of F respectively. The DD-plot is

$$DD(F, G) = \{(M_h D_F(\mathbf{z}), M_h D_G(\mathbf{z})), \text{ for all } \mathbf{z} \in \mathbf{Z}\}, \quad (3.2)$$

where $M_h D_F(\mathbf{z})$ and $M_h D_G(\mathbf{z})$ are depth values of \mathbf{z} with respect to F and G respectively. Since F and G are unknown, we construct a DD-plot using a sample version of (3.2):

$$DD(F_m, G_n) = \{(M_h D_{F_m}(\mathbf{z}_i), M_h D_{G_n}(\mathbf{z}_i)), \quad i = 1, \dots, N\}, \quad (3.3)$$

where

$$\begin{aligned} M_h D_{F_m}(\mathbf{z}_i) &= \left\{ 1 + (\mathbf{z}_i - \bar{\mathbf{X}}) \hat{\Sigma}_X^{-1} (\mathbf{z}_i - \bar{\mathbf{X}})' \right\}^{-1} \text{ and} \\ M_h D_{G_n}(\mathbf{z}_i) &= \left\{ 1 + (\mathbf{z}_i - \bar{\mathbf{Y}}) \hat{\Sigma}_Y^{-1} (\mathbf{z}_i - \bar{\mathbf{Y}})' \right\}^{-1}. \end{aligned} \quad (3.4)$$

In (3.4), $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are sample mean vectors and $\hat{\Sigma}_{\mathbf{X}}$ and $\hat{\Sigma}_{\mathbf{Y}}$ are sample covariance matrices of \mathbf{X} and \mathbf{Y} , respectively.

A MATLAB implementation is given on the accompanying CD in the folder Data Depth.

3.2 Tests Based on the Empirical Distribution Function (EDF)

These tests compare two multivariate samples by assessing the proximity of their sample EDFs. To describe the test statistic, the definition of the EDF is critical and, therefore, I begin by giving it here.

In one dimension, the cumulative distribution function is defined as $F(x) = P(X \leq x)$ and is estimated from the sample by the EDF

$$\begin{aligned} F_n(x) &= \frac{\text{number of observations } \leq x}{n} \\ &= \frac{\sum_j^n I(X_j \leq x)}{n}. \end{aligned} \quad (3.5)$$

In (3.5), $I(X_j \leq x)$ is an indicator function which assumes the value one, when the inequality is satisfied, and zero when it is not. Therefore, in one dimension, the EDF is a step function with jumps of size $1/n$ at every observed point. In more than one dimensional the cumulative distribution function (CDF) $F(x, y, \dots)$, analogous to the one dimension case, could be defined as

$$F(x, y, \dots) = P(X \leq x, Y \leq y, \dots). \quad (3.6)$$

The definition of the cumulative distribution function in (3.6) is non-unique because the direction of ordering of the observations $\{x, y, \dots\}$ is arbitrary. In one dimension, the direction of ordering is immaterial because $P(X \leq x) = 1 - P(X \geq x)$, so that the only two realistic data orderings give equivalent distribution functions. In two dimensions there are four evident ways of ordering the observations, given by $(X \leq x, Y \leq y)$, $(X \leq x, Y \geq y)$, $(X \geq x, Y \leq y)$, and $(X \geq x, Y \geq y)$, and each

is equally valid for the definition of the cumulative distribution function (Peacock, 1983). The corresponding forms of the CDFs are given by

$$\begin{aligned}
 F^1(x, y) &= P(X \leq x, Y \leq y), \\
 F^2(x, y) &= P(X \leq x, Y \geq y), \\
 F^3(x, y) &= P(X \geq x, Y \leq y), \\
 F^4(x, y) &= P(X \geq x, Y \geq y).
 \end{aligned} \tag{3.7}$$

The corresponding EDFs are defined as:

$$\begin{aligned}
 \widehat{F}^1(x, y) &= \frac{1}{N} \sum_j^N I(X_j \leq x, Y_j \leq y), \\
 \widehat{F}^2(x, y) &= \frac{1}{N} \sum_j^N I(X_j \leq x, Y_j \geq y), \\
 \widehat{F}^3(x, y) &= \frac{1}{N} \sum_j^N I(X_j \geq x, Y_j \leq y), \\
 \widehat{F}^4(x, y) &= \frac{1}{N} \sum_j^N I(X_j \geq x, Y_j \geq y),
 \end{aligned} \tag{3.8}$$

where $I(\cdot, \cdot)$ is an indicator function, which assumes the value one, when the argument is true, and zero, when the argument is false. The empirical distribution functions defined in (3.8) are all consistent estimators for the corresponding CDFs in (3.7). By contrast with the one-dimensional case they are *not* all equivalent. Justel, Peña and Zamar (1997) presents an alternative procedure for defining higher dimensional empirical distribution functions.

One example of a test statistic based on the empirical distribution function is the simplified Kolmogorov-Smirnov form described in subsection 3.2.1. It is the only test investigated in this thesis which involves the empirical distribution functions of the samples. In the test, the goal is to find the largest difference between the two empirical distribution functions of the samples and this is adopted as the test statistic (Greenberg, 2006).

3.2.1 The Simplified Kolmogorov-Smirnov Test

The two-sample Kolmogorov-Smirnov test was generalized to two dimensions originally postulated by Peacock (1983) and later modified by Fasano and Franscechini (1987). In Peacock's procedure, one searches for the largest difference between the two empirical distribution functions of the two dimensional samples. Implementation of his test requires that the EDFs of the two samples be evaluated in all the N^2 points $\mathbf{z} = (z_{k1}, z_{\ell 2})$ ($k, \ell = 1, \dots, N$) where $\mathbf{z} \in \mathbf{Z}$. Therefore, the test of Peacock (1983) is computationally prohibitive especially when the sample sizes are large. In dimensions higher than two the computational problem is exacerbated further. Therefore, Fasano and Franscechini (1987) proposed a variant of Peacock's test which requires the evaluation of the empirical distribution functions of the two samples only in the N observed points. Their test is significantly quicker, computationally, and in fact it has similar power characteristics (Greenberg, 2006). They adopted as a test statistic the largest cumulative difference evaluated by ranging over the two samples in turn in all the four quadrants around observed sample points i.e. using all four definitions of the EDFs in (3.8). Computer routines for their test are given in Press, Teukolsky, Vetterling and Flannery (1992). Greenberg (2006) further simplified the forms of the statistic by restricting evaluation of the EDF to $\hat{F}^1(x, y)$ in (3.8). Therefore, for two samples with a combined sample of size N , Greenberg's simplified Kolmogorov-Smirnov (SKS) test requires only N evaluations of each of the two EDFs. Obviously, this is a huge improvement as regards the computation burden involved compared to the tests by Peacock (1983) and Fasano and Franscechini (1987). Nevertheless, it comes at the expense of power because results from an empirical investigation into the power performance of the three versions of the test, as reported in Greenberg's thesis (Greenberg, 2006), indicate that the SKS test possesses the lowest power. The lower power of the SKS test is attributable to fact that less information from the data is used, as compared with the Peacock (1983) and Fasano and Franscechini (1987) statistics. Nonetheless, empirical studies suggest that the SKS test is consistent and has reasonable power properties (Greenberg, 2006). The SKS test is preferable for application in more than two dimensions because it is currently the only computationally feasible form.

To see the convenience of the SKS test in more than two dimensions, consider the case of two trivariate samples with combined size N . Peacock's test will require that

the EDFs of the two samples be evaluated in $8N^3$ points. For the test of Fasano and Franscechini (1987), the evaluations reduce to $2 \times 8N$, whereas, for the SKS test, only $2N$ evaluations are required. In general, $2^{d+1}N^d$ evaluations of the EDFs are needed for the test of Peacock (1983), $2^{d+1}N$ for the test of Fasano and Franscechini (1987), and $2N$ for the SKS test. The computational burden of the other tests (Peacock, 1983 and Fasano and Franscechini, 1987) in arbitrary dimensions renders them impracticable.

Formally, for bivariate samples $\mathbf{X} = \{(x_{1j}, y_{1j}); 1 \leq j \leq m\}$ and $\mathbf{Y} = \{(x_{2k}, y_{2k}); 1 \leq k \leq n\}$, with respective empirical distribution functions F_m and F_n ,

$$T_{\text{SKS}}^i = \sqrt{\frac{mn}{m+n}} \sup_{(x,y) \in Z} \left| \widehat{F}_m^i(x, y) - \widehat{F}_n^i(x, y) \right|. \quad (3.9)$$

For the bivariate SKS statistic investigated in the thesis only the form $\widehat{F}^4(x, y)$ of the EDF is used [see (3.8)]. For convenience T_{SKS}^4 will simply be denoted by T_{SKS} .

The MATLAB routine for calculating the statistic T_{SKS} is `SKS_perm_test.m` (Greenberg, 2006), given in the directory `Statistics` on the accompanying CD.

The test statistic (3.9) is used to assess the hypothesis (2.1) against the alternative hypothesis (2.2). The null distribution of T_{SKS} is not known and, therefore, the critical value and the p-value of T_{SKS} are estimated by the permutation method described in Section 2.4.1. Values of T_{SKS} greater than the critical value, establishes the difference between population distributions of the observed multivariate samples.

3.3 Statistics Based on Interpoint Distance

The majority of multivariate tests investigated in this thesis are based on interpoint distances of the samples. Some typical examples of distance functions are

$$\max_{1 \leq i \leq d} |u_i - v_i|, \quad (3.10)$$

$$\sum_{i=1}^d |u_i - v_i|, \quad (3.11)$$

$$\left\{ \sum_{i=1}^d (u_i - v_i)^2 \right\}^{\frac{1}{2}}, \quad (3.12)$$

where u_i and v_i are components of d -dimensional vectors \mathbf{u} and \mathbf{v} , observations from multivariate samples \mathbf{X} or \mathbf{Y} or \mathbf{Z} . The Euclidean metric in (3.12) is used throughout the thesis, unless stated otherwise.

Henze (1988), Schilling (1986), Weiss (1960), Hall and Tajvidi (2002), and Friedman and Rafsky (1979) have used interpoint distances for determining nearest neighbours in their proposed multivariate two-sample tests of comparison. Other tests discussed in this chapter which are based on interpoint distances are those proposed by Baringhaus and Franz (2001) and Rosenbaum (2005), while the work of Maa, Pearl and Bartoszyński (1996) is a theoretical framework for dimension reduction which results in univariate distributions of interpoint distances. The tests are explained in detail in the following sections.

3.3.1 The Henze Nearest Neighbour Statistic

The statistics proposed by Henze (1988) and Schilling (1986) are quite similar. The test statistic by Henze (1988) is preferable because unlike Schilling's (Schilling, 1986) which is restricted to the Euclidean metric for determination of nearest neighbours, it is available for general distance metrics [see equations (3.10), (3.11) and (3.12) for some examples of distance metrics available]. Schilling (1986) studied the theoretical properties of his statistics, including consistency and power. Power performance of the various statistics introduced in his paper was studied in a simulation experiment in which conditions were matched with those used by Friedman and Rafsky (1979) in their power studies. The conclusions which were drawn from their investigations are also true for the statistic proposed by Henze (1988), as stated in the latter paper. In this thesis, the test statistic by Henze (1988) is preferred.

The test proposed by Henze (1988) is defined in the following way (see Section 2.2 for notation). Let $\|\cdot\|$ represent a general norm on \mathbb{R}^d . Define the r th nearest neighbour of \mathbf{Z}_i by $N_r(\mathbf{Z}_i)$, as that observation \mathbf{Z}_j which is such that $\|\mathbf{Z}_\nu - \mathbf{Z}_i\| \leq \|\mathbf{Z}_j - \mathbf{Z}_i\|$

for exactly $r - 1$ values of ν , $1 \leq \nu \leq N$; $\nu \neq i, j$. Define the indicator function

$$I_i(r) = \begin{cases} 1 & , \quad \text{if } \mathbf{Z}_i \text{ and } N_r(\mathbf{Z}_i) \text{ are from the same sample,} \\ 0 & , \quad \text{otherwise.} \end{cases}$$

Let K be a small integer (typically $1 \leq K \leq 6$). To test the null hypothesis H_0 , we use the statistic given by

$$T_{H(K)} = \sum_{j=1}^N \sum_{i=1}^K I_j(i),$$

that is, the number of same-type nearest neighbours amongst the K nearest neighbours, and summed over the pooled samples. If the two populations are not identical, samples from one population will tend to cluster together in d -space. Therefore, large values of $T_{H(K)}$ are expected under the alternative hypothesis (2.2). Henze (1988) showed that for large samples the probability of the error of the first kind does not depend on the hypothesized distribution and, therefore, the test is asymptotically distribution-free. Further, he showed that when the null hypothesis is true, conditionally on the pooled sample, and for a general distance metric, the statistic $T_{H(K)}$ is asymptotically normal. The asymptotic distribution of $T_{H(K)}$ is calculated as follows:

- (i) Define an indicator variable a_{ij}^+ by

$$a_{ij}^+ = \begin{cases} 1 & , \quad \text{if } \mathbf{z}_j \text{ is amongst the set of } K \text{ nearest neighbours of } \mathbf{z}_i, \\ 0 & , \quad \text{otherwise;} \end{cases}$$

- (ii) For each observation \mathbf{z}_j in the K nearest neighbour graph of $\mathbf{z}_1, \dots, \mathbf{z}_N$, the indegree is given by

$$d_j^{(K)} = \sum_{i=1}^N a_{ij}^+, \quad 1 \leq j \leq N;$$

- (iii) Define the quantities $C_N^{(K)}$ and $V_N^{(K)}$ by

$$C_N^{(K)} = \frac{1}{NK} \sum_{i=1}^N \sum_{j=1}^N a_{ij}^+ a_{ji}^+ \quad \text{and} \quad V_N^{(K)} = \frac{1}{NK} \sum_{i=1}^N (d_i^{(K)} - K)^2.$$

- (iv) The parameters of the asymptotic distribution of $T_{H(K)}$ are approximated by

$$E(T_{H(K)}) = K \left\{ \frac{m(m-1) + n(n-1)}{N-1} \right\}, \quad (3.13)$$

$$\text{Var}(T_{H^{(k)}}) = K \frac{mn}{N-1} \times \left\{ \frac{4(m-1)(n-1)}{(N-2)(N-3)} \left(1 + V_N^{(k)} - \frac{2K}{N-1} \right) + A \right\}, \quad (3.14)$$

where

$$A = \left(1 - \frac{4(m-1)(n-1)}{(N-2)(N-3)} \right) C_N^{(k)}.$$

For sufficiently large sample sizes m and n ,

$$\tilde{T}_{H^{(k)}} = \frac{T_{H^{(k)}} - E(T_{H^{(k)}})}{\sqrt{\text{Var}(T_{H^{(k)}})}}$$

is approximately standard normal. The null hypothesis (2.1) is rejected at the nominal significance level α , if

$$\tilde{T}_{H^{(k)}} \geq C_\alpha,$$

where C_α is the $100(1 - \alpha)$ th percentile of the standard normal cumulative distribution function (Henze, 1988).

UNIVERSITY of the

The MATLAB computer routines for computing the statistic $T_{H^{(k)}}$ are given on the accompanying CD in the directory `Statistics`. `HenzeNN_perm_test.m` is the permutation implementation of the test which was used. `HenzeNN_Asy_test.m` is the asymptotic implementation of the test.

3.3.2 The Hall-Tajvidi Statistics

In a somewhat similar procedure to the work by Henze (1988) and Schilling (1986), Hall and Tajvidi (2002) made use of interpoint distances to determine the number of nearest neighbours of each observations in the pooled sample. The interpoint distances can be computed by any of the equations (3.10), (3.11), and (3.12). Their test statistic as defined at (3.15) is a weighted sum of absolute deviations of the number of nearest neighbours of each observation from the respective samples, from their respective expected values deduced by permutation argument. The power of the statistics for various combination of weights are investigated in a simulation study. Hall and Tajvidi (2002) performed the study of power of the two statistics at (3.15) for location as well as scale alternatives in a multivariate setting, and included Mann-Whitney and two-sample Kolmogorov-Smirnov for the same distri-

butional characteristics in a univariate setting.

The computational procedure of the statistics is described in the following way. The distance measure denoted by $D(\mathbf{u}, \mathbf{v})$ on the sample space, is the basis for the test. Compute distances of each observation to all other observations in the pooled sample \mathbf{Z} , i.e. compute $D(\mathbf{X}_i, \mathbf{Z}_k)$ for $\mathbf{Z}_k \in \mathbf{Z} \setminus \mathbf{X}_i$ $\{i = 1, \dots, m\}$, and $D(\mathbf{Y}_i, \mathbf{Z}_k)$ for $\mathbf{Z}_k \in \mathbf{Z} \setminus \mathbf{Y}_i$ $\{i = 1, \dots, n\}$. For $\{j = 1, \dots, m+n-1\}$, define the following quantities:

- (i) $M_i(j)$ is the number of observations in \mathbf{Y} that are among the j nearest neighbours of \mathbf{X}_i in $\mathbf{Z} \setminus \mathbf{X}_i$ ($i = 1, \dots, m$);
- (ii) $N_i(j)$ is the number of observations in \mathbf{X} that are among the j nearest neighbours of \mathbf{Y}_i in $\mathbf{Z} \setminus \mathbf{Y}_i$ ($i = 1, \dots, n$).

$M_i(j)$ and $N_i(j)$ are computed using the upper portion of the column vector of ordered distances. For distances of observations from sample \mathbf{X} , the number of nearest observations up to the size of sample \mathbf{Y} are used while distances involving observations from sample \mathbf{Y} , number of nearest observations up to the size of sample \mathbf{X} , are used. Hall and Tajvidi (2002) showed that conditional on the pooled sample \mathbf{Z} , $M_i(j)$ and $N_i(j)$ are hypergeometrically distributed random variables when the null hypothesis is true, with means

$$E_0(M_i(j)|Z) = \frac{nj}{m+n-1} \quad \text{and} \quad E_0(N_i(j)|Z) = \frac{mj}{m+n-1},$$

where E_0 is the expectation when the null hypothesis H_0 , is true.

Let DM and DN denote the deviations of M and N from their mean values under H_0 , then

$$DM_i(j) = \left| M_i(j) - \frac{nj}{m+n-1} \right| \quad \text{and} \quad DN_i(j) = \left| N_i(j) - \frac{mj}{m+n-1} \right|.$$

Statistics T_{HT} and S_{HT} for testing the null hypothesis against the omnibus alternative are given by

$$\begin{aligned} T_{\text{HT}} &= \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^n [DM_i(j)]^\gamma w_1(j) + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m [DN_i(j)]^\gamma w_2(j) \\ S_{\text{HT}} &= \sum_{j=1}^n w_1(j) \sup_{1 \leq i \leq m} [DM_i(j)]^\gamma + \sum_{j=1}^m w_2(j) \sup_{1 \leq i \leq n} [DN_i(j)]^\gamma, \quad (3.15) \end{aligned}$$

where γ is a positive exponents ($1 \leq \gamma \leq 2.5$ in Hall and Tajvidi (2002)) and $w_k(j)$ ($k = 1, 2$) are weight functions. Hall and Tajvidi (2002) suggested the possibilities:

- (i) $w_1(j) = 1$ and $w_2(j) = 1$;
- (ii) $w_1(j) = j$ and $w_2(j) = j$;
- (iii) $w_1(j) = n + 1 - j$ and $w_2(j) = m + 1 - j$.

The sampling distribution and critical values of T_{HT} and S_{HT} under the null hypothesis (2.1) are evaluated by the permutation method. When the samples are from identical populations, small values of both statistics T_{HT} and S_{HT} , are expected.

The MATLAB program `Hall_Tajvidi_perm_test.m`, in the folder `Statistics`, on the accompanying CD was used for the permutation implementation of the two statistics.

3.3.3 The Friedman-Rafsky Statistic

The test also referred to as the "multivariate runs test" is a proposition of Friedman and Rafsky (1979). In essence, it is a multivariate version of the Wald-Wolfowitz runs test. Friedman and Rafsky (1979) suggested a sorting scheme for higher dimensional random variables which is analogous to a sorted list in the univariate case. They used the minimum spanning tree (MST), constructed from interpoint distances of the pooled multivariate sample points, as a generalization of the univariate sorted list. Their test statistic is the number of subtrees which result when incompatible connections (edges connecting points from different samples) are removed. Some theoretical properties of their statistic were investigated by Henze and Penrose (1999). Henze and Penrose (1999) confirmed its asymptotic normality and showed theoretically that the multivariate two-sample tests based on it are universally consistent as conjectured by Friedman and Rafsky (1979). Friedman and Rafsky (1979) compared the power of their statistics to parametric competitors (normal likelihood ratio and normal scores test) for location and scale alternatives.

Given a finite set \mathbf{Z} of d -dimensional points ($d \geq 1$), define the spanning tree on \mathbf{Z} as the set of points all of which are connected, such that the connections (called **edges**) have no loops. In other words, starting from any node on the spanning

tree, it is impossible to return to that point in any way except by backtracking i.e. retracing the path you have taken. The tree length is the total of its Euclidean edge lengths. Therefore, an MST is the spanning tree for which the total Euclidean length of the connections is the smallest possible. That is, if each edge (i, j) of a spanning tree has a Euclidean length δ_{ij} , a spanning tree which minimises the sum $\sum \delta_{ij}$ is called an MST. A MATLAB routine for MST is available on the internet from http://www.models.kvl.dk/users/fans/Some_matlab/MST/index.asp. In principle the MST is not necessarily unique, since there may be more than one spanning tree with the same minimal Euclidean length, if there are two or more edges of identical Euclidean length.

To perform the test we proceed as follows:

1. Construct the minimum spanning tree of the pooled sample points \mathbf{Z} ;
2. Remove all edges which connect a point in \mathbf{X} to a point in \mathbf{Y} ;
3. Define the Friedman-Rafsky statistic T_{FR} , as the number of disjoint subtrees (runs) that results.

Equivalently, T_{FR} is one more than the number of edges in the minimum spanning tree which joins observations from different samples. We can compute T_{FR} by counting the number of edges linking observations from different samples and then add one to the total. If samples are from the same population, observations will be well mixed and large values of the statistic T_{FR} are expected. Hence, small values of T_{FR} provide evidence against the null hypothesis (2.1).

