Rank Test for Multivariate Two Sample Data using Projection Pursuit

by

Unawatuna Gamage Asiri Gunathilaka, B.S.

A Thesis

in

Statistics

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for the Degree of

Master of Science

Approved

Prof. Frits Ruymgaart

Dr. Petros Hadjicostas

John Borrelli
Dean of the Graduate School

August, 2007

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor Professor Frits Ruymgaart for his unconditional support during my work. Your enthusiasm and avid interest in the subject promoted a very exciting atmosphere. Whenever I was facing a difficulty, your advice always came at the right time.

My sincere gratitude goes to the Dr. Petros Hadjicostas for his great efforts to provide me constructive comments during my thesis time as well as on the preliminary version of this thesis. Also I wish to thank Dr. Robert E. Byerly for his creative suggestions on the Latex draft to improve the visual appearence of my thesis.

During my time at Texas Tech University I had the opportunity to learn and work with many great teachers. I am indebted to them for inspiring me to see the beauty of Mathematics and Statistics. Furthermore, I am grateful to all those who helped me to complete this thesis, especially my roommates Hemal, Parakrama and Nadeeka for their stimulating suggestions and continuous encouragements.

Last but not all the least, I am deeply indebted to my mother Indra. She raised me, supported me, taught me and loved me through all my ups and downs in my life. To her, I dedicated this thesis.

TABLE OF CONTENTS

ABSTRACT

Construction of an asymptotically distribution free test for the hypothesis that two multivariate random samples are identically distributed has been a topic among many statisticians for a long time. Although this problem has been solved for random samples of multivariate normal data within the parametric setting, there are not many studies in the literature for treating this problem with random samples from arbitrary unknown distributions. This thesis sheds a new light on this topic proposing an innovative nonparametric procedure which can be applied for any two random samples from unknown distributions.

In our approach we propose to establish a multiple direction rank statistic developed based on the projected data towards some arbitrary directions. Next we develop the test statistic in terms of this multiple direction rank statistic, which can be used to test whether the two samples have the same underlying distribution or not.

Finally we investigate the asymptotics of our model under the null hypothesis and the local alternatives.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The work of this thesis provides a basic framework for constructing asymptotically distribution free tests for the hypothesis that two multivariate samples have the same underlying density with respect to Lebesgue measure. More precisely, let $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be two random samples in $\mathbb{R}^\nu$ which are from densities f and g, respectively. (These densities f and g are with respect to the Lebesgue measure.) The null hypothesis to be tested is

$$H_0 : f = g. \tag{1.1}$$

It is well known how this problem can be solved when both f and g are $\nu-$variate normal densities with the same covariance structure and with arbitrary means. In such a case, the null hypothesis (1.1) is equivalent to the equality of the two means.

Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be an iid random sample from a $\nu-$variate normal population with mean vector μ_1 and covariance matrix Σ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be an iid random sample from a $\nu-$variate normal population with mean vector μ_2 and covariance matrix Σ. Also assume these two samples are mutually independent. We wish to test the equality of means. In particular, we would test the null hypothesis

$$H_0 : \mu_1 = \mu_2. \tag{1.2}$$

Suppose e is an unit vector of $\mathbb{R}^\nu$, say a direction. Now projecting the data X_{ij}'s towards to the direction e we define

$$X_{ij,e} = X_{ij}^T e \tag{1.3}$$

for $\quad i = 1, 2 \quad$ and $\quad j = 1, 2, \ldots, n_i$.

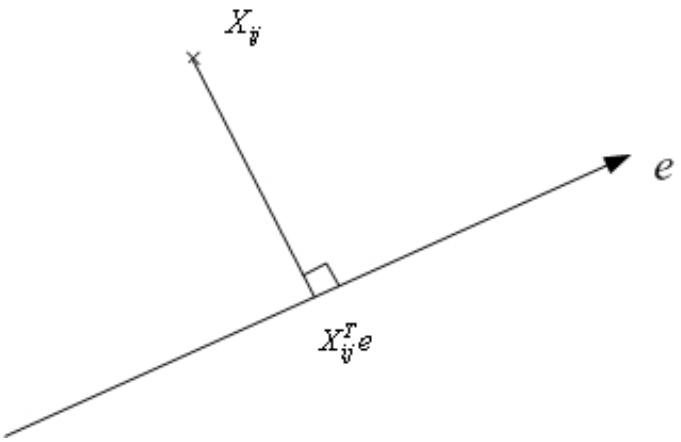

Figure 1.1. Projection of the data point X_{ij}

Thus as a result of similar type of projections of the observed data towards to the direction e, we get the univariate normal random variables $X_{11,e}, X_{12,e}, \ldots, X_{1n_1,e}$ and $X_{21,e}, X_{22,e}, \ldots, X_{2n_2,e}$, which are linear combinations of the coordinates of the observed random variables $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$, respectively. These two new random samples $X_{11,e}, X_{12,e}, \ldots, X_{1n_1,e}$ and $X_{21,e}, X_{22,e}, \ldots, X_{2n_2,e}$ have sample means and covariances $e^T\bar{X}_1$, $e^T S_1 e$ and $e^T\bar{X}_2$, $e^T S_2 e$ respectively, where

$$\bar{X}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{1j}, \tag{1.4}$$

$$\bar{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}, \tag{1.5}$$

$$S_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}\left(X_{1j} - \bar{X}_1\right)\left(X_{1j} - \bar{X}_1\right)^T, \tag{1.6}$$

$$S_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(X_{2j} - \bar{X}_2\right)\left(X_{2j} - \bar{X}_2\right)^T. \tag{1.7}$$

Now let μ_{1e} and μ_{2e} be the population means of these new random samples $X_{11,e}$, $X_{12,e}, \ldots, X_{1n_1,e}$ and $X_{21,e}, X_{22,e}, \ldots, X_{2n_2,e}$ respectively. Then considering these univariate normal quantities, the test statistic for the equality of means can be derived. In fact, the student T^2-statistic T_e^2 for the direction e can be defined in terms of $X_{1i,e}$'s and $X_{2i,e}$'s as follows.

$$T_e^2 = \frac{e^T \left(\bar{X}_1 - \bar{X}_2\right) \left(\bar{X}_1 - \bar{X}_2\right)^T e}{e^T \left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{pooled} e} \tag{1.8}$$

where

$$S_{pooled} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 + n_2 - 2)}. \tag{1.9}$$

Here we reject the null hypothesis

$$H_{0e} : \mu_{1e} = \mu_{2e} \tag{1.10}$$

if

$$T_e^2 > c \tag{1.11}$$

where the critical value c is determined from the distribution of T_e^2.

In particular, $c = F_{1,n_1+n_2-2}(\alpha)$, where F_{1,n_1+n_2-2} is an F distribution with degrees of freedom 1 and $n_1 + n_2 - 2$, and α is the prespecified significance level set up by the experimenter.

Now let us fix an arbitrary finite $k \in \mathbb{N}$ and define k different unit vectors in $\mathbb{R}^\nu$, say directions;

i.e. let $e_1, e_2, \ldots, e_k \in \mathbb{R}^\nu$ be such that $\|e_i\| = 1 \quad$ for $\quad i = 1, 2, \ldots, k$.

Following the above procedure, for each direction we can derive test statistics $T_{e_1}^2, T_{e_2}^2, \ldots, T_{e_k}^2$ to test the null hypothesis $H_{0e_1}, H_{0e_2}, \ldots, H_{0e_k}$ respectively, where $H_{0e_i} : \mu_{1e_i} = \mu_{2e_i}$ for $i = 1, 2, \ldots, k$.

More importantly, in this situation the null hypothesis (1.2) for the multivariate problem can be written as

$$H_0 : \bigcap_{\substack{i=1 \\ \|e_i\|=1}}^{k} H_{0e_i} \tag{1.12}$$

Thus according to the union−intersection principle, the test statistic T^2 for testing above null hypothses (1.2) is

$$T^2 = \max_{i=1,\ldots,n} T^2_{e_i}. \tag{1.13}$$

This method can also be exploited even if the data are not necessarily normal. Also the mean and covariance need not exist. In this case the asymptotic distribution of the above type of test statistic can be obtained.

Here, however, we want to extend the latter case even further without making the assumption that any moment exists. This leads to the null hypothesis in (1.1) that the completely arbitrary densities are the same.