Under the null hypothesis the permutation distribution of the statistic is asymptotically normal with mean and variance given by

$$\begin{aligned} E(T_{FR}) &= \frac{2mn}{N} + 1, \\ \text{Var}(T_{FR}|\mathbf{Z}) &= \frac{2mn(2mn - N)}{N^2(N - 1)} + \\ &\quad \frac{2mn(C - N + 2)[N(N - 1) - 4mn + 2]}{N(N - 1)(N - 2)(N - 3)}. \end{aligned} \quad (3.16)$$

The parameter C is dependent on the configuration of the MST. It is the number of edge pairs sharing a common node and is given by

$$C = \frac{1}{2} \sum_{i=1}^N d_i(d_i - 1),$$

where d_i is the degree of node i . The degree of a node is the number of edges incident on it.

The standardized statistic is given by

$$\tilde{T}_{\text{FR}} = \frac{T_{\text{FR}} - E(T_{\text{FR}})}{\sqrt{\text{Var}(T_{\text{FR}}|\mathbf{Z})}}, \quad (3.17)$$

which is asymptotically standard normal.

The Friedman-Rafsky statistic T_{FR} was calculated using the MATLAB routine `Friedman-Rafsky_Asy_test.m` in the directory `Statistics` on the accompanying CD. `Friedman-Rafsky_perm_test.m` is a permutation implementation of the same test.

3.3.4 An Interpoint Distance Distribution Test

The work by Maa, Pearl and Bartoszyński (1996) proposes a theoretical framework for dimension reduction of two multivariate samples into three sets of univariate samples of interpoint distances. Motivated by the recognition that most multivariate two-sample tests are based on interpoint distances of observations in the samples, they showed that under mild conditions, the parent distributions of the two multivariate samples are different, if and only if the distributions of interpoint distances differ within and between distributions. They further suggested using any three-sample statistic (see Kiefer (1959) for some of the appropriate statistics) for testing the homogeneity hypothesis that the three univariate samples have the same distribution. For this thesis, statistics by Kiefer (1959) and Fisz (1963) were preferred because they are consistent and have good power properties.

To compute the test statistic we need a distance function h defined on \mathbb{R}^d . The function h must satisfy some mild assumptions (see *lemma 1* on page 1071 in Maa,

Pearl and Bartoszyński (1996)). Some suitable examples of h are given in equations (3.10) to (3.12).

The hypothesis of equality of two independent continuous multivariate populations is equivalently formulated in terms of the equality of the univariate distributions of interpoint distances as theorem 3.3.1 due to Maa, Pearl and Bartoszyński (1996) shows.

Theorem 3.3.1 : *Let $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ be independently and identically distributed d -dimensional random variables with density f and cumulative distribution function F , let $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ be independently and identically distributed d -dimensional random variables with density g and cumulative distribution function G , and suppose that the \mathbf{X} 's and \mathbf{Y} 's are independent. If the densities f and g , and h satisfy the conditions of lemma 1 of Maa, Pearl and Bartoszyński (1996), then*

$$h(\mathbf{X}_1, \mathbf{X}_2) =_{\ell} h(\mathbf{Y}_1, \mathbf{Y}_2) =_{\ell} h(\mathbf{X}_3, \mathbf{Y}_3) \text{ if and only if } F = G, \quad (3.18)$$

where $=_{\ell}$ indicates equality of distributions.

Maa, Pearl and Bartoszyński (1996) conjectured that the hypothesis of equality of the distributions of interpoint distances (equation (3.18)) is true for all distributions F and G , and every h . It is noteworthy that the three sets of interpoint distances are *not* independent. This has implications for assessing the differences between the three underlying distributions.

To test the hypothesis (3.18), choose a function h , such as the Euclidean metric (equation 3.12) and compute the following pairwise distances

$$h(\mathbf{x}_k, \mathbf{x}_{\ell}) = \left\{ \sum_{i=1}^d (x_{ki} - x_{\ell i})^2 \right\}^{\frac{1}{2}} \quad k = 1, \dots, m-1; \ell = k+1, \dots, m; \quad (3.19)$$

$$h(\mathbf{y}_k, \mathbf{y}_{\ell}) = \left\{ \sum_{i=1}^d (y_{ki} - y_{\ell i})^2 \right\}^{\frac{1}{2}} \quad k = 1, \dots, n-1; \ell = k+1, \dots, n; \quad (3.20)$$

$$h(\mathbf{x}_k, \mathbf{y}_{\ell}) = \left\{ \sum_{i=1}^d (x_{ki} - y_{\ell i})^2 \right\}^{\frac{1}{2}} \quad k = 1, \dots, m; \ell = 1, \dots, n. \quad (3.21)$$

Then, any omnibus univariate test for assessing the equality of three distributions is used to test the hypothesis formulated in theorem 3.3.1, namely, $h(\mathbf{X}_k, \mathbf{X}_\ell) =_\ell h(\mathbf{Y}_k, \mathbf{Y}_\ell) =_\ell h(\mathbf{X}_k, \mathbf{Y}_\ell)$. Rejection of the hypothesis is evidence against the equality of the underlying distributions of the two independent multivariate populations.

One possible statistic for testing the hypothesis at (3.18) is the three-sample Kolmogorov - Smirnov test proposed by David (1958). However, the statistic is very restrictive because it requires that the number of observations in the two samples be equal. This requirement makes it unsuitable for implementation in the IPDD (Interpoint distance distribution) test because the sizes of the three univariate samples of interpoint distances resulting from the multivariate samples are always unequal. However, a number of suitable tests are available in the literature (Fisz, 1963; Kiefer, 1959). Fisz (1963) discusses tests for assessing the equality of distributions of k independent samples. These tests are applicable to the problem above (see Section 10.13 in Fisz (1963)). Kiefer (1959) also gives a method for testing the null hypothesis of equality of k univariate populations. The tests given by both Fisz (1963) and Kiefer (1959) are designed for independent samples. Of course, the three sets of interpoint distances in (3.19) to (3.21) are not mutually independent. This is not important in the context of this test, as permutation tests rather than asymptotic formulae are used to calculate significance levels.

Let $S_{j,n_j}(x)$, $j = 1, 2, 3$ be the EDFs of the three samples. Define the following quantities

$$D_1(n_1, n_2) = \max_x |S_{1,n_1}(x) - S_{2,n_2}(x)| \quad \text{and}$$

$$D_2(n_1, n_2, n_3) = \max_x \left| S_{3,n_3}(x) - \frac{n_1 S_{1,n_1}(x) + n_2 S_{2,n_2}(x)}{n_1 + n_2} \right|.$$

The statistic given by Fisz (1963) is

$$T_F = \max\{A_1, A_2\}, \quad (3.22)$$

where

$$A_1 = \sqrt{\frac{n_2 n_1}{n_1 + n_2}} D_1(n_1, n_2) \quad \text{and} \quad A_2 = \sqrt{\frac{n_3(n_2 + n_1)}{n_1 + n_2 + n_3}} D_2(n_1, n_2, n_3).$$

The Kiefer (1959) statistic for testing the equality of the three populations is defined by

$$T_K = \left\{ \max_x \sum_{j=1}^3 n_j [S_{j,n_j}(x) - \hat{S}(x)]^2 \right\}^{\frac{1}{2}} \quad (3.23)$$

where

$$\hat{S}(x) = \frac{1}{n_1 + n_2 + n_3} \sum_{j=1}^3 n_j S_{j,n_j}(x).$$

The statistics T_F and T_K will be referred to as "interpoint distance distribution" (IPDD) statistics in subsequent chapters.

IPDD_Asy_test.m and IPDD_perm_test.m are the MATLAB programs in the folder **Statistics**, on the accompanying CD in which the statistic T_F and T_K were implemented. In the power studies, the permutation implementation IPDD_perm_test.m, was used.

A variant of the above is to consider the univariate distributions of the nearest neighbour interpoint distances only, instead of the full sets of interpoint distances. The test is referred to as the *Nearest neighbour distance distribution test* (NNDD).

Denoting by $\|\cdot\|$ the Euclidean distance in d -dimensional space \mathbb{R}^d , the three sets of distances are:

$$\begin{aligned} d_j &= \min_{i \neq j} \|\mathbf{x}_i - \mathbf{x}_j\| & i, j = 1, \dots, m; \\ d_k &= \min_{k \neq \ell} \|\mathbf{y}_k - \mathbf{y}_\ell\| & k, \ell = 1, \dots, n; \\ d_{k\ell} &= \min \|\mathbf{x}_k - \mathbf{y}_\ell\| & k = 1, \dots, m; \ell = 1, \dots, n \end{aligned} \quad (3.24)$$

where $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$. The dimension reduction results in three univariate samples of sizes m , n and $m + n$. The statistics (3.22) and (3.23) are used to test the equality of the three univariate distributions. In the sequel, the notation T_F^{NN} and T_K^{NN} is used for the NNDD statistics.

The MATLAB routine NNDD_perm_test.m in the folder **Statistics**, on the accompanying CD, was used to implement the NNDD test via the statistics T_F^{NN} and T_K^{NN} .

3.3.5 The Baringhaus-Franz Statistic

The statistic due to Baringhaus and Franz (2001) was motivated by a conjecture by Deuber (see Morgenstern (2001) for references). Deuber conjectured that:

(A) *For equal numbers of black and white points randomly distributed in Euclidean space the sum of the pairwise distances between points of equal colors is less than or equal to the sum of the pairwise distances between points of different colour;*

(B) *Equality holds only in the case when black and white points coincide.*

The result is stated equivalently as

$$\int \|\mathbf{u} - \mathbf{v}\| dF_n \otimes G_n(\mathbf{u}, \mathbf{v}) - \frac{1}{2} \int \|\mathbf{u}_1 - \mathbf{u}_2\| dF_n \otimes F_n(\mathbf{u}_1, \mathbf{u}_2) - \frac{1}{2} \int \|\mathbf{v}_1 - \mathbf{v}_2\| dG_n \otimes G_n(\mathbf{v}_1, \mathbf{v}_2) \geq 0, \quad (3.25)$$

where \mathbf{U} and \mathbf{V} represent positions of the black and white points with respective empirical distributions F_n and G_n (Baringhaus and Franz, 2001), and $\mathbf{u}_1, \mathbf{u}_2 \in \mathbf{U}$; $\mathbf{v}_1, \mathbf{v}_2 \in \mathbf{V}$.

For independent $\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_1, \mathbf{Y}_2$ Baringhaus and Franz (2001) deduced the inequality

$$E\|\mathbf{X}_1 - \mathbf{Y}_1\| - \frac{1}{2}E\|\mathbf{X}_1 - \mathbf{X}_2\| - \frac{1}{2}E\|\mathbf{Y}_1 - \mathbf{Y}_2\| \geq 0. \quad (3.26)$$

Equality holds only if the two populations are identical (Baringhaus and Franz, 2001).

The proof by Morgenstern (2001) of the conjecture (3.25) motivated the test of Baringhaus and Franz (2001). Their test statistic is a weighted sum of interpoint distances within and between samples. It is shown to be consistent against all alternatives and has good power performance against some parametric and non-parametric competitors for location and scale alternatives.

With the assumption that \mathbf{X} and \mathbf{Y} have finite expectation, Baringhaus and Franz (2001) proposed using the sample version of (3.26) to assess the validity of the

hypothesis of equality of the distributions i.e.

$$\begin{aligned} T_{\text{BF}} = & \frac{1}{m+n} \sum_{i=1}^m \sum_{j=1}^n \|\mathbf{X}_i - \mathbf{Y}_j\| - \frac{mn}{2(m+n)m^2} \sum_{i=1}^m \sum_{j=1}^m \|\mathbf{X}_i - \mathbf{X}_j\| - \\ & \frac{mn}{2(m+n)n^2} \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{Y}_i - \mathbf{Y}_j\|. \end{aligned} \quad (3.27)$$

The hypothesis (2.1) is rejected for large values of T_{BF} .

The critical value and p -value of the statistic are obtainable by either bootstrapping or the permutation method (Baringhaus and Franz, 2001). In this thesis, the test is implemented using the permutation method.

The MATLAB routine `Baringhaus_Franz_perm_test.m`, in the directory `Statistics` on the accompanying CD, was used for the implementation of the test.

3.3.6 The Weiss Statistic

The work of Weiss (1960) is similar to the approaches of Henze (1988) and Schilling (1986). Weiss (1960) used interpoint distances to construct non-overlapping spheres around observations of one sample and their nearest neighbour from the same sample. The test statistic is the number of spheres which contain no observations from the other sample. Few theoretical properties of the statistic are known.

The procedure for computing the statistic is as follows:

- (i) For each observation \mathbf{X}_i , calculate the Euclidean distance

$$R_i = \frac{1}{2} \min_{i \neq j} \{ \|\mathbf{X}_i - \mathbf{X}_j\|, \dots, \|\mathbf{X}_i - \mathbf{X}_m\| \}, \quad i = 1, \dots, m.$$

- (ii) Denote by S_i the number of $\mathbf{Y}_k \in \{\mathbf{Y}_1, \dots, \mathbf{Y}_n\}$ which are contained in the open sphere

$$\{\mathbf{u} : \|\mathbf{u} - \mathbf{X}_i\| < R_i\},$$

i.e. the number of \mathbf{Y}_k lying completely inside the sphere of radius R_i centered on \mathbf{X}_i .

(iii) For a non-negative integer r , define the indicator function

$$I_i(r) = \begin{cases} 1 & , \text{ if } S_i = r \\ 0 & , \text{ otherwise.} \end{cases}$$

(iv) Taking the case where $r = 0$, the test statistic is given by

$$T_m(0) = \frac{1}{m} \sum_{i=1}^m I_i(0).$$

When sample \mathbf{X} is different from sample \mathbf{Y} , the combined sample is not properly mixed. Many observations in \mathbf{X} are isolated from the observations in \mathbf{Y} and, as a result a large value of the test statistic $T_m(0)$ is expected. Therefore, the test is for large values of the statistic $T_m(0)$. If the role of \mathbf{X} and \mathbf{Y} are interchanged, the test statistic is denoted by $T_n(0)$. Asymptotically, the null hypothesis (2.1) is rejected if

$$\begin{aligned} T_m(0) &> \frac{2^{d\gamma}}{1 + 2^{d\gamma}} \text{ or} \\ T_n(0) &> \frac{2^{d\frac{1}{\gamma}}}{1 + 2^{d\frac{1}{\gamma}}}, \end{aligned} \tag{3.28}$$

where $\gamma = \frac{m}{n}$ and d is the dimension (Weiss, 1960).

3.3.7 The Cross-Match Test

In a proposed test similar to that of Friedman and Rafsky (1979), Rosenbaum (2005) used the interpoint distances to construct an optimal non-bipartite matching (ONM) of observations from the pooled sample. An ONM is a procedure for matching observations into disjoint pairs that minimizes the total sum of distances within pairs. The number of pairs made up of observations from different samples, known as "cross-matches", is of interest. The number of cross-matches is the test statistic. The test is distribution-free and the null distribution of the test statistic is known for small samples. For large samples, asymptotic normality applies (Rosenbaum, 2005). The power performance of the test statistic was investigated empirically in the univariate case. However, it is unknown whether the test statistic is universally consistent, or has satisfactory power in the multivariate setting because such properties were not ascertained in the study.

The computation of the test statistic proceeds as follows:

- (i) Firstly, the components of the pooled sample \mathbf{Z} are ranked individually from 1 to N . The vector \mathbf{R}_i is the d -tuple of ranks of the components of \mathbf{Z}_i .
- (ii) The distance $D(\mathbf{R}_i, \mathbf{R}_j)$ is defined to be the Mahalanobis distance between vectors \mathbf{R}_i and \mathbf{R}_j i.e.

$$D(\mathbf{R}_i, \mathbf{R}_j) = (\mathbf{R}_i - \mathbf{R}_j)^T \mathbf{S}_{\mathbf{R}}^{-1} (\mathbf{R}_i - \mathbf{R}_j) \quad i < j,$$

where $\mathbf{S}_{\mathbf{R}}$ is the sample covariance matrix of the ranks. Clearly, there are $\binom{N}{2}$ distinct pairwise distances $D(\mathbf{R}_i, \mathbf{R}_j)$.

- (iii) Using the $\binom{N}{2}$ interpoint distances, construct an ONM of the observations (Rosenbaum, 2005). The procedure requires that N is an even integer. If N is an odd integer, an $(N + 1)$ th pseudo-observation is created with distances $D(\mathbf{R}_i, \mathbf{R}_{N+1}) = 0$ for $i = 1, \dots, N$. Construct an ONM with $N + 1$ observations, and discard the pair containing the pseudo-observation.
- (iv) To define the cross-match test statistic, let $T_k^{\mathbf{X}}$ be the number of pairs with k observations from sample \mathbf{X} , $k = 0, 1, 2$. Interchanging the role of \mathbf{X} and \mathbf{Y} , $T_0^{\mathbf{X}} = T_2^{\mathbf{Y}}$, $T_1^{\mathbf{X}} = T_1^{\mathbf{Y}}$ and $T_2^{\mathbf{X}} = T_0^{\mathbf{Y}}$ all hold. The number of cross-matches $T_1^{\mathbf{X}}$, henceforth denoted by T_1 , is invariant and therefore, is taken to be the test statistic (Rosenbaum, 2005).
- (v) If samples \mathbf{X} and \mathbf{Y} are from identical populations, a large number of \mathbf{X}_i are optimally matched to \mathbf{Y}_i . Consequently, small values of T_1 are significant (Rosenbaum, 2005).

Under the null hypothesis, the exact small sample permutation distribution of T_1 is given by

$$P(T_1 = t_1 | \mathbf{Z}) = \frac{2^{t_1} \left(\frac{N}{2}\right)!}{{}^N C_m t_0! t_1! t_2!}.$$

The same distribution is obtained when m and n are interchanged. For sufficiently large samples, Rosenbaum (2005) showed that the asymptotic distribution of T_1 is normal with parameters approximated by

$$E(T_1) = \frac{mn}{N-1} \quad \text{and}$$

$$\text{Var}(T_1) = \frac{2mn(m-1)(n-1)}{(N-3)(N-1)^2}. \quad (3.29)$$

3.3.8 Notes

- (i) Multivariate two-sample tests described in this chapter can be used to perform goodness-of-fit tests. To perform the goodness-of-fit test, given a sample \mathbf{X} , a Monte Carlo sample \mathbf{Y} is drawn from the specified distribution, and the hypothesis of equality of the two multivariate samples is tested. Of course, the size of sample \mathbf{Y} would need to be generally large, which may render such an approach cumbersome in practical applications.
- (ii) Most of the test statistics described above are based on the distances between observations. Changing a measure of distance between sample points can potentially influence the value of the test statistic and therefore the result of the test. The Euclidean distance was used because it is invariant to orthogonal and some affine transformations. The power against some specific alternatives may be affected by the choice of the distance metric. All test statistics discussed, except the SKS and the cross-match statistics, satisfy the invariance property under orthogonal and some affine transformations. The SKS statistic is invariant to transformations which preserve the ordering of the sample points in d -space, for example, componentwise standardization and scaling. The invariance of the cross-match statistic is with respect to transformations that preserve componentwise rankings of the observations and the Mahalanobis interpoint distance of the ranks.
- (iii) The Hall-Tajvidi statistics T_{HT} and S_{HT} allow for choices of the exponents γ and weights $w_k(j)$. Empirical studies of the two statistics, for many combinations of exponents and weights, have shown that there are only minor differences in power properties of the different versions of the statistics. Thus, the simplest versions, with constant weights $w_k(j) = 1$ and exponent $\gamma = 1.0$, can be used confidently with minimal loss of power (Hall and Tajvidi, 2002). In the power studies reported in Chapter 4, the simplest versions of the statistics T_{HT} and S_{HT} were used.
- (iv) Excepting the test statistic by Weiss (1960), all the test statistics discussed are symmetric - they give the same value when the roles of \mathbf{X} and \mathbf{Y} are

interchanged. Clearly, generally $T_m(0) \neq T_n(0)$ in (3.28). To remove the lack of symmetry, Weiss (1960) suggested using the mean of $T_m(0)$ and $T_n(0)$ as a test statistic.

- (v) Only seven of the nine multivariate two-sample tests discussed above were considered in the power studies. The test proposed by Weiss (1960) has some unknown theoretical properties while the complexity of implementation of the statistic by Rosenbaum (2005) prompted its omission.
- (vi) Results from initial simulations suggested that the powers of the statistics T_F and T_K are similar, as are powers of T_F^{NN} and T_K^{NN} . Therefore, in subsequent chapters, only the powers of T_K and T_K^{NN} are reported, for IPDD and NNDD tests respectively, in all power studies where the similarity in power was observed.



Chapter 4

Power Studies

This chapter gives a discussion of the power performance of some of the multivariate two-sample test statistics described in Chapter 3. Various conditions which were used in the reported power studies are discussed. The results are reported in Section 4.4. Additionally, a graphical exploration for compatibility between two multivariate samples based on the DD-plots is included in Section 4.5. The MATLAB routines used in the power analyses are given in the directory **Power Studies**, on the accompanying CD.

A major concern in application of the proposed multivariate two-sample tests investigated in this study is the limited information on their performance. In some studies, this concern was addressed in a limited fashion. For example, Baringhaus and Franz (2001) studied the power of the test statistic T_{BF} under various conditions, including dimensionality and distributional characteristics like location and scale. The authors considered sampling from multivariate normal as well as non-normal populations, including the multivariate logistic population. They compared the power of the statistic T_{BF} to other test statistics in both the univariate and multivariate settings. Their results suggest that the power of T_{BF} is very close to Hotelling's T^2 statistic for multivariate normal location alternatives, and has considerably more power than the statistic $T_{H(K)}$ of Henze (1988) for multivariate logistic alternatives. Friedman and Rafsky (1979) investigated the power of the T_{FR} relative to parametric competitors. They studied the sensitivity of T_{FR} , among other statistics, to the combination of dimensionality and distributional characteristics for multivariate normal samples. The results revealed that the power of T_{FR} generally

improves when more than one minimum spanning tree (preferably three or more) are used in high dimensions, when compared to other parametric and non-parametric competitors, for both multivariate normal location and scale alternatives. Schilling (1986) conducted a power study based on a combination of dimensionality, distributional characteristics and the number of nearest neighbours. Schilling (1986) sampled from the multivariate normal population. Information about the power of $T_{H(K)}$ can be deduced from his results. The powers of T_{HT} and S_{HT} were studied by Hall and Tajvidi (2002) for bivariate normal scale alternatives and choice of distance metric [see (3.10) to (3.12)]. The results of the study suggest that S_{HT} has slightly better power than T_{HT} when variables are correlated ($\rho = 0.5$) (Hall and Tajvidi, 2002).

The power of the test statistics depends not only on the factors such as dimensionality, type of parent distribution, and sample differences, but also on the sample sizes and level of significance. With increased sample sizes (m and n), test statistics will detect differences between two samples with higher probability. In other words, the test statistics have power approaching one when the sample sizes are increased, a property known as consistency. This property holds for all test statistics described in Chapter 3. The difficulty regarding the large sample size required to attain good power was emphasized by Schmidt (1996). Schmidt (1996) suggested that scientific inquiry can be retarded because many worthwhile research projects cannot be conducted, since the sample sizes required to achieve adequate power of some test statistics may be difficult, if not impossible, to attain. The power problem of the test statistics can be ameliorated by capitalizing on the fact that some statistics are more powerful in detecting specific deviations from the null. As a result, an investigation into the powers of the test statistics described in the preceding chapter is worthwhile.

4.1 Sampled Populations

There are many distributions that are of practical interest. The selection of three distributions for this study at least reflects some variety of properties of distributions. The power of some of the test statistics described in Chapter 3 was studied for samples drawn from three bivariate populations. The populations sampled were:

bivariate normal; bivariate uniform; and bivariate exponential. The populations are discussed in the following sections.

4.1.1 Bivariate Normal Population

The bivariate normal population is a symmetric and mesokurtic distribution. Location, scale, and correlation alternatives were considered. In all studies for bivariate normal populations, sample \mathbf{X} was drawn from the standard bivariate normal distribution $\text{BVN}(\mathbf{0}, \mathbf{I})$. For location alternatives, sample \mathbf{Y} was drawn from the bivariate normal distribution $\text{BVN}(\mu, \mathbf{I})$ with mean vector $\mu = \begin{pmatrix} \Delta \\ 0 \end{pmatrix}$, where Δ ranged over the interval $[0, 2]$. For the scale alternatives, sample \mathbf{Y} was drawn from the $\text{BVN}(\mathbf{0}, \Sigma_s)$ with covariance matrix Σ_s of the form $\begin{pmatrix} \sigma & 0 \\ 0 & 1 \end{pmatrix}$, where σ was varied from 1 to 6.5. In the case of the correlation alternatives, sample \mathbf{Y} were drawn from the $\text{BVN}(\mathbf{0}, \Sigma_C)$ with covariance matrix Σ_C of the form $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where ρ was ranged over the interval $[0, 0.99]$. The results are reported in Section 4.4.3.

4.1.2 Bivariate Uniform Population

The bivariate uniform population is a symmetric but highly platykurtic distribution. The standard bivariate uniform distribution is one with observations uniformly distributed in the unit square and is denoted by $\text{BVU}[0, 1]$. The power performance of some of the test statistics was studied for location, scale, and correlation alternatives, as in the bivariate normal case. In all simulation experiments of bivariate uniform population, sample \mathbf{X} was drawn from the standard bivariate uniform distribution $\text{BVU}[0, 1]$. For location alternatives, sample \mathbf{Y} was drawn from the bivariate uniform distribution with the mean vector shifted by $\begin{pmatrix} \Delta \\ 0 \end{pmatrix}$. The parameter Δ was varied from 0 to 0.5. In the case of the scale and correlation alternatives, \mathbf{Y} was sampled from bivariate uniform populations differing from $\text{BVU}[0, 1]$ only by covariance matrices. The conditions used for the scale and correlation alternatives are the identical to those used for the similar alternatives in the bivariate normal cases. Results are presented in Section 4.4.4.

4.1.3 Bivariate Exponential Population

A highly skewed population which was studied, is the bivariate exponential population. Scale/location and correlation alternatives were considered. In both cases, sample \mathbf{X} was sampled from the standard bivariate exponential distribution BVE($\mathbf{1}$), with independent marginals and marginal means $\lambda_1 = \lambda_2 = 1$. Sample \mathbf{Y} was drawn from the BVE distribution with covariance matrix Σ_s of the form $\begin{pmatrix} \sigma & 0 \\ 0 & 1 \end{pmatrix}$, where σ was varied from 1 to 6.5, for location/scale alternatives. Correlation alternatives were dealt with as for the bivariate normal case. The results are reported in Section 4.4.5.

4.2 Estimation of Power for Finite Samples

The power of a multivariate two-sample test statistic is the probability of rejecting the null hypothesis (2.1), given that it is false. For complex non-parametric multivariate two-sample test statistics studied, published tables or commercial software (e.g. SAS, S-Plus, SPSS) are not available like there is for most univariate parametric tests. In this case, Monte Carlo simulations provide a very useful way of estimating power. In the simulation experiments, the simulated samples \mathbf{X} and \mathbf{Y} were generated independently.

The algorithm for estimating power of any test statistic numerically is as follows:

1. Simulate a sample, \mathbf{X} of size m according to a standard multivariate distribution F , and a sample \mathbf{Y} of size n according to a specified multivariate distribution G ;
2. Calculate the multivariate two-sample test statistic;
3. If the test statistic is statistically significant at the pre-specified α -level, the result is noted;
4. Return to step one and repeat the procedure a large number of times W .

The estimated power \hat{h} , the probability of a statistically significant result, is obtained by computing the proportion of the runs (replicates) which produced significant

results:

$$\hat{h} = \frac{\text{Number of times } H_0 \text{ is rejected at } \alpha\text{-level in } W \text{ replications}}{\text{Total number of replications } (W)}. \quad (4.1)$$

By sampling theory, \hat{h} is a binomial random variable. Therefore, for sufficiently large W , the distribution of \hat{h} is approximately normal with mean h and standard deviation (also known as the standard error of the proportions) of

$$\begin{aligned} \sigma_{\hat{h}} &\approx \sqrt{\frac{h(1-h)}{W}}, \\ &\approx \sqrt{\frac{\hat{h}(1-\hat{h})}{W}}. \end{aligned} \quad (4.2)$$

Equation (4.2) indicates explicitly the dependence of the estimated power on the number of replicates W . Thus, for a better approximation of the power of a test statistic, the number of replicates must be sufficiently large.

4.3 Computational Details

Power simulations were done using the MATLAB software package on a 3 Giga-Hertz Pentium 4 computer. Every point in the parameter range considered, for all the alternatives reported in the subsequent sections, represents a specific number of replicates W and permutations B . Due to considerations of computing time and computing resources, power was approximated for a fairly small number of replications $W = 500$ and permutations $B = 500$, and few points in the parameter ranges were chosen. For the ranges of location, scale and correlation, ten equally spaced points were used, hence the non-smooth nature of the reported power functions appearing below. Therefore, under the conditions for which power studies were conducted, the tests can be arranged in the following ascending order of computational times: Baringhaus-Franz, SKS, IPDD, NNDD, Henze, Hall-Tajvidi, and Friedman-Rafsky tests. In general, the computational time for each set of results in Figures 4.9 to 4.17 was approximately 6 days.

The multivariate two-sample tests studied are intensive computationally because of the nature of the algorithms required to compute the statistics, for example, the

Friedman-Rafsky and Hall-Tajvidi statistics. If the computer implementation is not efficient computationally, the demand is exacerbated further. The implementation of the tests was done by the permutation method except for the Friedman-Rafsky statistic T_{FR} for which the asymptotic result was used. The latter strategy was supported by results of trial simulations.

4.4 Power Comparisons

Most of the power studies of multivariate two-sample tests reported in the literature concentrated on null hypotheses defined by the standard multivariate normal distribution against location or scale alternatives or both. The Friedman-Rafsky and Hall-Tajvidi tests are examples. In this thesis, the power performance of some multivariate two-sample statistics discussed in Chapter 3 viz. Baringhaus-Franz T_{BF} , Friedman-Rafsky T_{FR} , Hall-Tajvidi T_{HT} and S_{HT} , Henze $T_{H(K)}$, IPDD statistics T_F and T_K , NNDD statistics T_F^{NN} and T_K^{NN} , and SKS statistic T_{SKS} , are studied for various alternative distributions discussed in Section 4.1. Figures of the power functions, in different colours, are given in `Power_Figures.pdf` in the directory `Power Studies`, on the accompanying CD.

4.4.1 Estimates of Type I Error Rates

Under the null hypothesis, the power \hat{h} must be equal to the nominal significance level α (Thas, 2001). In this study, some statistical tests investigated are implemented with approximate critical values and significance level α . The implication is that the exact p -values were replaced with values approximated from the samples by the permutation method. As a result, power evaluated under the null hypothesis may be slightly different from the nominal significance level α for some test statistics. The lack of conformity of the approximated power \hat{h} to α under the null hypothesis is known as the *bias* of a test statistic with respect to the given nominal significance level α (Greenberg, 2006). Different test statistics deviate differently from the nominal significance level α , that is, some test statistics underestimate while others overestimate the power under the null hypothesis. Bias is caused by several factors viz. sample size, number of permutations and number of replications among others. The bias is significantly reduced by using large sample sizes m and n , and a sufficiently large number of permutations B as well as replications W (Thas, 2001).

To check for accuracy of type I error probabilities for the studied test statistics, simulations were done for samples of sizes $m = 60$ and $n = 50$ with $W = 2000$ replications. The nominal significance levels used were 0.010, 0.050, and 0.100. The p -values were approximated by using the empirical results of $B = 500$ permutations, conditioned on the pooled samples. The estimate of type I error probabilities are proportions of the 2000 replicates which were declared significant at the indicated nominal significance level. Table 4.1 shows empirical levels of all the test statistics for the populations studied. The results generally indicate good approximations to nominal significance levels. Approximations for nominal level 0.050 seem equally good across all test statistics in Table 4.1. On the whole, the deviations of the estimated probabilities from nominal values are satisfactorily small.

Table 4.1: Estimates of type I error probabilities

Statistics	Bivariate Normal			Bivariate Exponential			Bivariate Uniform		
	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$
T _{BF}	0.005	0.043	0.097	0.010	0.049	0.098	0.007	0.049	0.101
T _{FR}	0.007	0.052	0.077	0.008	0.056	0.085	0.007	0.047	0.075
T _{HT}	0.009	0.044	0.098	0.010	0.049	0.098	0.008	0.050	0.108
S _{HT}	0.009	0.050	0.097	0.010	0.051	0.095	0.009	0.049	0.097
T _{H(4)}	0.016	0.058	0.108	0.017	0.062	0.108	0.012	0.043	0.091
T _F	0.009	0.063	0.109	0.010	0.048	0.092	0.008	0.051	0.096
T _K	0.010	0.062	0.108	0.009	0.048	0.093	0.008	0.050	0.096
T _F ^{NN}	0.013	0.055	0.100	0.006	0.050	0.101	0.012	0.049	0.096
T _K ^{NN}	0.013	0.059	0.105	0.009	0.050	0.105	0.009	0.047	0.095
T _{SKS}	0.007	0.041	0.099	0.010	0.046	0.090	0.013	0.050	0.097

Preliminary empirical studies of the multivariate two-sample tests have shown that sufficient accuracy in estimating α with $W = 500$ replicates is guaranteed for a moderate size of $B = 500$ permutation resamples. Therefore, in the simulation studies of the tests reported subsequently, sample sizes were $m = 60$ and $n = 50$, while the number of replicates W and permutation resamples P were fixed at 500. The power properties of the test statistics were investigated for a nominal significance level $\alpha = 0.050$.

4.4.2 The Parameter K in the Henze statistic $T_{H(K)}$

Henze's statistic $T_{H(K)}$ is a function of K , the number of nearest neighbours taken into account. When the value of K is changed, the statistical properties of $T_{H(K)}$ are significantly influenced. Particularly, the power performance of $T_{H(K)}$ improves with increasing K . However, beyond a certain value, further increase of K produces a diminishing return on the power. Figures 4.1 to 4.8 show power functions of $T_{H(K)}$ at a nominal significance level $\alpha = 0.05$, for all populations. The value of K was range from 1 to 9.

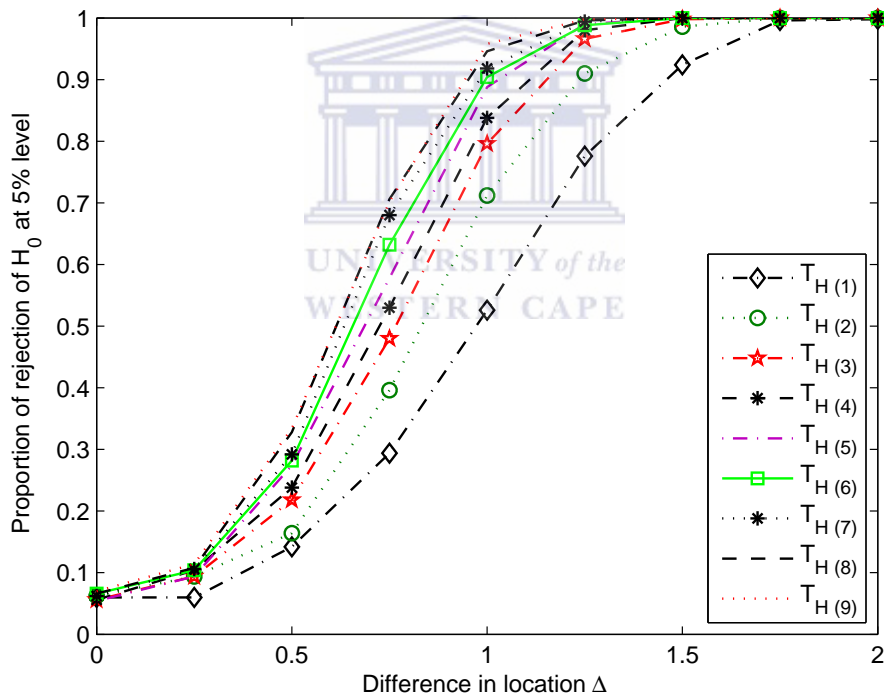


Figure 4.1: Power functions for $T_{H(K)}$ for bivariate normal location alternatives.

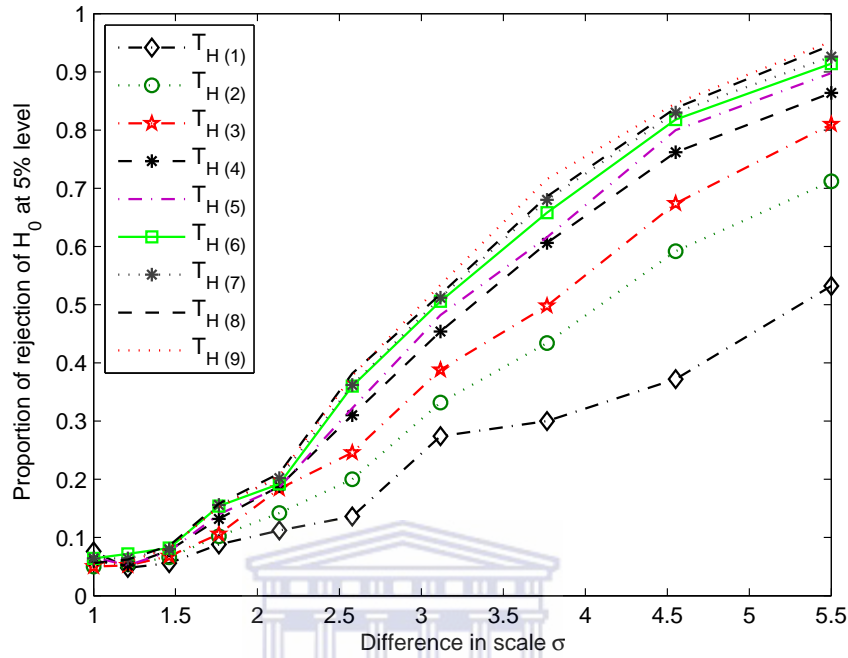


Figure 4.2: Power functions for $T_{H(k)}$ for bivariate normal scale alternatives.

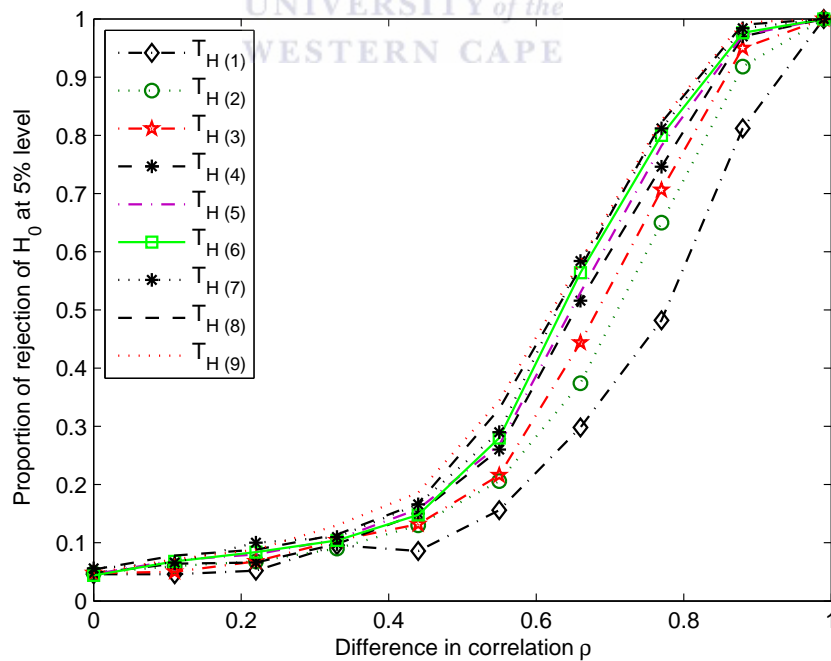


Figure 4.3: Power functions for $T_{H(k)}$ for bivariate normal correlation alternatives.

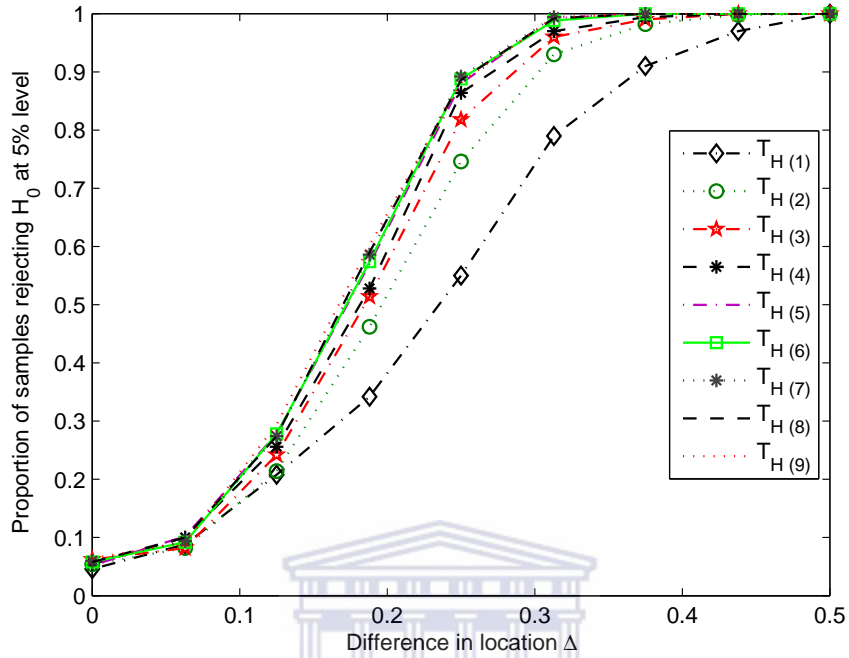


Figure 4.4: Power functions for $T_{H(K)}$ for bivariate uniform location alternatives.

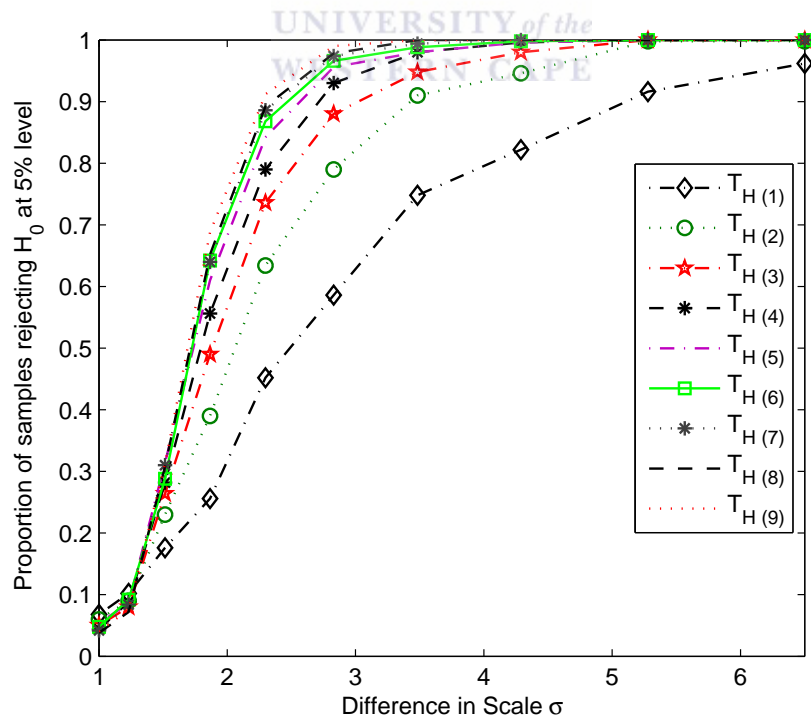


Figure 4.5: Power functions for $T_{H(K)}$ for bivariate uniform scale alternatives.

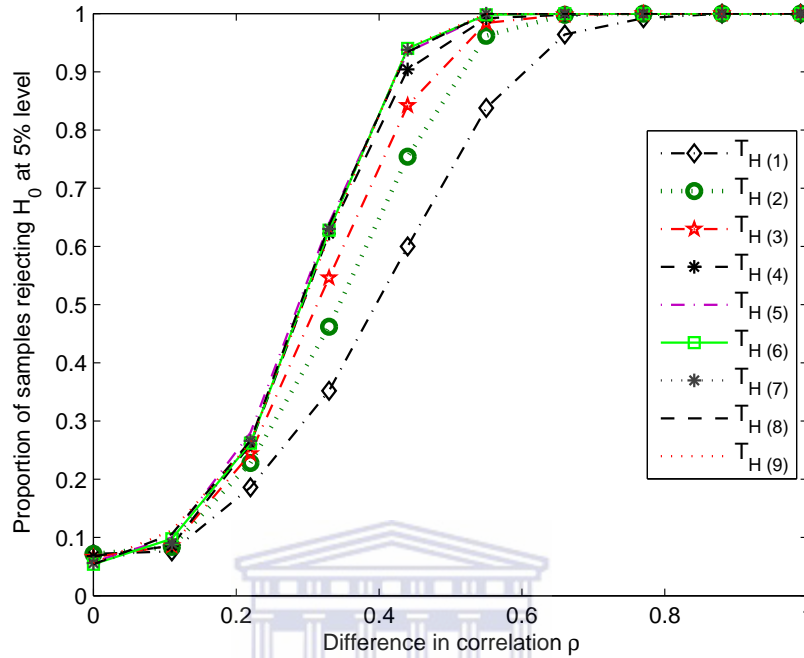


Figure 4.6: Power functions for $T_{H(K)}$ for bivariate uniform correlation alternatives.

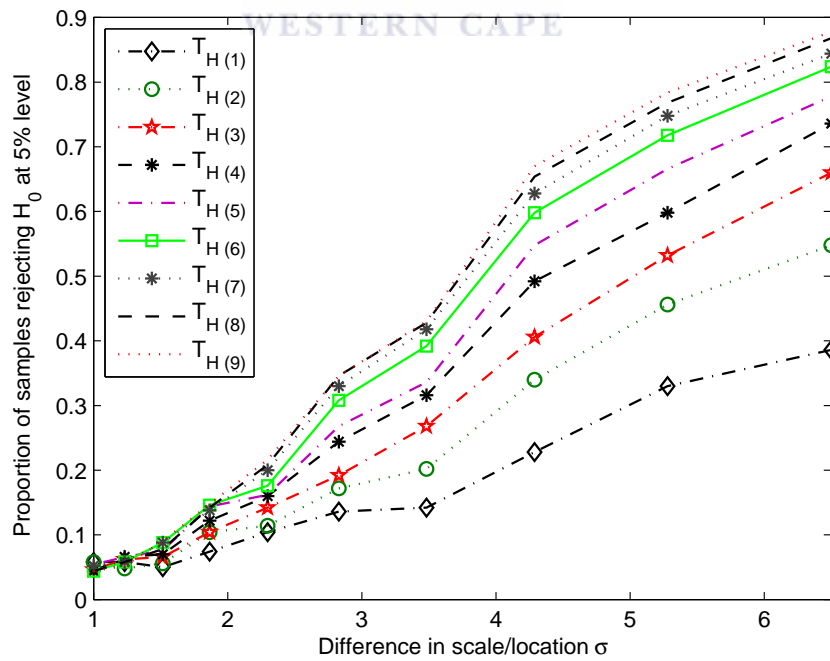


Figure 4.7: Power functions for $T_{H(K)}$ for bivariate exponential scale/location alternatives.

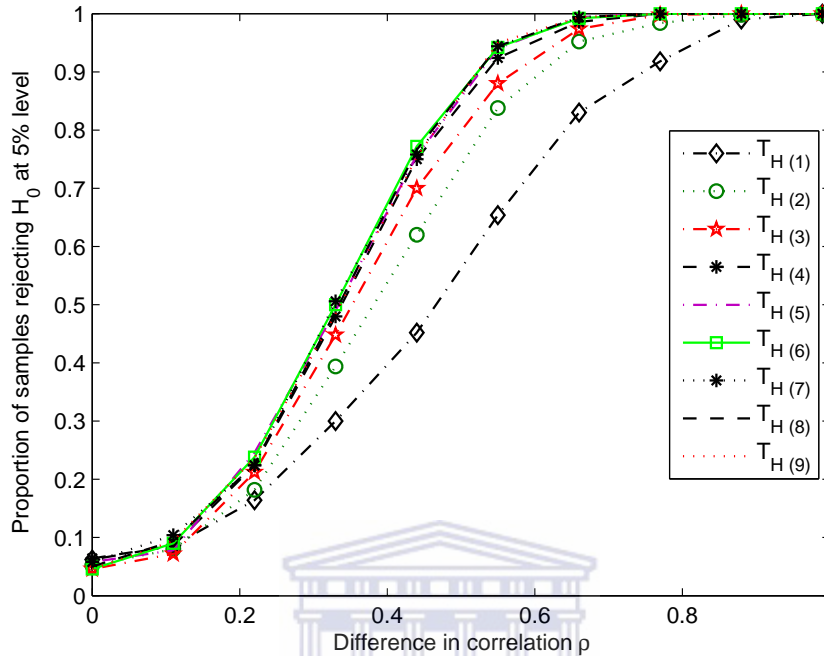


Figure 4.8: Power functions for $T_{H(K)}$ for bivariate exponential correlation alternatives.

Some observations can be made based on the results in Figures 4.1 to 4.8. The number of the nearest neighbours K up to 3 produced test statistics $T_{H(K)}$ with power functions which are clearly distinguishable from each other. Generally, for $K \geq 4$, the power performance increased more slowly with increasing K except for the case of bivariate exponential scale/location alternatives (Figure 4.7). This suggests that when the number of nearest neighbours K is at least 4, the power of the test statistic $T_{H(K)}$ is minimally affected by the increase in K . The results provide a useful guideline when selecting the value used in the power study because no criterion for choosing an optimal value of K is available (Schilling, 1986). Therefore, for simulations reported subsequently, $K = 1$ (for comparison) and $K = 4$ are used.

4.4.3 Bivariate Normal Distribution

This section discusses result of the power studies when the populations sampled are normal differing in locations (Figure 4.9), scale (Figure 4.10), and correlations (Figure 4.12). The test statistics compared are T_{BF} , T_{FR} , T_{HT} , S_{HT} , $T_{H(K)}$, T_K^{NN} , T_K , and T_{SKS} .

Location Differences

Figure 4.9 shows results from the power studies of the bivariate normal location alternatives. Empirical results suggest that when there is a location difference between the two bivariate normal populations, the Baringhaus-Franz statistic T_{BF} performs better than every other test statistic for the whole range of the location shifts. This result is not surprising because T_{BF} is known to be relatively sensitive to location differences between multivariate normal populations (Baringhaus and Franz, 2001). Baringhaus and Franz (2001) showed empirically that T_{BF} compares satisfactorily well to the parametric competitor, Hotelling's T^2 statistic, for a similar setting. As Figure 4.9 shows, the performances of T_{SKS} , T_{HT} and S_{HT} are virtually the same. The statistic $T_{H(4)}$ showed moderate power, while the remainder - particularly T_K^{NN} - performed poorly.

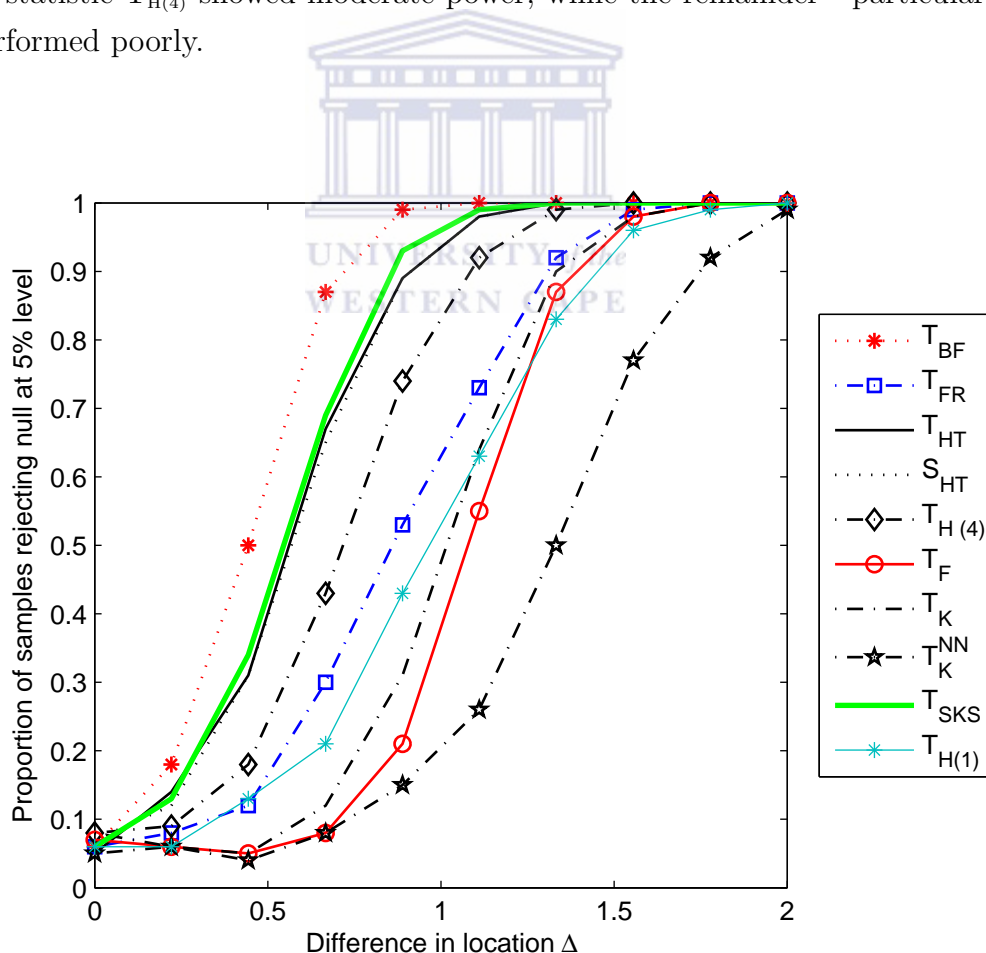


Figure 4.9: Power functions for bivariate normal location alternatives.

Scale Differences

Univariate tests of Kolmogorov-Smirnov type are known to be generally sensitive non-parametric tests for differences in scale (Hall and Tajvidi, 2002). However, the theoretical property is not obviously generalizable to the higher dimensional type of Kolmogorov-Smirnov statistics as attested by the poor performance of T_{SKS} for this setting. The statistic T_K is seen to dominate the other non-parametric statistics for normal scale alternatives. It is not surprising that the statistic on the full set of interpoint distances T_K is much more sensitive to scale differences than the one based only on the nearest neighbour distances, T_K^{NN} . Performances of the T_{HT} , S_{HT} , $T_{H(4)}$ and T_{BF} statistics are similar, with T_{FR} somewhat worse. The performances of the T_{SKS} , $T_{H(1)}$ and T_K^{NN} statistics are poor.

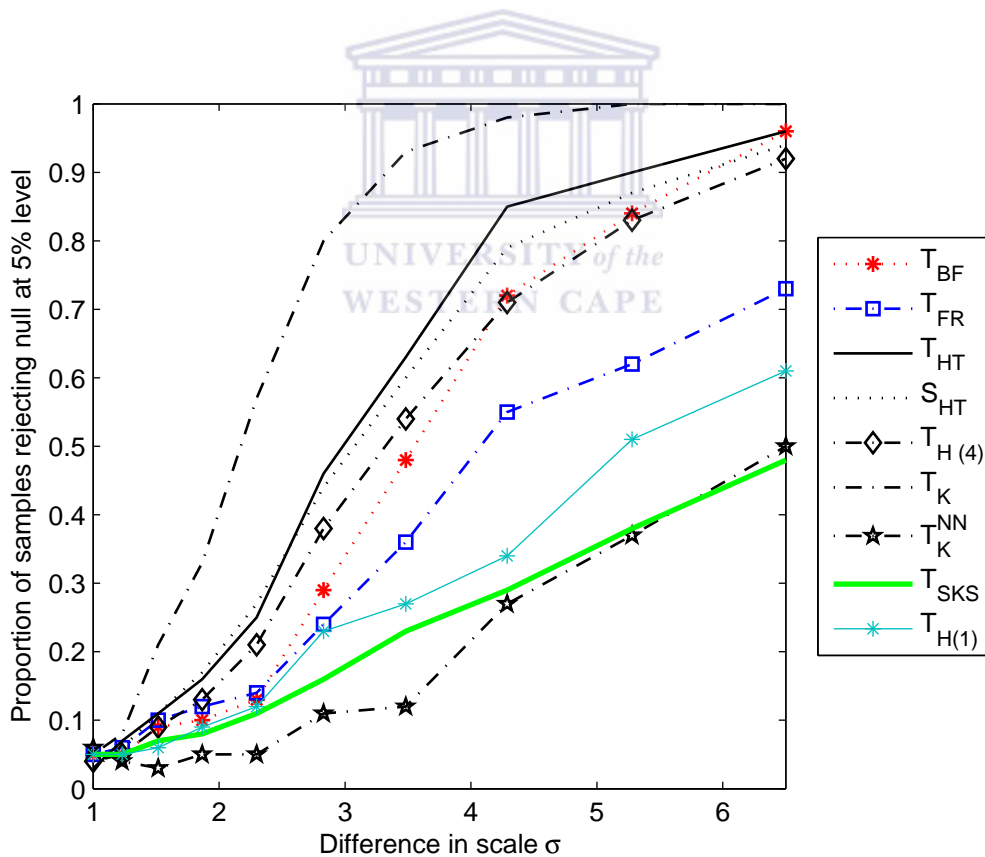


Figure 4.10: Power functions for bivariate normal scale alternatives.

Sometimes, particular deviations from the null hypothesis are not of interest. I digress slightly to consider the case where the two sample means are set equal in

order to eliminate the possibility of a significant result due to different population means (Figure 4.11).

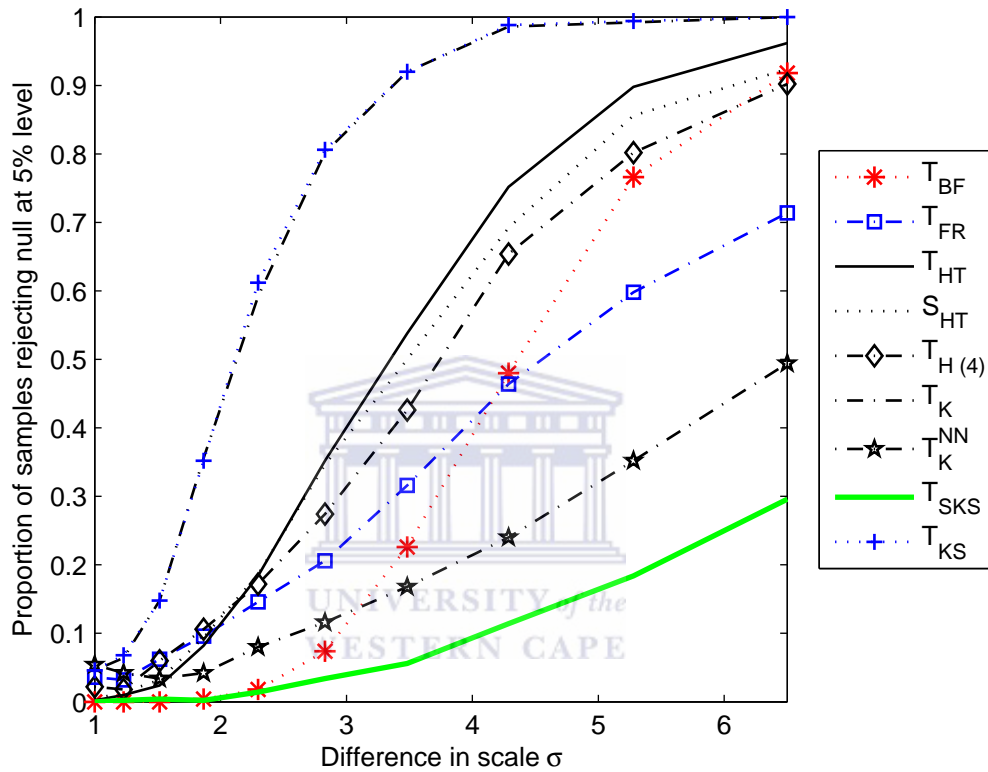


Figure 4.11: Power functions for location-adjusted bivariate normal scale alternatives.

Additional power studies done for the normal scale alternatives, in which the two bivariate normal samples were mean centered, are reported in Figure 4.11. In the case of the IPDD statistics, the distribution of the between-sample distances [see (3.21)] of the adjusted samples was ignored and only the distributions of within-sample distances [see (3.19) and (3.20)] were considered. To assess the equality of the two resulting univariate distributions, the two-sample univariate Kolmogorov-Smirnov test statistic was used. Results indicate that the Kolmogorov-Smirnov statistic is especially sensitive against the scale alternatives. The performance ranking of the statistics is similar to that in Figure 4.10, with the power of the usual Kolmogorov-Smirnov statistic matching that of the statistic T_K . Noticeable in Figure 4.11 is the poor performances of the statistics T_{BF} and T_{SKS} for small scale differences.

Correlation Differences

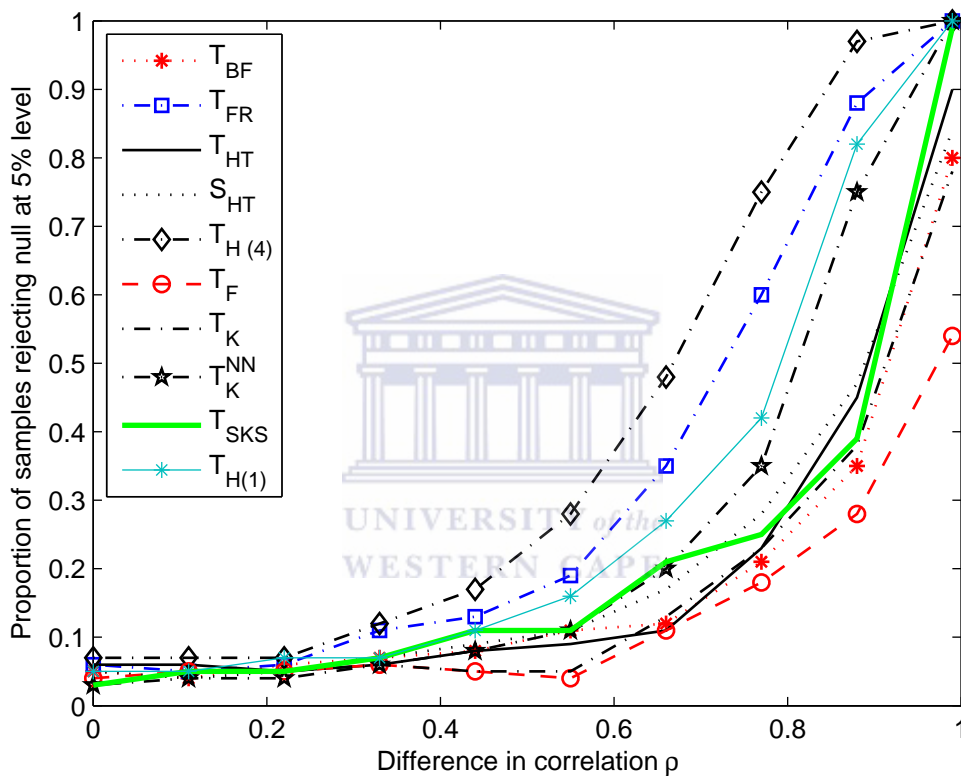


Figure 4.12: Power functions for bivariate normal correlation alternatives.

The results in Figure 4.12 show that the performance of statistics $T_{H(4)}$ and T_{FR} are substantially better than those of the other statistics for the correlation alternatives, with $T_{H(1)}$ and T_K^{NN} next best. A common feature of these four statistics is that they are all based in some way on nearest neighbour distances of the sample observations. Therefore, this suggests that differences in dependence structure could more easily be detected using nearest neighbour based test statistics. It is clear the test statistics T_{FR} and $T_{H(4)}$ are recommended for normal correlation alternatives. Note that the power curves for T_F is shown explicitly, as the power for this IPDD statistic is considerably lower than that of T_K in this instance.

4.4.4 Bivariate Uniform Distribution

Location Differences

The powers of the test statistics are considerably different for sufficiently large location differences (Figure 4.13). Generally, T_{BF} is the most powerful. The statistics T_K^{NN} and T_K are poorest.

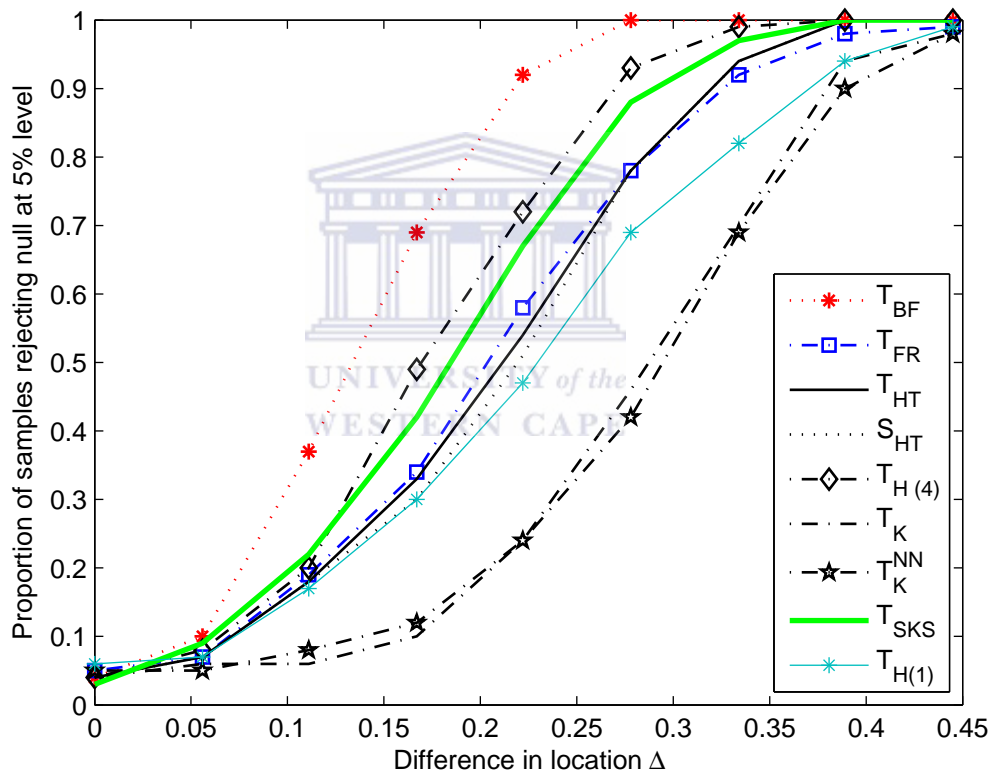


Figure 4.13: Power functions for bivariate uniform location alternatives.

Scale Differences

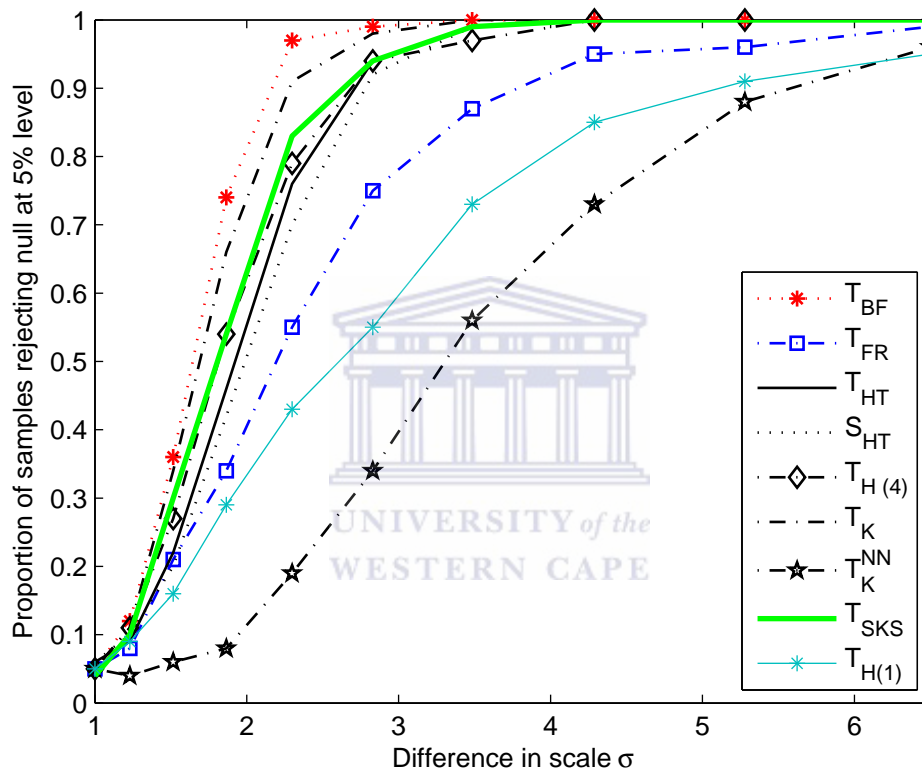
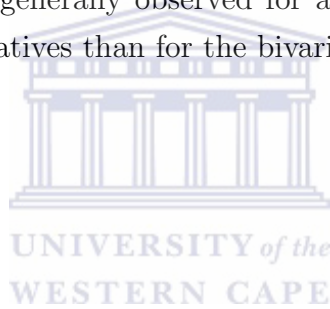


Figure 4.14: Power functions for bivariate uniform scale alternatives.

As is apparent from Figure 4.14, the powers of T_{BF} , T_K , T_{SKS} , T_{HT} , and S_{HT} are high for this setting. For T_{FR} the performance was moderate while that of the statistic T_K^{NN} was clearly the worst. Thus, based on these results, it is clear that all the test statistics except $T_{H(1)}$, T_K^{NN} and T_{FR} , have good power against bivariate uniform scale differences.

Similarities and differences between the results from two scale alternatives in Figures 4.10 and 4.14 are:

- (i) the statistics T_F and T_K performs very well against scale alternatives for both populations;
- (ii) T_K^{NN} had the lowest power against scale alternatives for both populations. The powers of T_{HT} , S_{HT} , $T_{H(4)}$ and T_{BF} are generally high against scale alternatives for both populations;
- (iv) the performance of T_{SKS} is very good against bivariate uniform scale alternatives but poor against bivariate normal scale alternatives;
- (v) higher powers were generally observed for all the statistics against bivariate uniform scale alternatives than for the bivariate normal scale alternatives.



Correlation Differences

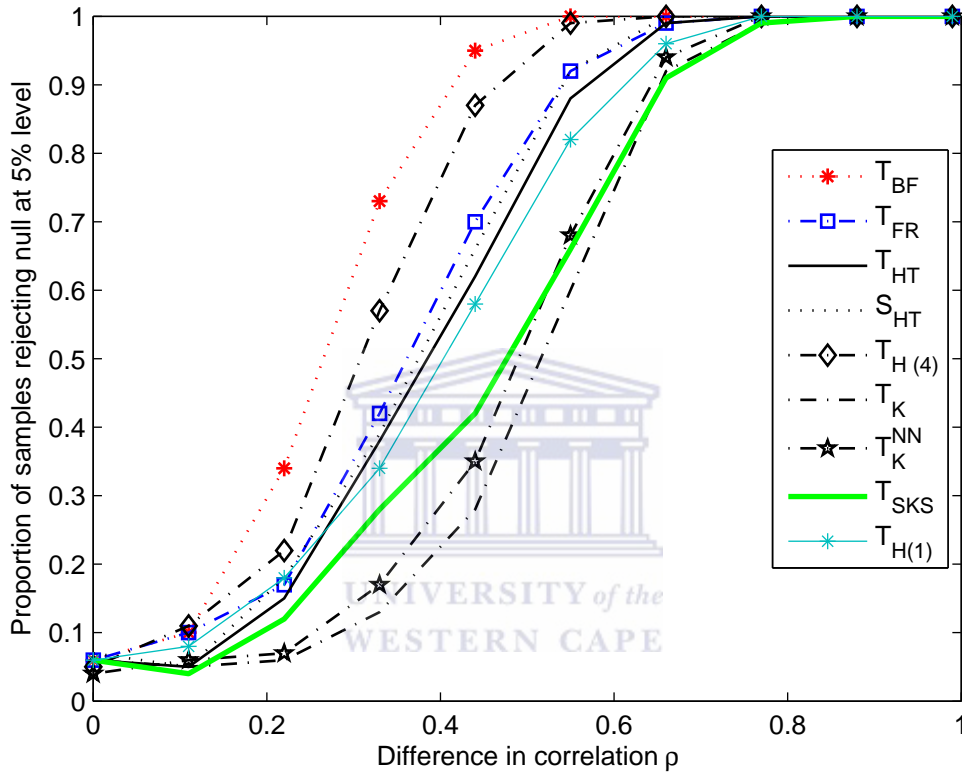


Figure 4.15: Power functions for bivariate uniform correlation alternatives.

The results are shown in Figure 4.15. In general, the power functions show that the performances of T_{BF} and $T_{H(4)}$ were noticeably better and that of T_K was the poorest for this setting.

4.4.5 Bivariate Exponential Distribution

Scale/Location Differences

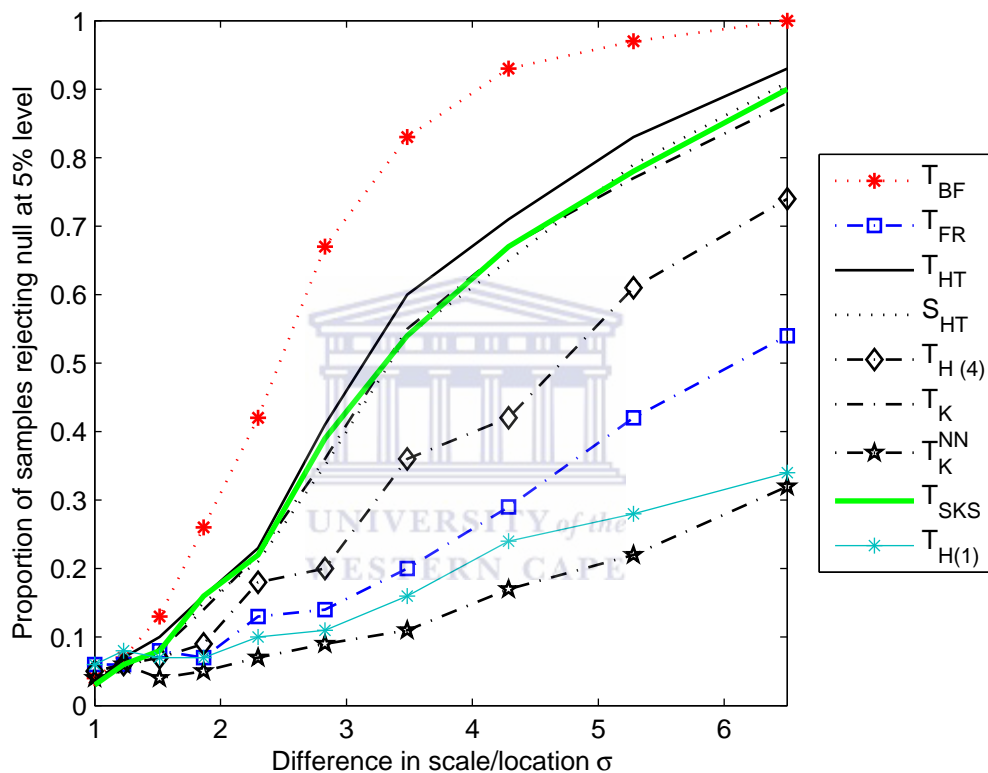


Figure 4.16: Power functions for bivariate exponential scale/location alternatives.

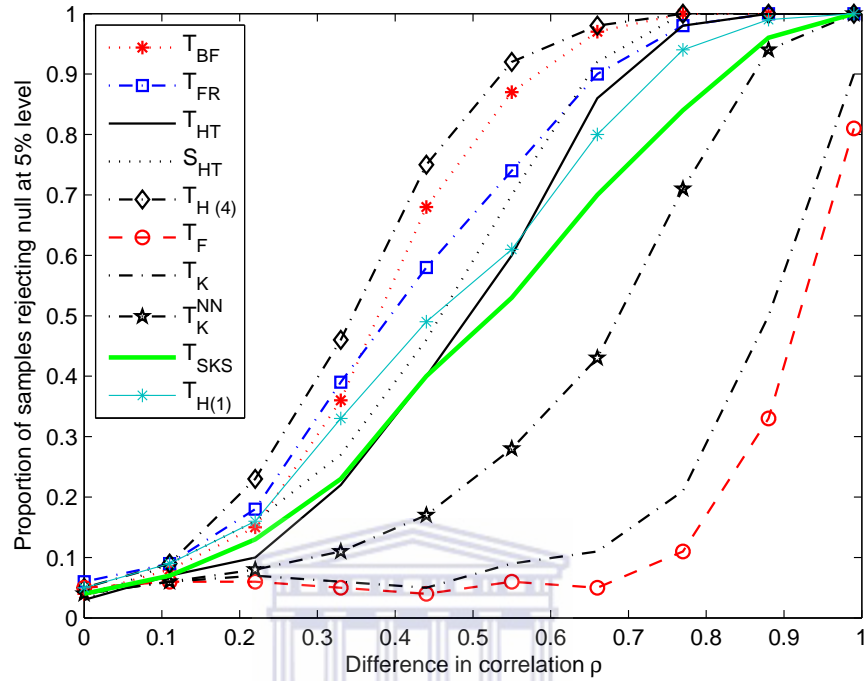
Figure 4.16 shows the results from the power studies for exponential scale/location alternatives. There are large differences in power between statistics. Clearly, T_{BF} has the highest power while $T_{H(1)}$ and T_K^{NN} have the lowest power for this setting. The performances of T_{SKS} , T_K , T_{HT} and S_{HT} are intermediate and very similar. Noticeable in Figure 4.16, is the underperformance of most nearest neighbour based statistics T_{FR} , $T_{H(4)}$, $T_{H(1)}$ and T_K^{NN} for this setting. This implies that nearest neighbour based statistics are not appropriate for scale problems when samples are drawn from highly skewed (exponential) populations.

Major similarities and differences between the result of the exponential scale/location alternatives in Figure 4.16 and those observed for location alternatives of the normal (Figure 4.9) and uniform (Figure 4.13) populations are:

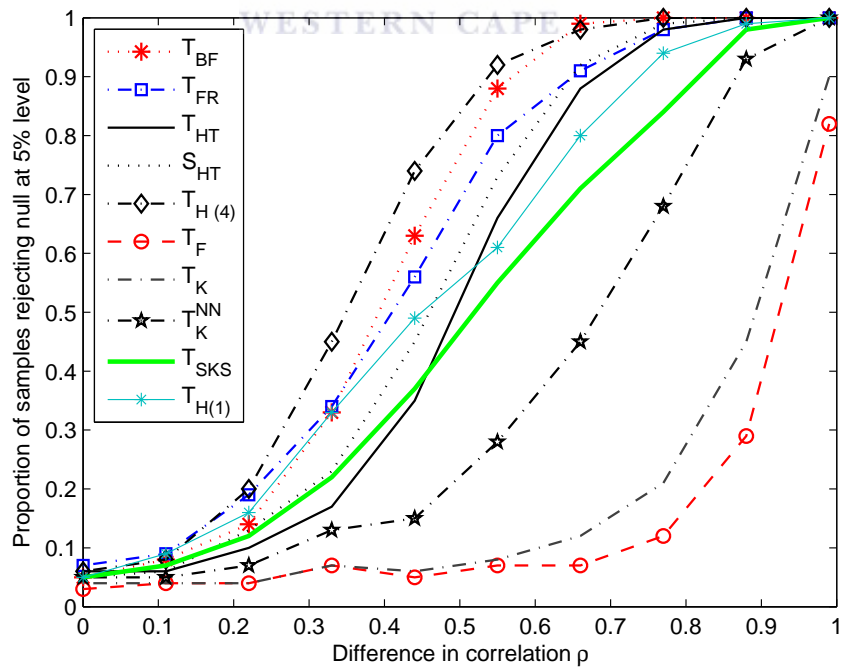
- (i) T_{BF} is the most powerful statistic against location alternatives for all three populations;
- (ii) the power of T_K is similar to that of T_K^{NN} for bivariate uniform location alternatives but very different for the case of the bivariate normal location alternatives and bivariate exponential scale/location alternatives. The latter performed particularly poorly against location alternatives across all populations;
- (iii) the powers of T_{SKS} , T_{HT} and S_{HT} are high against bivariate normal location alternatives and bivariate exponential scale/location alternatives but mediocre against the bivariate uniform location alternatives;
- (iv) $T_{H(4)}$ performed very well against bivariate normal and bivariate uniform location alternatives, but only moderately against bivariate exponential scale/location alternatives.

Correlation Differences

Figures 4.17 (a) and (b) show estimated power functions of all test statistics when populations sampled are bivariate exponential differing in dependence structures. The results in Figure 4.17 (b) were obtained to verify those in (a) as the curve shapes for several of the statistics (particularly T_K , T_F and $T_{H(1)}$) appear unusual. The powers of most test statistics are generally high with $T_{H(4)}$ best for this setting. Clearly, over the whole range of the correlation, T_F and T_K have far less power than other test statistics. There are several examples of changes in power of statistics with changes in the magnitude of correlations (for example T_{FR} and T_{BF} ; T_{SKS} and T_{HT} ; T_{HT} and $T_{H(1)}$).



(a)



(b)

Figure 4.17: Power functions for bivariate exponential correlation alternatives.

Some noticeable differences and similarities among the results for correlation differences in Figures 4.17, 4.12 and 4.15 are:

- (i) the powers of $T_{H(4)}$ and T_{BF} were generally the highest in detecting correlation differences, but the latter performed very poorly against bivariate normal correlation alternatives;
- (ii) the performances of T_F and T_K were the poorest against correlation differences across all the populations;
- (iii) generally, the powers of all the statistics were high against the bivariate uniform correlation alternatives;
- (iv) the powers of all the statistics are generally low against bivariate normal correlation alternatives. The performances are very similar for small differences in correlation.

4.4.6 General Discussion and Recommendations

It is not possible to recommend a particular multivariate two-sample test statistic as having the highest power in all instances discussed. However, some test statistics were shown to have good power against specific types of alternatives for all populations. Therefore, based on the results from the power studies, recommendations about the power of the test statistics against specific departures from the null hypothesis are made with regard to the type alternatives:

- (i) The powers of statistics T_{BF} , T_{SKS} , $T_{H(4)}$, T_{HT} and S_{HT} were generally high against location alternatives. This is true regardless of the distribution sampled. These test statistics showed robustness to distributional geometry. The statistic T_{BF} by Baringhaus and Franz (2001) should be preferred to other statistics for location-shift problems.
- (ii) The statistics T_K , T_{BF} , $T_{H(4)}$, T_{HT} and T_{SK} were shown to generally be powerful for scale alternatives. However, the power properties of the statistics exhibited dependence on the distributional geometry of the sampled populations. The power of T_{BF} was low for samples from the bivariate normal distribution. Overall, the statistic T_K is good across all populations and therefore should be given preference for scale problems.

- (iii) The statistics $T_{H(4)}$, T_{FR} and $T_{H(1)}$ were generally powerful for the correlation alternatives. T_{BF} performed well for uniform and exponential distributions but very poorly for normal distributions. Particularly, $T_{H(4)}$ and T_{FR} are considerably robust to all the populations investigated and should therefore be preferred to other test statistics for correlation problems.

4.5 The Depth-Depth Plots

The concept of data-depth has been used for various multivariate analysis techniques, among them multivariate comparison, multivariate classification and multivariate outlier detection. In this section, multivariate comparisons of two distributions based on the data-depth metric are discussed. The technique is illustrated via the DD-plot. The Mahalanobis depth was used to quantify the depth of the sample points. DD-plots, which show the depth values of the pooled sample relative to the two sample centroids are reported. Distributional characteristics studied include location and scale (for studies of other characteristics see Liu, Parelius and Singh (1999) and references therein). Sample sizes of $m = 100$ and $n = 100$ were used in all cases.

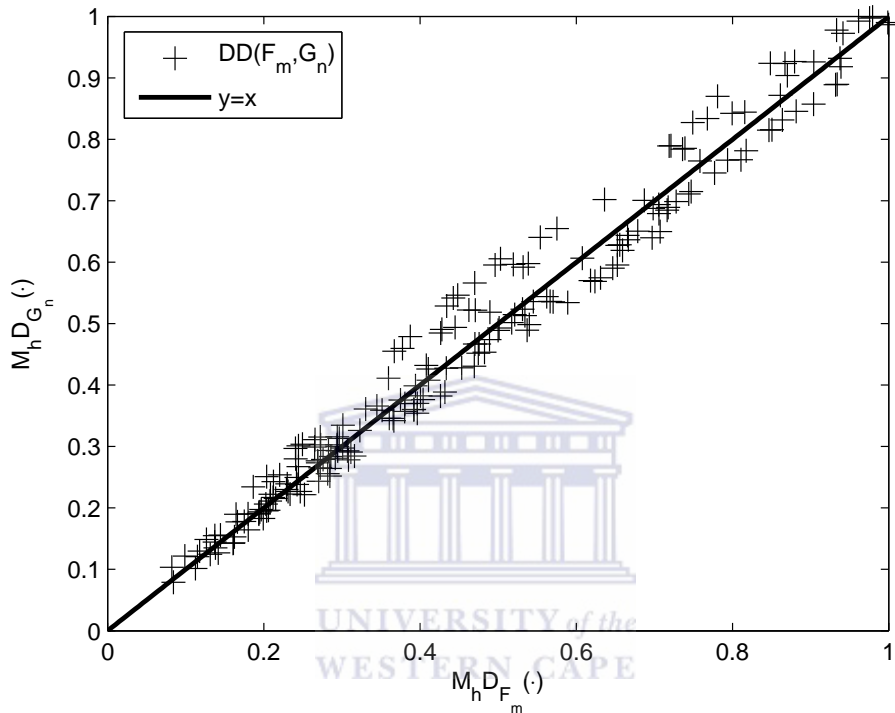


Figure 4.18: DD-plot for identical distributions.

If the null hypothesis (2.1) is true, the DD-plot defined in Section 3.1 should be clustered along the line $y = x$, as Figure 4.18 shows. The two samples were drawn from the standard bivariate normal population $\text{BVN}(\mathbf{0}, \mathbf{I})$. This pattern is expected irrespective of the sampled population (Liu, Parelius and Singh, 1999).

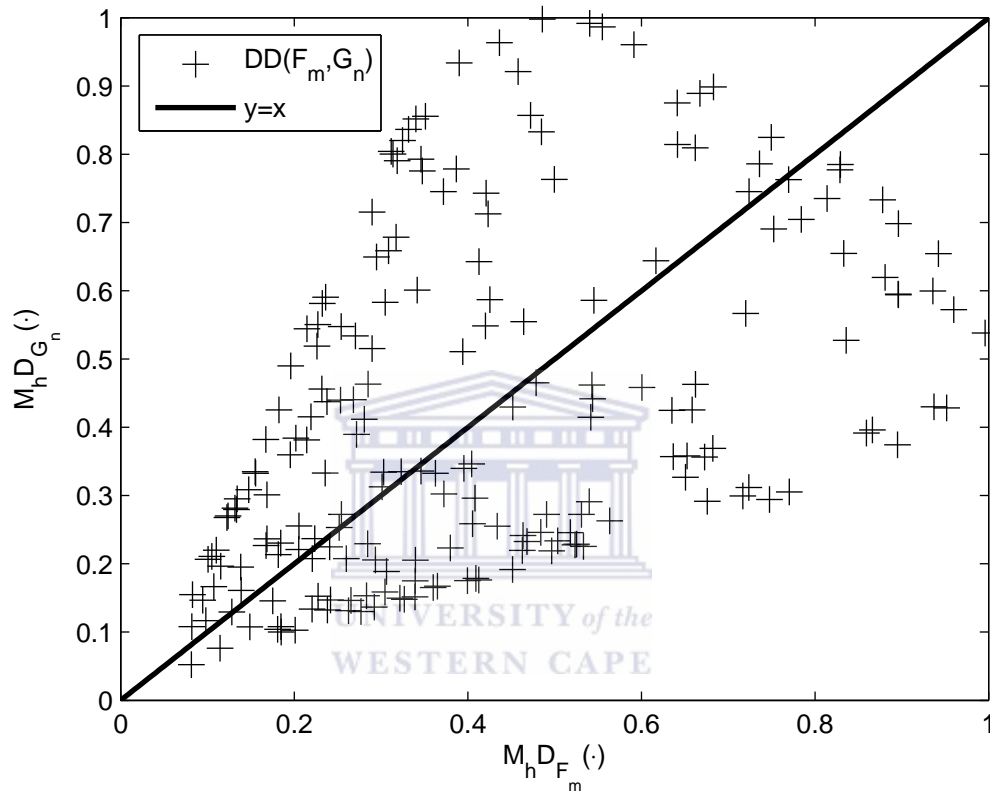


Figure 4.19: DD-plot for distributions with location difference.

Figure 4.19 shows the DD-plot with one sample from $\text{BVN}(\mathbf{0}, \mathbf{I})$ and the other sample from $\text{BVN}(\mu, \mathbf{I})$ with the location shifted to $\mu = (1, 0)^T$. In this case, the DD-plot shows an obvious deviation from the line $y = x$, in a symmetric fashion as if the DD-plot were a scatter plot. The pattern of departure from linearity characterizes the location difference (Liu, Parelius and Singh, 1999).

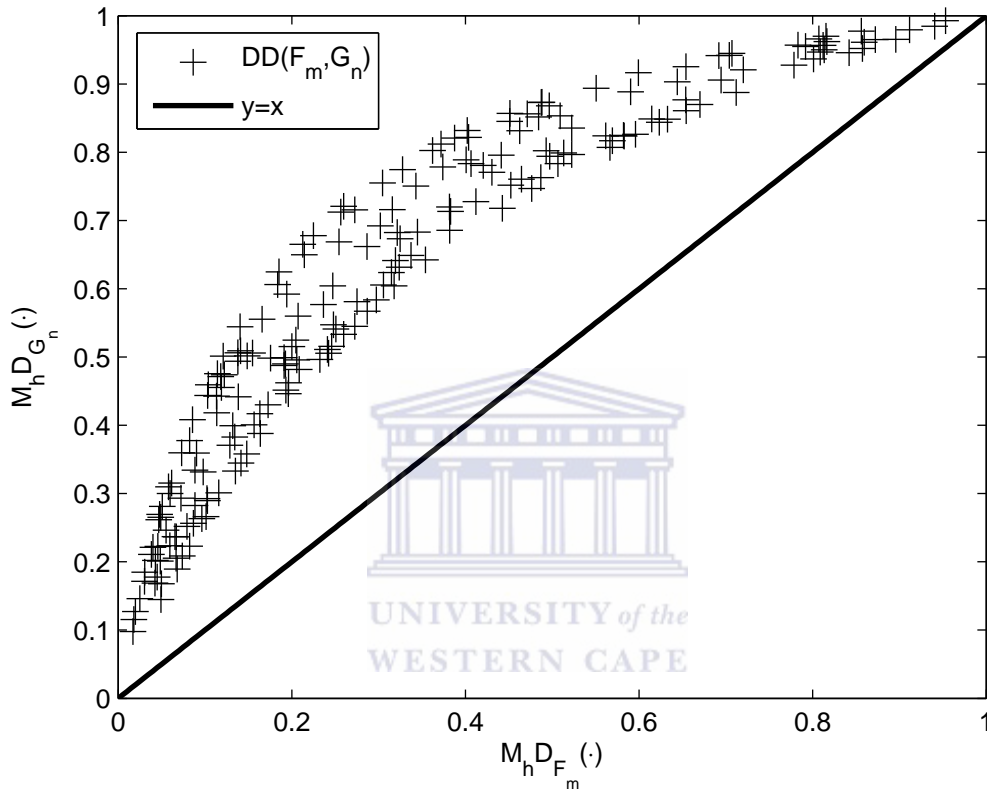


Figure 4.20: DD-plot for distributions with scale difference.

Figure 4.20 shows the DD-plot with one sample from $\text{BVN}(\mathbf{0}, \mathbf{I})$ and the other sample from the $\text{BVN}(\mathbf{0}, 4\mathbf{I})$. Notice the arching of points above the diagonal line ($y = x$). In Figure 4.20, the depth values calculated with respect to $\text{BVN}(\mathbf{0}, 4\mathbf{I})$ were plotted as x -co-ordinates. Typically, this pattern of deviation from linearity, or its reflection with respect to the line $y = x$, serves as an indicator of scale differences in multivariate settings (Liu, Parelius and Singh, 1999).

Chapter 5

Analysis of Cluster Data

A globular cluster is a spherical collections of typically tens of thousands of stars, placed closely together in space. It has relatively high stellar density toward the centre. In this chapter, data sets used consist of brightness measurements of globular cluster stars to illustrate comparison of high dimensional data by means of the multivariate two-sample test statistics. The MATLAB programmes, as well as the datasets used in the analyses below, are given in the folder **Cluster Analysis Routines**, on the accompanying CD.

Piotto *et al.* (2002) used these data in the investigation of stellar dynamics and stellar evolution in globular clusters. The data sets (available at the Padova Globular Cluster Group archives at <http://dipastro.astro.unipd.it/globular>) contain brightness data of the stars in globular clusters. The measurements were recorded from the Wide Field and Planetary Camera 2 (WFPC2) images: WFPC2 is a camera installed on the Hubble Space Telescope (HST). The camera features four detectors. Three of these, arranged in a reverse L-formation, comprise the Wide Field Camera (WFC) and adjacent to them is the Planetary Camera (PC), a fourth detector with different optics to afford more detailed view over a smaller region of the visual field¹. WFC and PC images are typically combined, producing the WFPC2's characteristic image shape, such as Figure 5.1, for the cluster NGC 4833. PC image recordings are identified by “chip number 1” (where “chip” refers to the detector). Measurements from the WFCs are referred to as chip numbers 2, 3 and 4, depending on which of the three WFC detectors was used. Photometric (i.e. brightness or intensity) data

¹see http://en.wikipedia.org/wiki/Wide_Field_and_Planetary_Camera_2

from the four detectors are stored in a single file known as a “4-chip-stack file”, for each globular cluster. Table 5.1 shows a partial 4-chip-stack photometric file of the NGC 4833 cluster. For each data set, the positions (x, y) of the stars reported in the photometric files were extracted by chip number and then an appropriate co-ordinate transformation was applied to find relative spatial positions of the stars. Figure 5.1 shows the orientations of the four images from the PC, WFC2, WFC3 and WFC4 cameras for the NGC 4833 cluster.

Figure 5.2 illustrates selected stellar positions of the observations on PC, WF2, and WF4, for the cluster NGC 4833 which were used in the analyses. Geometrically, the three portions considered in Figure 5.2, for the NGC 4833 cluster, are congruent. The purpose of the statistical analysis is to study the homogeneity of the stellar brightness properties across the globular clusters. This aim is facilitated by first comparing stars from two regions far from the centre - the outer quarters of chips 2 and 4 are used for this purpose [step (i)]. If the null hypothesis of equal populations is accepted, these two sets of data are combined and compared with the photometric properties of the stars from the globular cluster centre, that is, chip 1 stars [step (ii)].

Single-chip data sets contain measurements of each star on the chip through two different filters denoted by F439W and F555W. Analysis concentrated on the F439W and F555W brightness data. These were analyzed as bivariate data on the variables *colour index*, $X_1 = F439W - F555W$ and *brightness*, $X_2 = F555W$.

The procedures were performed for all the clusters analyzed. Figures 5.3 to 5.22 show scatter plots of the data sets which were analyzed, with vertical axes inverted. Tables 5.2 and 5.3 list the cluster ID numbers and the results from the analyses using the statistics investigated in the power studies. The p -values of all the test statistics were obtained by 1000 permutation resamplings, for sufficient accuracy of approximation, except the Friedman-Rafsky statistic T_{FR} , for which p -values were determined from its asymptotic distribution.

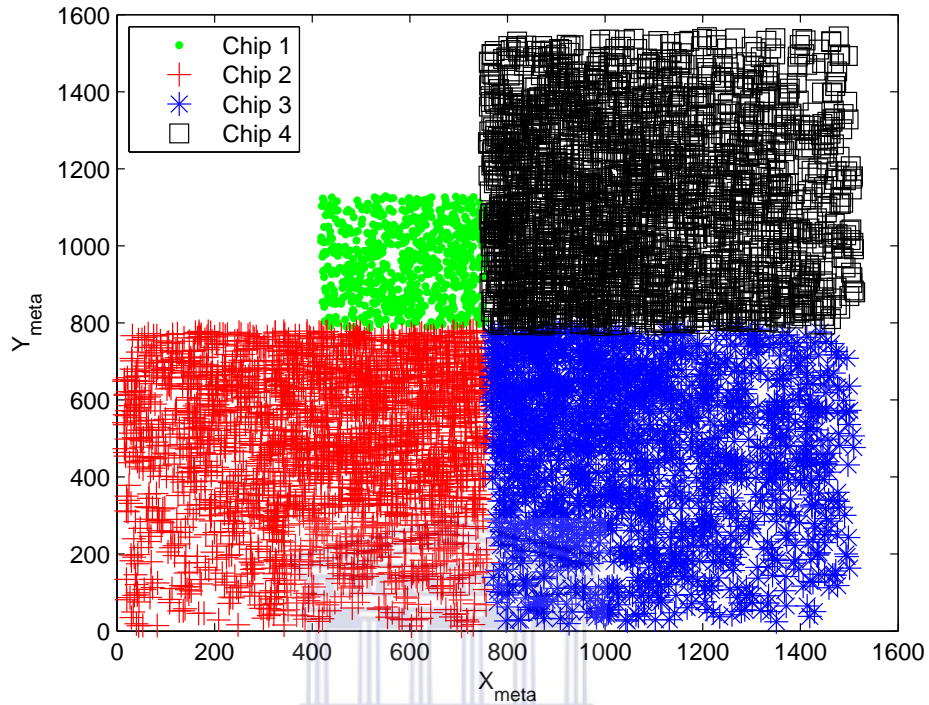


Figure 5.1: Orientation of chips for the NGC 4833 cluster.

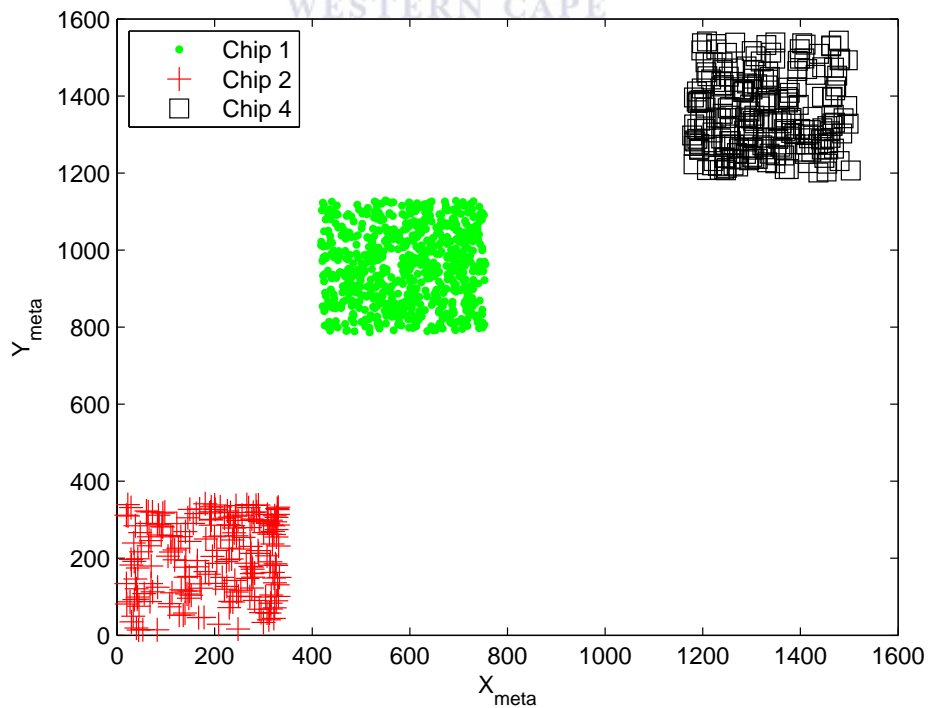
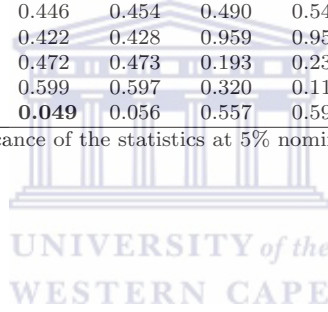


Figure 5.2: Parts of the chips analyzed for NGC 4833 cluster.

Table 5.2: Step (i) p -values of the globular cluster test statistics

	T_{BF}	T_{FR}	T_F	T_K	T_F^{NN}	T_K^{NN}	T_{SKS}	$T_{H(4)}$	T_{HT}	S_{HT}
IC 1257	0.759	0.190	0.769	0.716	0.049	0.085	0.728	0.721	0.781	0.757
IC 4499	0.705	0.615	0.737	0.743	0.455	0.587	0.207	0.829	0.393	0.293
NGC 3201	0.011	0.500	0.032	0.032	0.194	0.250	0.004	0.443	0.002	0.003
NGC 4147	0.259	0.018	0.143	0.121	0.210	0.227	0.756	0.118	0.281	0.426
NGC 4372	0.914	0.916	0.597	0.597	0.081	0.031	0.656	0.972	0.923	0.836
NGC 4590	0.117	0.057	0.067	0.067	0.270	0.115	0.001	0.001	0.196	0.126
NGC 4833	0.050	0.547	0.005	0.005	0.230	0.239	0.030	0.241	0.560	0.531
NGC 5634	0.371	0.752	0.865	0.868	0.305	0.331	0.268	0.702	0.490	0.396
NGC 6171	0.013	0.260	0.525	0.529	0.767	0.809	0.011	0.048	0.049	0.046
NGC 6218	0.753	0.302	0.790	0.799	0.256	0.150	0.594	0.558	0.522	0.370
NGC 6235	0.015	0.022	0.037	0.042	0.663	0.760	0.375	0.186	0.098	0.129
NGC 6256	0.446	0.165	0.855	0.850	0.791	0.733	0.439	0.281	0.210	0.399
NGC 6287	0.002	0.001	0.360	0.304	0.238	0.109	0.001	0.003	0.001	0.001
NGC 6325	0.001	0.285	0.117	0.072	0.457	0.403	0.008	0.015	0.032	0.055
NGC 6342	0.103	0.005	0.346	0.376	0.593	0.628	0.002	0.002	0.276	0.223
NGC 6355	0.001	0.640	0.446	0.454	0.490	0.546	0.003	0.803	0.005	0.001
NGC 6362	0.541	0.281	0.422	0.428	0.959	0.952	0.094	0.116	0.687	0.555
NGC 6380	0.143	0.004	0.472	0.473	0.193	0.237	0.019	0.121	0.073	0.026
NGC 6401	0.161	0.698	0.599	0.597	0.320	0.118	0.109	0.622	0.112	0.087
NGC 6838	0.340	0.514	0.049	0.056	0.557	0.598	0.236	0.234	0.353	0.268

Bold p -values indicate significance of the statistics at 5% nominal level.

**Table 5.3:** Step (ii) p -values of the globular cluster test statistics

	T_{BF}	T_{FR}	T_F	T_K	T_F^{NN}	T_K^{NN}	T_{SKS}	$T_{H(4)}$	T_{HT}	S_{HT}
IC 1257	0.368	0.015	0.352	0.394	0.043	0.010	0.015	0.028	0.433	0.197
IC 4499	0.001	0.271	0.159	0.052	0.656	0.413	0.001	0.003	0.001	0.001
NGC 3201	0.001	0.003	0.096	0.098	0.800	0.606	0.004	0.001	0.003	0.001
NGC 4147	0.038	0.006	0.016	0.019	0.270	0.317	0.018	0.281	0.056	0.031
NGC 4372	0.001	0.076	0.101	0.089	0.095	0.090	0.011	0.001	0.050	0.014
NGC 4590	0.001	0.017	0.267	0.308	0.424	0.242	0.001	0.001	0.001	0.001
NGC 4833	0.001	0.000	0.079	0.104	0.002	0.001	0.001	0.001	0.001	0.001
NGC 5634	0.001	0.000	0.751	0.718	0.082	0.285	0.001	0.001	0.002	0.003
NGC 6171	0.002	0.032	0.147	0.184	0.914	0.948	0.029	0.011	0.008	0.009
NGC 6218	0.001	0.000	0.016	0.021	0.290	0.014	0.001	0.001	0.001	0.001
NGC 6235	0.001	0.010	0.099	0.148	0.004	0.001	0.001	0.039	0.002	0.001
NGC 6256	0.001	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
NGC 6287	0.001	0.000	0.024	0.052	0.001	0.001	0.001	0.002	0.001	0.001
NGC 6325	0.006	0.003	0.094	0.072	0.016	0.103	0.001	0.001	0.002	0.001
NGC 6342	0.001	0.010	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001
NGC 6355	0.001	0.008	0.120	0.177	0.363	0.106	0.001	0.004	0.010	0.001
NGC 6362	0.003	0.330	0.196	0.189	0.148	0.152	0.002	0.279	0.027	0.008
NGC 6380	0.001	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
NGC 6401	0.003	0.202	0.679	0.749	0.023	0.036	0.001	0.010	0.008	0.001
NGC 6838	0.416	0.326	0.341	0.332	0.971	0.949	0.254	0.128	0.201	0.259

Bold p -values indicate significance of the statistics at 5% nominal level.

Interpretation of the results in Tables 5.2 and 5.3 is aided by summarizing as follows:

- (a) count the number of rejections at the given nominal level α for step (i) results in Table 5.2 and compute the percentage for each test statistic;
- (b) repeat the procedure for step (ii) results in Table 5.3.

Table 5.4: Percentages (%) of data sets in Tables 5.2 and 5.3 for which the null hypothesis was rejected at the 5% level for each test statistic

	T_{BF}	T_{FR}	T_F	T_K	T_F^{NN}	T_K^{NN}	T_{SKS}	$T_{H(4)}$	T_{HT}	S_{HT}
Step (i)	35	25	20	15	5	5	45	25	25	25
Step (ii)	90	75	30	25	45	45	95	85	85	90

An examination of Table 5.4 suggests that the test statistics T_{BF} , T_{FR} , T_{SKS} , $T_{H(4)}$, T_{HT} , and S_{HT} have similar discriminating ability. The large percentages of rejection for these test statistics in step (ii) analyses suggest that they are sensitive against the type of departures from equality that are common in the cluster colour - brightness data. Moreover, the six test statistics were shown in Chapter 4 to have similarly high powers against location-shift alternatives.

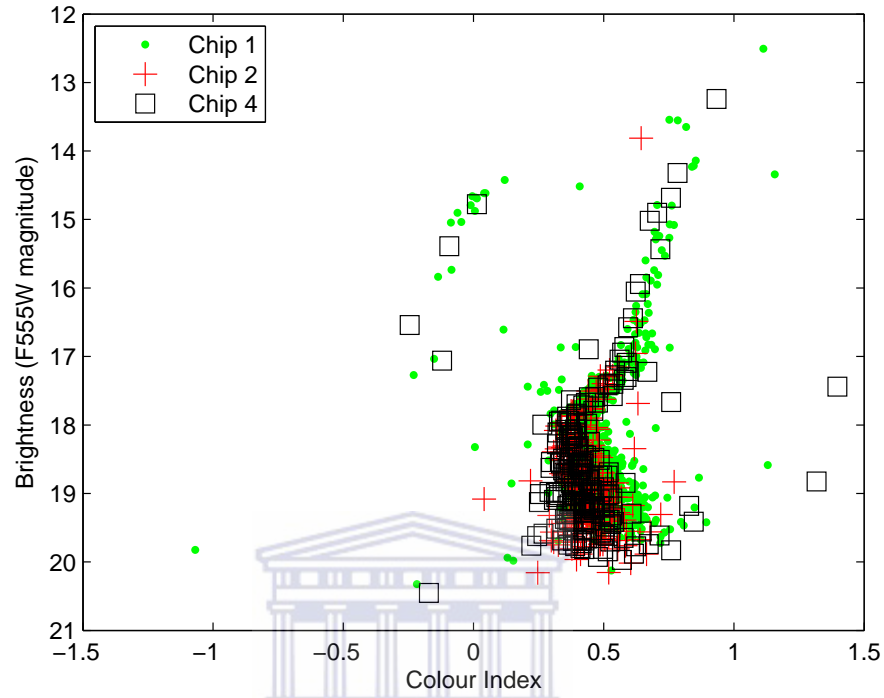


Figure 5.3: Scatter plot for NGC 4833 cluster.

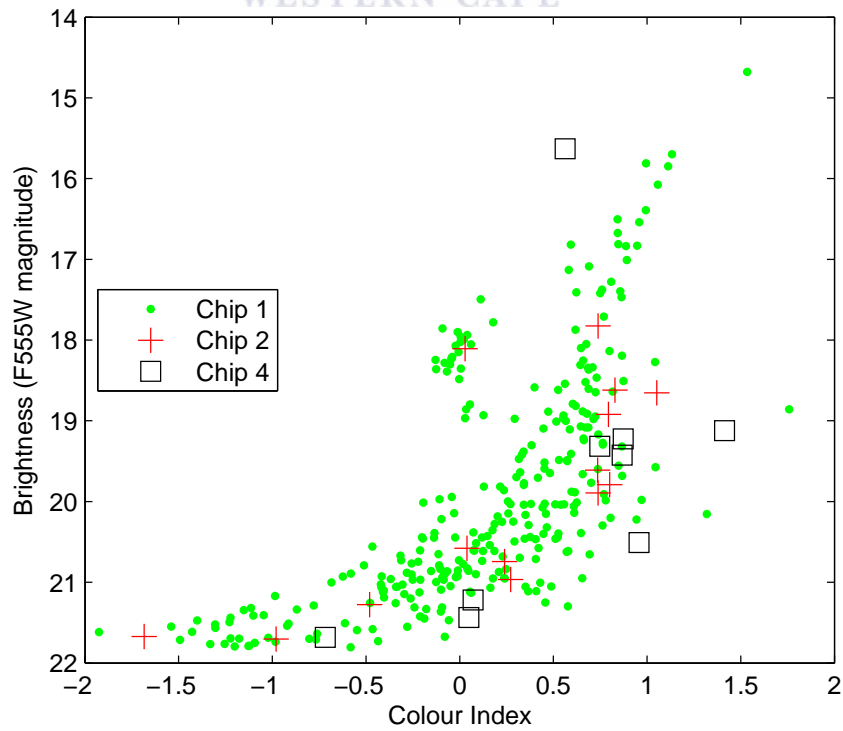


Figure 5.4: Scatter plot for IC 1257 cluster data.

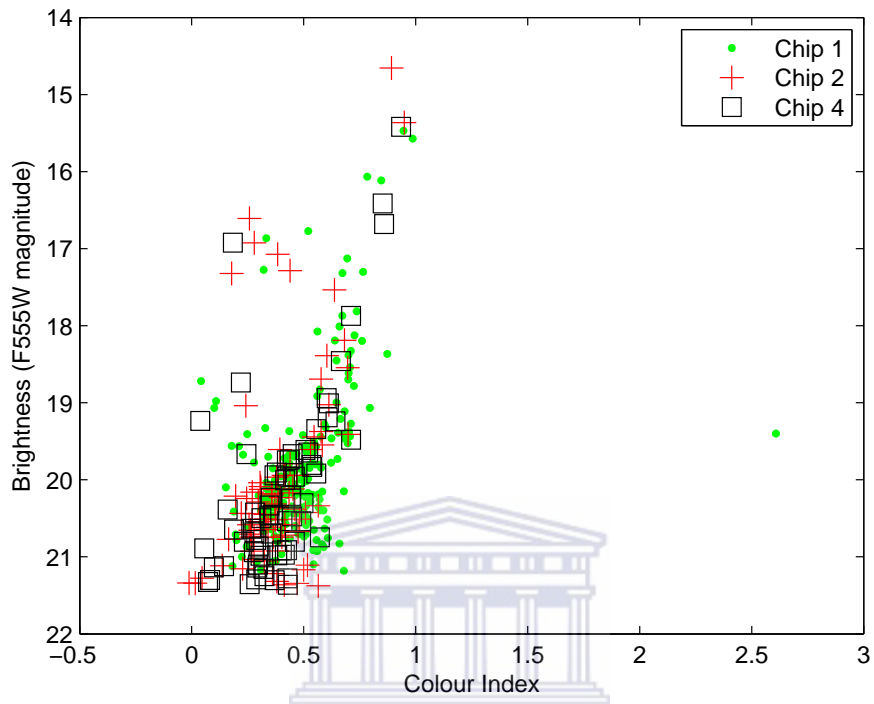


Figure 5.5: Scatter plot for IC 4499 cluster data.

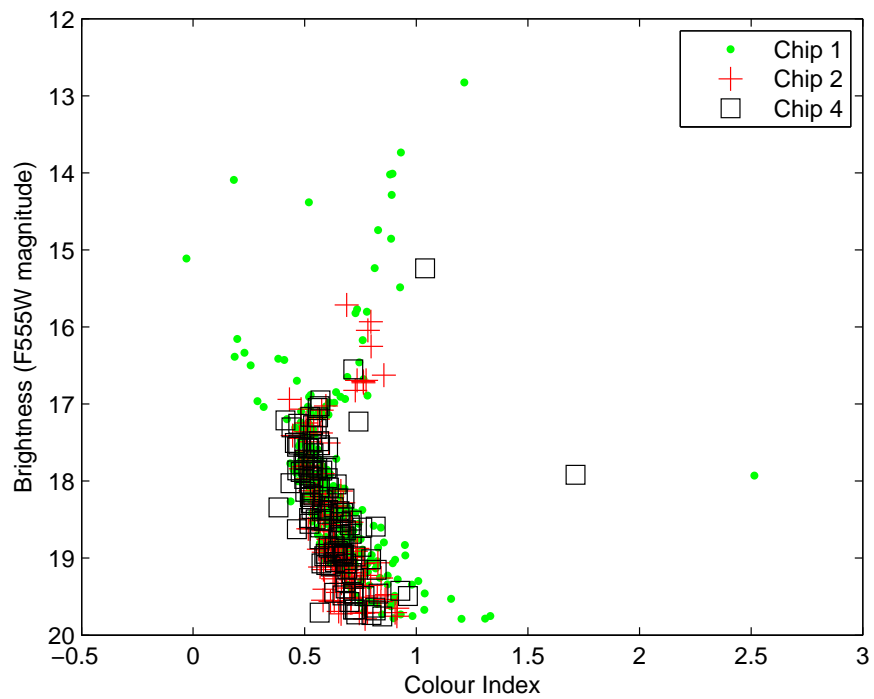


Figure 5.6: Scatter plot for NGC 3201 cluster data.

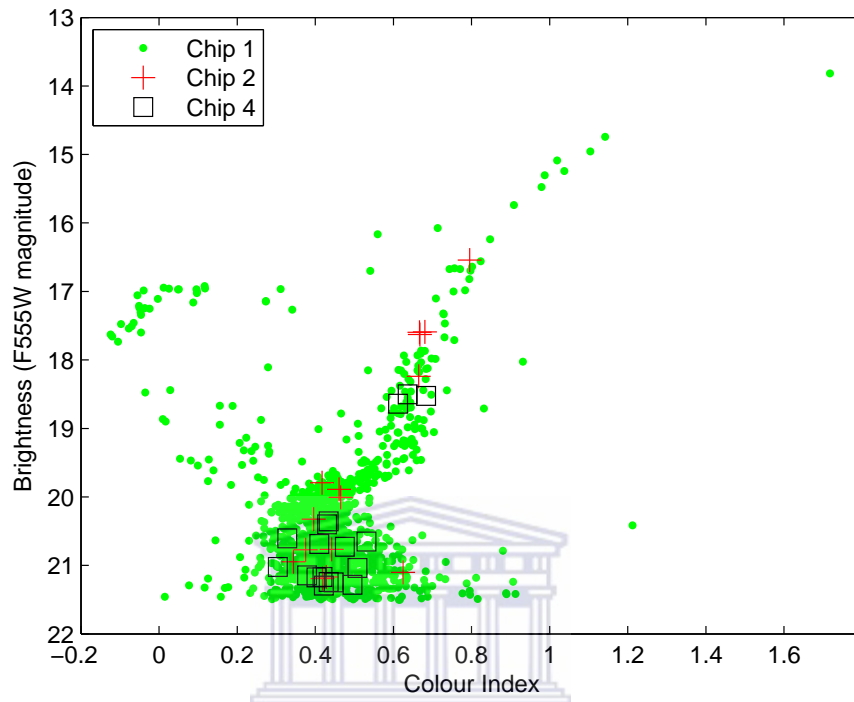


Figure 5.7: Scatter plot for NGC 4147 cluster data.

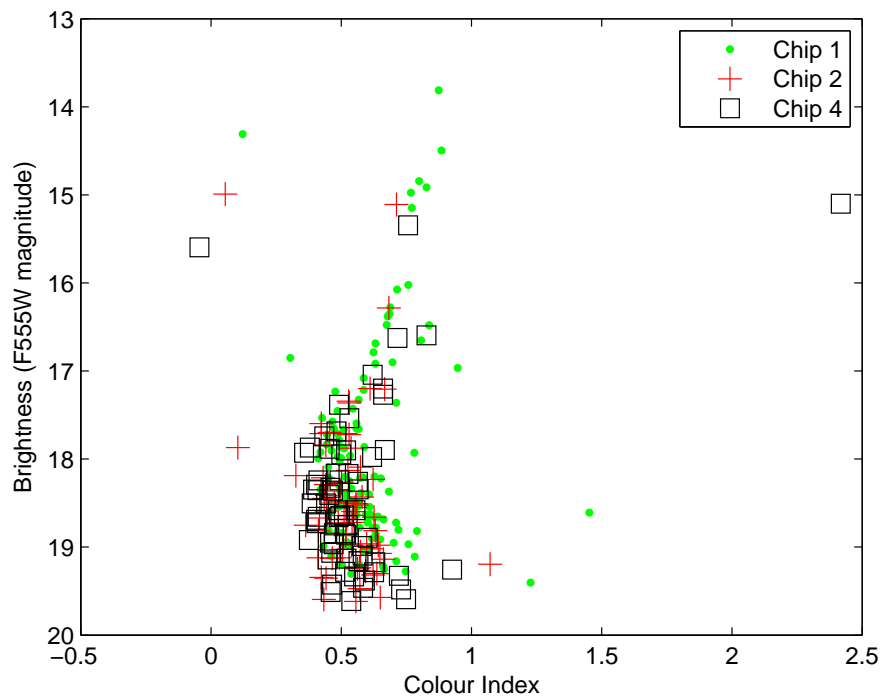


Figure 5.8: Scatter plot for NGC 4372 cluster data.

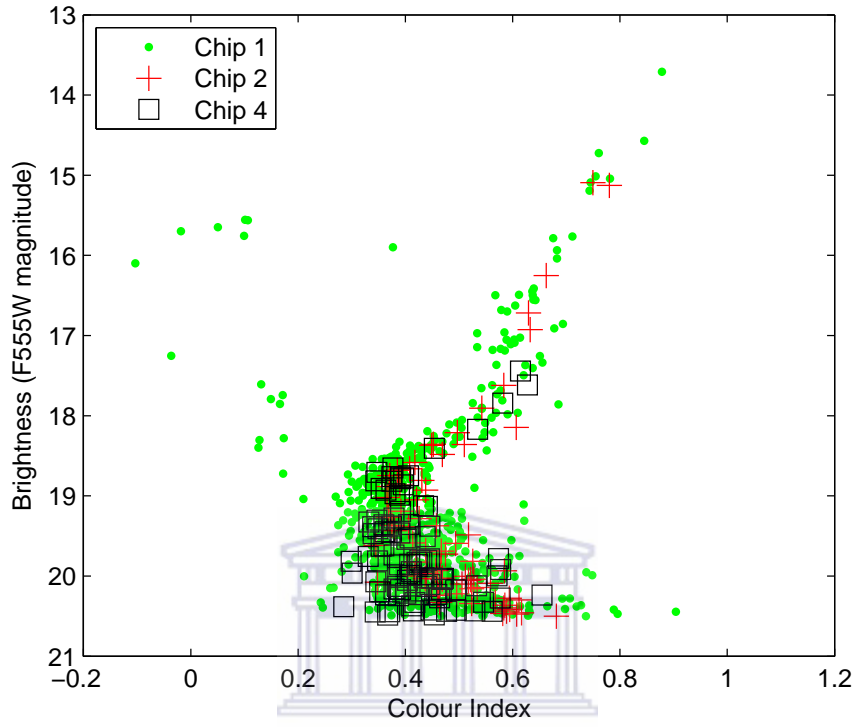


Figure 5.9: Scatter plot for NGC 4590 cluster data.

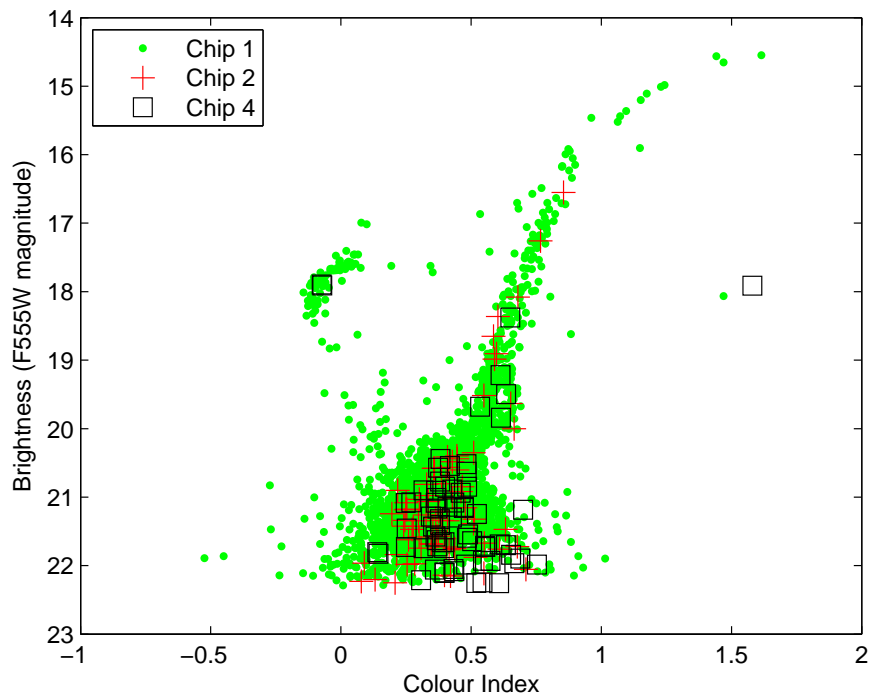


Figure 5.10: Scatter plot for NGC 5634 cluster data.

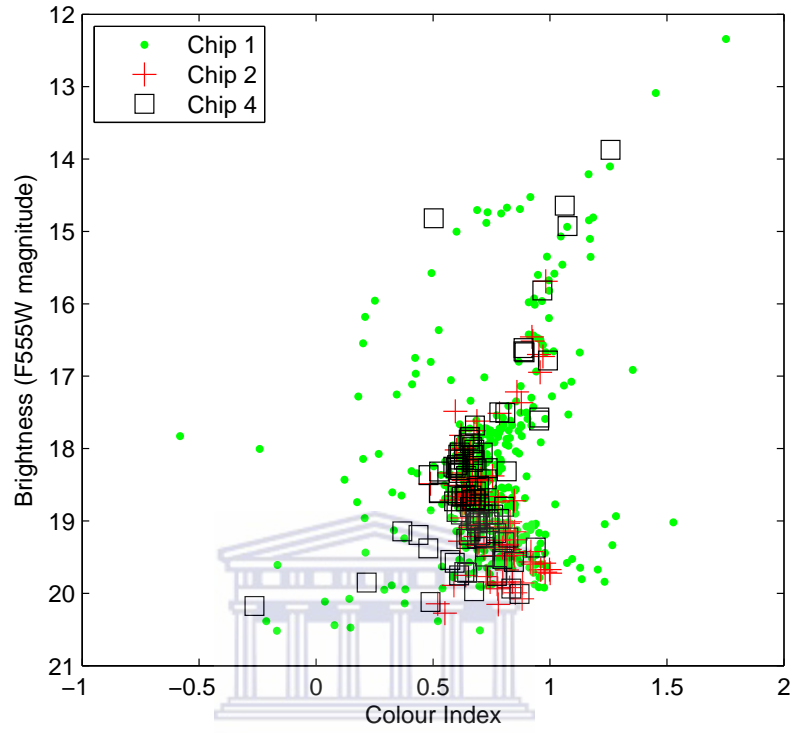


Figure 5.11: Scatter plot for NGC 6171 cluster data.

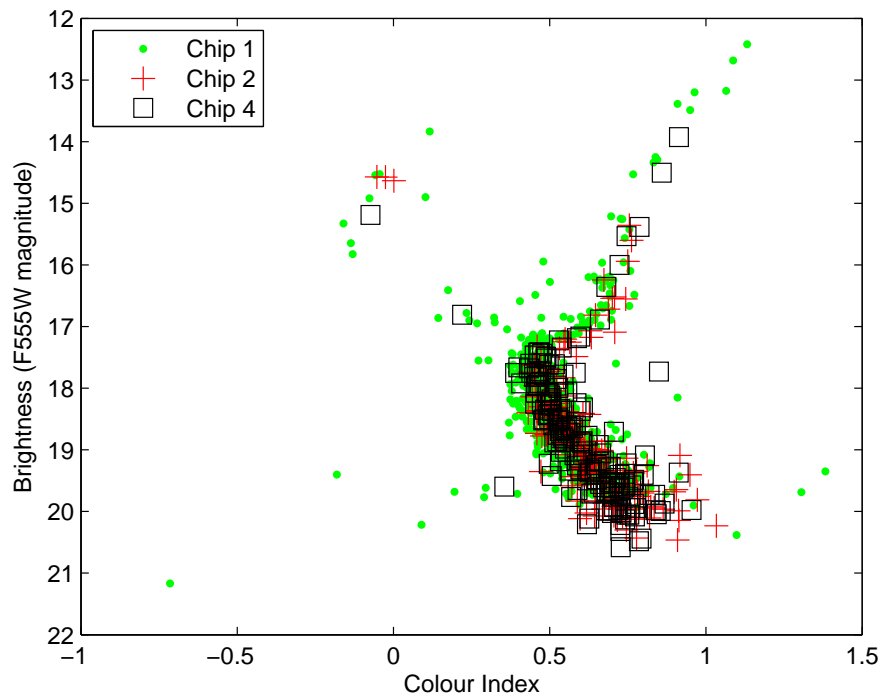


Figure 5.12: Scatter plot for NGC 6218 cluster data.

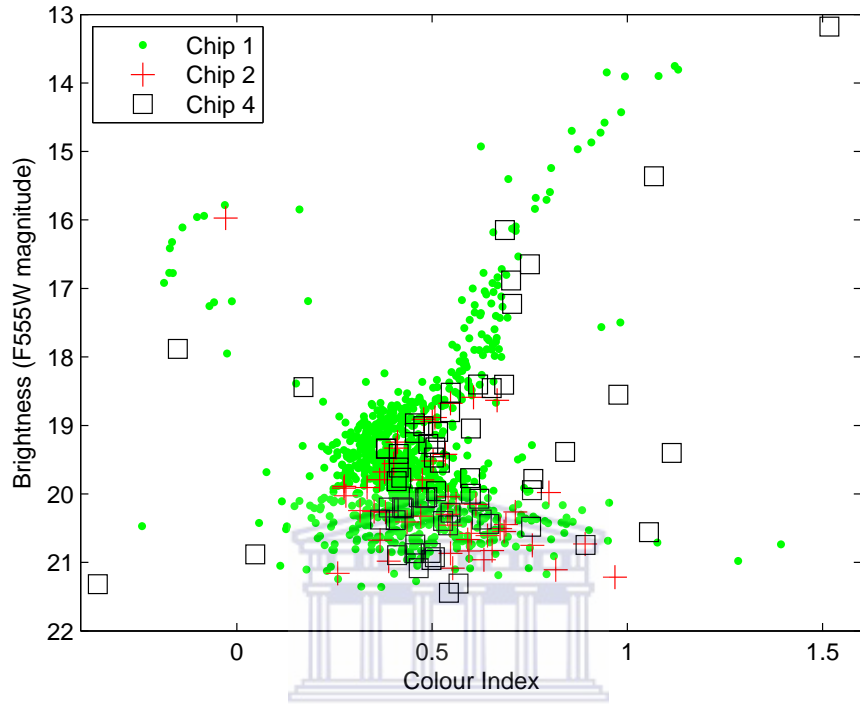


Figure 5.13: Scatter plot for NGC 6235 cluster data.

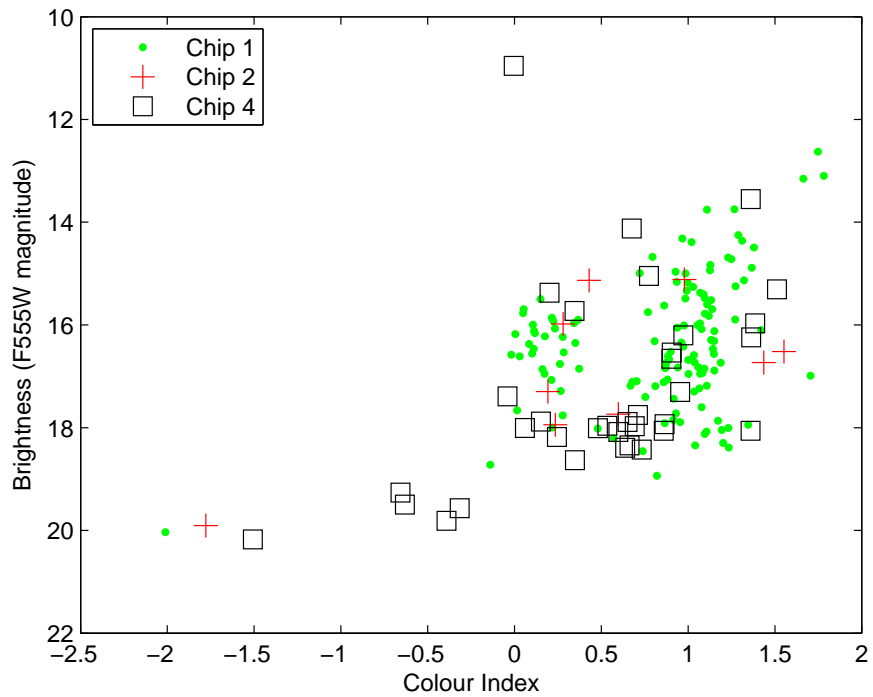


Figure 5.14: Scatter plot for NGC 6256 cluster data.

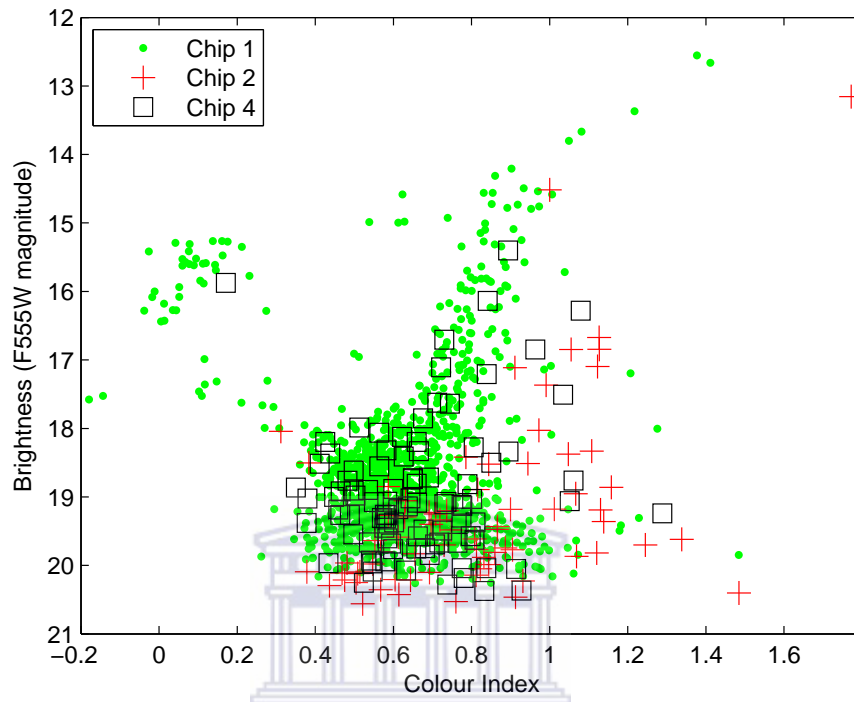


Figure 5.15: Scatter plot for NGC 6287 cluster data.

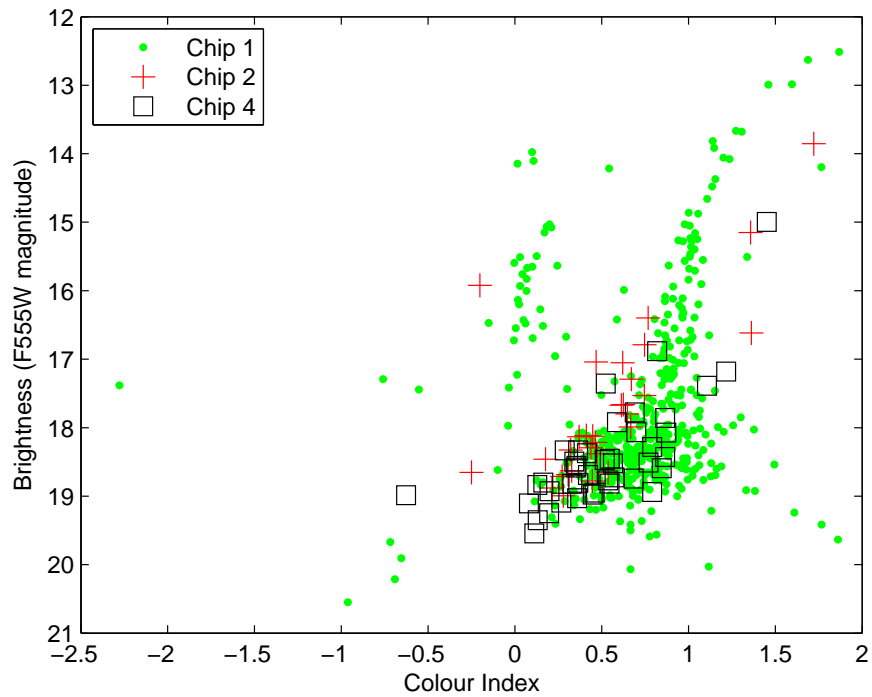


Figure 5.16: Scatter plot for NGC 6325 cluster data.

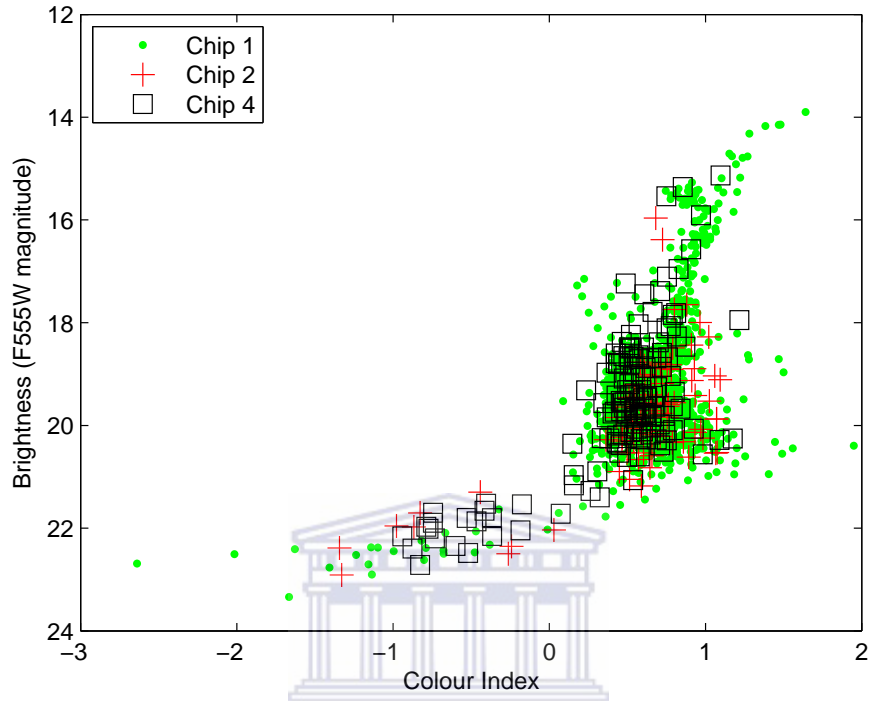


Figure 5.17: Scatter plot for NGC 6342 cluster data.

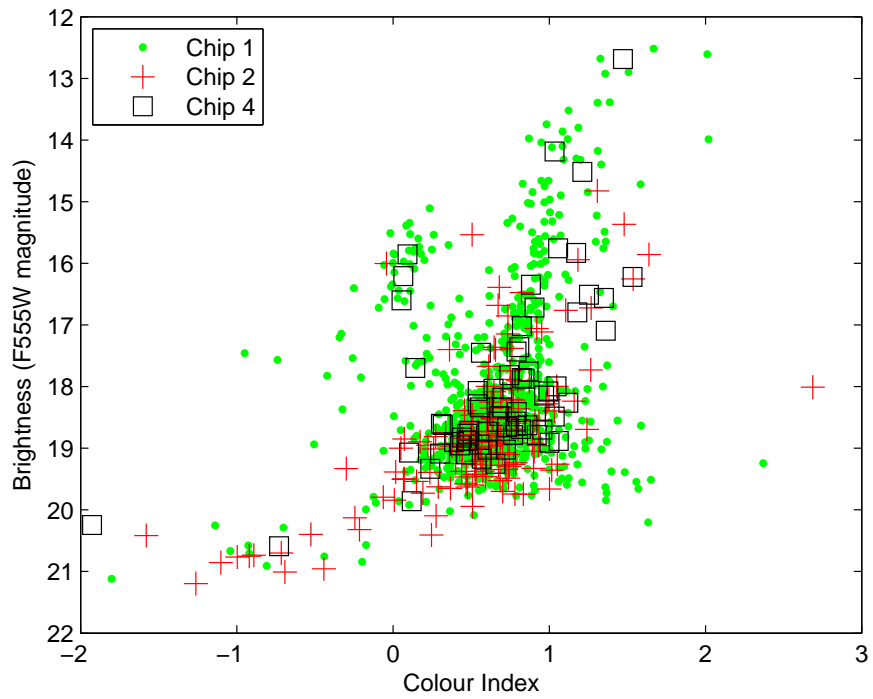


Figure 5.18: Scatter plot for NGC 6355 cluster data.

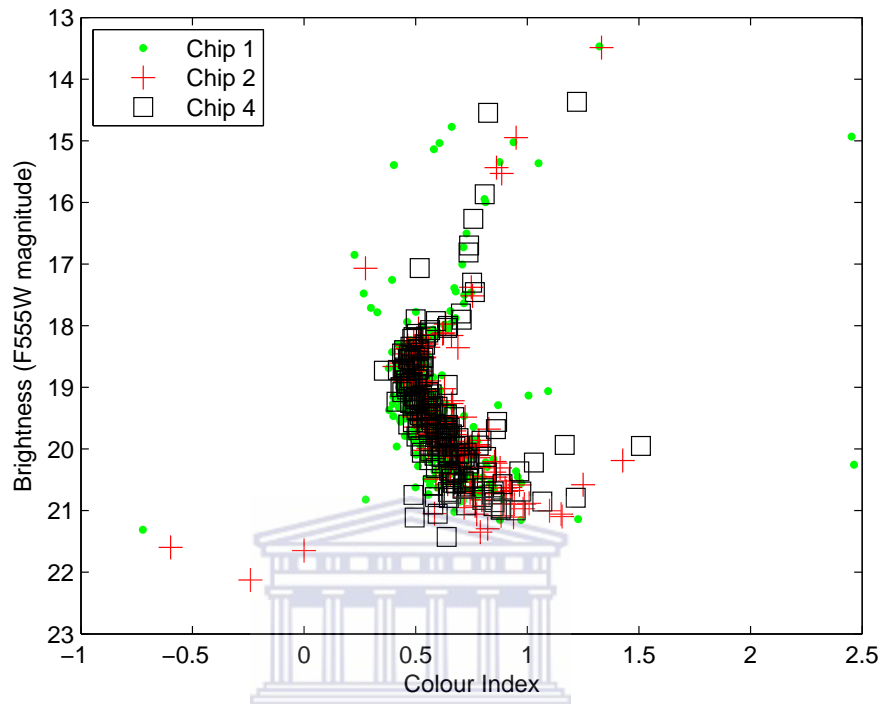


Figure 5.19: Scatter plot for NGC 6362 cluster data.

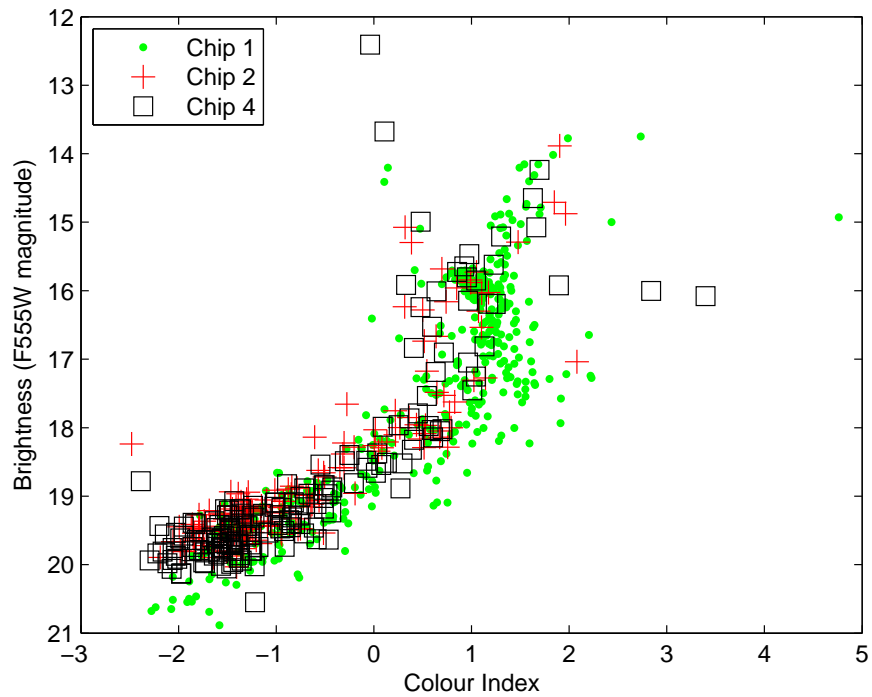


Figure 5.20: Scatter plot for NGC 6380 cluster data.

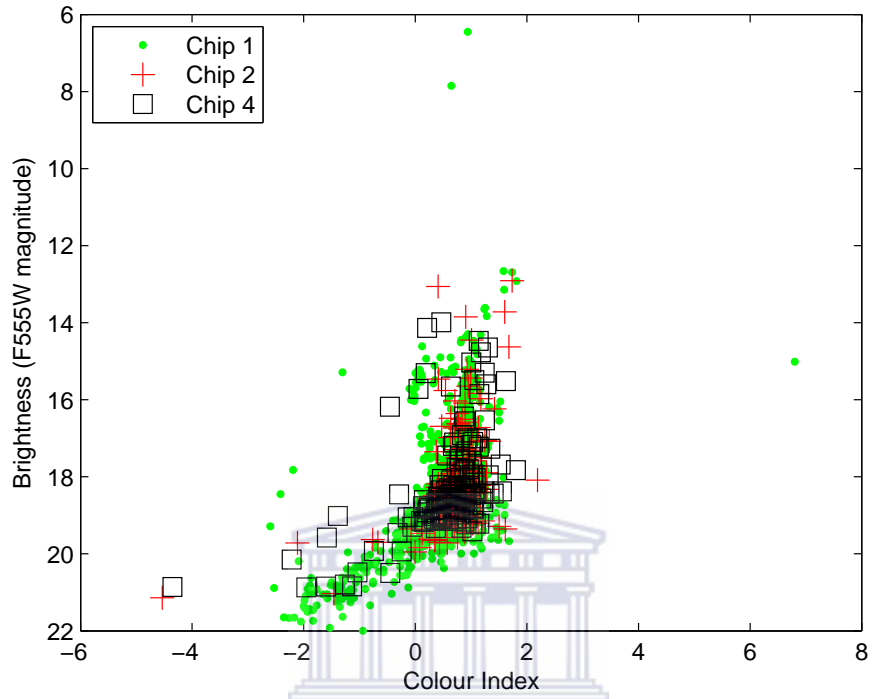


Figure 5.21: Scatter plot for NGC 6401 cluster data.

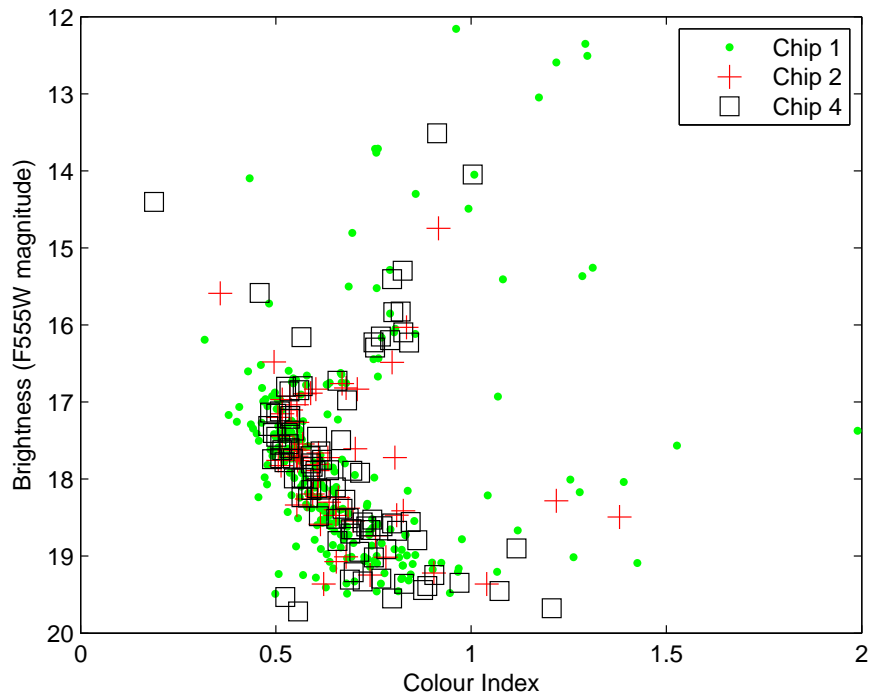


Figure 5.22: Scatter plot for NGC 6838 cluster data.

Chapter 6

Conclusion

Little is known about the practical application of multivariate two-sample tests. Possible reasons are: ignorance about the existence of the variety of statistical tests for multivariate two-sample problems; unavailability of ready-to-use software; and the reluctance by practitioners to use tests when little is known about their power and robustness. It was against this background that studies were done of the relative power of the selected tests. This thesis investigated the powers of the EDF and interpoint distance type tests for a range of alternatives from bivariate distributions of the exponential, normal, and uniform types. On the basis of the results from the power studies, it was established that it is not possible to make a general recommendation to always use a particular multivariate two-sample test statistic, irrespective of the sampled distribution. Table 6.1 shows a general summary of the recommendations based on the study in Chapter 4.

Table 6.1: Statistics recommended for analysis

Alternatives	Statistics
Location	T_{BF} T_{SKS} T_{HT} S_{HT}
Scale	T_F T_K
Correlation	$T_{H(4)}$ T_{FR}

Results from the power studies suggest that some tests have power against specific alternatives and may not be useful for other alternatives. Particular choices depend on the type of potential differences between the populations that are deemed

important to detect. If the user is going to rely on one and only one multivariate two-sample test, then the Baringhaus-Franz statistic T_{BF} is recommended for location alternatives; the IPDD test via either T_F or T_K should be preferred for scale problems; and the nearest neighbour test statistic $T_{H(4)}$ should be the choice for correlation alternatives. These recommendations are based on the good power, which is either comparable or superior to the other tests, against the entire range of alternatives considered in the power studies. Other multivariate two-sample test statistics which have good power are those shown in Table 6.1. Moreover, the implementation of these test statistics is fairly easy and computationally fast. The Baringhaus-Franz statistic is available as a ready-to-use test known as the *Cramer test* in the R language (Baringhaus and Franz, 2001).

Since the statistics are omnibus, they are not helpful in diagnosing the nature of the departure from the null hypothesis. The rejection of the hypothesis can be complemented with a non-parametric graphical procedure such as the DD-plots.

In the power studies, permutation approximations of the exact distributions for the selected test statistics were used, with the exception of the statistic T_{FR} for which the asymptotic distribution was used because of considerations of computation time. The large number of permutations used generally provides a more accurate approximation of the exact distribution of the test statistic than asymptotic forms. Furthermore, for some of the test statistics, asymptotic distributions are unavailable. However, the use of asymptotic distributions of the multivariate two-sample tests should be recommended if their accuracy is guaranteed for relatively small sample sizes, because the permutation method is very demanding computationally and could take hours to days to produce results. Besides, many potential users of the multivariate two-sample tests may have neither the necessary skills nor the inclination to empirically determine p -values each time they apply the tests.

The utility of the multivariate two-sample tests was demonstrated in the analysis of photometric data sets of twenty galactic globular clusters. An additional application of the test statistics is given by Koen and Siluyele (2007).

The determination of the power of the test statistics for different types of bivariate distributions, and the inclusion of the correlation alternatives, are the much needed extensions to published studies of the multivariate two-sample tests. However, substantial scope exists for further extensions:

- (i) other significance levels and sample sizes;
- (ii) more complicated alternatives to the null hypothesis;
- (iii) higher dimensionality of the samples;
- (iv) other distributions, which may include mixtures of distributions.



Bibliography

- [1] Annis C.P.E. (2006). *Statistical engineering: Curse of Dimensionality*. [Online]. Available www.statisticalengineering.com/curse_of_dimensionality.htm
- [2] Baringhaus L. & Franz C. (2001). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, **88**: 190-206.
- [3] Bauer H. (1972). *Probability theory and elements of measure theory*. New York: Holt, Rinehart and Winston, Inc., pp 129-148.
- [4] David H.T. (1958). A three-sample Kolmogorov-Smirnov test. *Annals of Mathematical Statistics*, **29**(2): 842-851.
- [5] Fasano G. & Franceschini A. (1987). A multidimensional version of the Kolmogorov-Smirnov test. *Monthly Notice of the Royal Astronomical Society*, **225**: 155-170.
- [6] Fisz M. (1963). *Probability theory and mathematical statistics* (3rd edition). Malabar: Robert E. Krieger Publishing Company, Inc.: 408-410.
- [7] Friedman H.J. & Rafsky C.L. (1979). A multivariate generalization of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, **7**: 697-717.
- [8] Gibbons D.J. (1985). *Nonparametric Methods for Quantitative Analysis* (2nd edition). Columbus: American Sciences Press, Inc.: 250-257.
- [9] Greenberg S.L. (2006). *Bivariate goodness-of-fit tests based on Kolmogorov-Smirnov type*. Unpublished Master's thesis. Johannesburg: University of Johannesburg.
- [10] Hall P. & Tajvidi N. (2002). Permutation test for equality of distributions in high-dimensional Settings. *Biometrika*, **89**(2): 359-374.
- [11] Henze N. & Penrose M. (1999). On the multivariate runs test. *The Annals of Statistics*, **27**(1): 290-289.
- [12] Henze N. (1988). A multivariate two-sample test based on the number of nearest neighbour type coincidences *The Annals of Statistics* **16**(2): 772-783.

- [13] Justel A., Peña D. & Zamar R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters*, **35**: 251-259.
- [14] Kiefer J. (1959). K-sample analogues of the Kolmogorov-Smirnov and Cramér-von Mises tests. *Annals of Mathematical Statistics*, **30**(2): 420-447.
- [15] Koen C. & Siluyele I. (2007). Multivariate comparisons of the period-light curve shape distributions of the Cepheids in five Galaxies. *Monthly Notices of the Royal Astronomical Society*, **377**, 1281-1286.
- [16] Liu Y., Parelius J.M. & Singh K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, **27**: 783-858.
- [17] Maa J., Pearl K.D. & Bartoszyński R. (1996). Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *The Annals of Statistics*, **24**(3): 1069-1074.
- [18] Mathews H.J. & Fink D.K. (1999). *Numerical methods using Matlab* (3rd edition). Upper Saddle River: Prentice Hall, Inc.: 101-186.
- [19] Metzger W.J. (1999). The minimum spanning tree as a two-sample test. [Online]. Available www.hef.ru.nl/~wes/papers/hen419.ps.gz
- [20] Morgenstern D. (2001). Proof of a conjecture by Walter Deuber concerning the distances between points of two types in \mathbb{R}^d . *Discrete Mathematics*, **226**: 347-349.
- [21] Park M.H. (2006). *Understanding the statistical power of a test*. [Online]. Available www.indiana.edu/~statmath/stat/all/power/power.html
- [22] Peacock J.A. (1983). Two-dimension goodness-of-fit testing in Astronomy. *Monthly Notices of the Royal Astronomical Society*, **202**: 615-627.
- [23] Press W.H., Teukolsky S.A., Vetterling W.T. & Flannery B.P. (1992). *Numerical recipes in Fortran 77: The art of scientific computing*. Cambridge: Cambridge University Press: 640 - 642.
- [24] Piotto G., King I., Djorgovski S.G., Sosin C., Zoccali M., Saviane I., De Angeli F., Riello M., Recio Blanco A., Rich R.M., Meylan G. & Renzini A. (2002). HST colour-magnitude diagrams of 74 Galactic Globular clusters in the HST F439W and F555W Bands. *Astronomy & Astrophysics*, **391**: 945-965.
- [25] Rohatgi K.V. (1984). *Statistical Inference*. New York: John Wiley & Sons, Inc.
- [26] Rosenbaum P.R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of Royal Statistical Society B*, **67**(Part 4): 515-530.

- [27] Schilling M.F. (1986). A multivariate two-sample test based on number of nearest neighbour. *Journal of the American Statistical Association*, **81**(395): 799-806.
- [28] Schmidt F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implication for training of researchers. *Psychological Methods*, **1**(2): 115-129.
- [29] Steele M.C. (2002). *The power of categorical goodness of fit test statistics*. PhD thesis. Brisbane: Griffith University: Australian School of Environmental Studies. [Online]. Available <http://www4.gu.edu.au:8080/adt-root/public/adt-QGU20031006.143823/index.html>
- [30] Thas, O. (2001). *Nonparametrical tests based on space partitions*. PhD thesis, Ghent: Ghent University: Faculty of Agriculture and Applied Biological Sciences. [Online]. Available http://biomath.rug.ac.be/publications/download/thasolivier_phd.pdf
- [31] Weiss L. (1960). Two-sample tests for multivariate distributions. *The Annals of Mathematical Statistics*, **31**(1): 159-164.
- [32] Wiesen C.A. & Schlenger W.E. (1998). *Statistical Power: A primer*. [Online]. Available www.workplace.samhsa.gov/WPResearch/Methodology-Data/primer.pdf