The previous construction of the test statistic in the multivariate situation using the union−intersection principle however can be extended in this more general situation by introducing a two−sample rank statistic for each direction e_i.

So, once again we look for a fix direction $e \in \mathbb{R}^\nu$ such that $\|e\| = 1$. Now we focus on the data X^*e's for each observed data $X \in \mathbb{R}^\nu$.

Given these data, we construct a two−sample rank statistic S^2_e, say for every direction e. We will present the detailed definition of S^2_e later. Naturally, we could use

$$S^2 = \max_{\|e\|=1} S^2_e \tag{1.14}$$

for testing the nonparametric null hypothesis.

Though this is the essential idea, instead of

$$\max_{\|e\|=1} S_e^2$$

(a kind of Kolmogrov-Smirnov statistic), which is hard to deal with asymptotically, we consider an integral

$$\int_{\|e\|=1} S_e^2 \, de \tag{1.15}$$

which is sort of Camér−von Mises device type. In this thesis we will elaborate on a simplified version of the latter type of statistic, by summing over only a finite number of given directions $e_1, e_2, \ldots, e_k \in \mathbb{R}^\nu$.

In other words, we will focus on a statistic of the type

$$\sum_{i=1}^{k} S_{e_i}^2 = \|S\|^2 \tag{1.16}$$

where S is the vector statistic $S = (S_{e_1}, S_{e_2}, \ldots, S_{e_k})$, with S_{e_i} being an i^{th} two−sample rank statistic.

As we noted in the above formulation of the vector statistic S, we have used a finite number of arbitrary directions. However, when it comes to a situation where we would apply this result for a real life problem, we need to decide how many number of directions are appropriate, and also which directions we should select for our analysis. Even though it is difficult to answer these questions for higher dimensional data, this type of questions can be answered properly for two dimensional data by investigating some plotted graphs. Thus, in order to explain the process of selecting an appropriate number of directions for a given problem, we present few examples here for two dimensional data.

First we consider Figure 1.2, which is a plotted graph of the data from two independent samples.

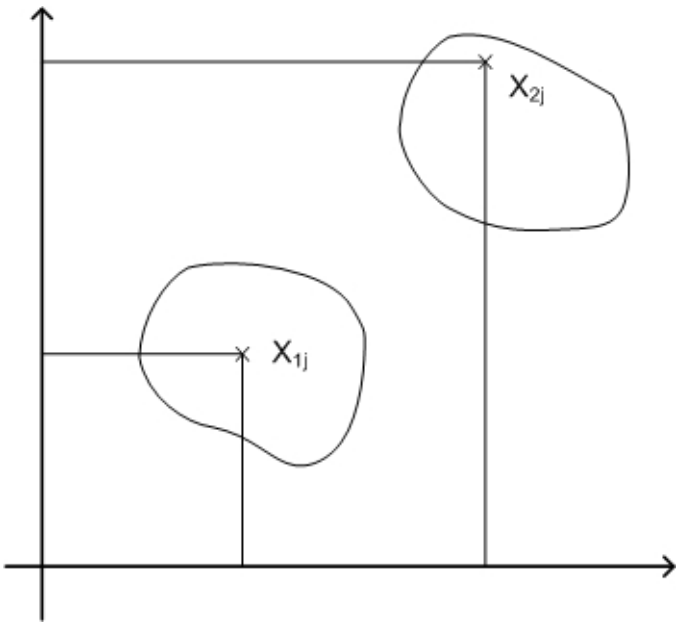

Figure 1.2. Non−overlapped projected data

In Figure 1.2, we see the projections of the original data towards the two axes form non overlapped univariate data which can be ranked. These ranks clearly identify the location difference of two random samples. Therefore selecting two axes as our directions seems to be a reasonable choice for this case. On the other hand, in Figure 1.3, which is also a graph of two independent samples, we have overlapped data as a result of projection of the original data towards to two axes. Ranks of these overlapped data do not identify the location difference correctly.

In situations like in Figure 1.3, the experimentor might prefer to add another direction in between two axes such that the projected data towards to that particlar direction are uncorrelated. For example, in the situation shown in Figure 1.3, the experimenter may prefer to use an additional direction which makes 45° angle with the horizontal axes, in addition to the directions along the main axes as shown in Figure 1.4.

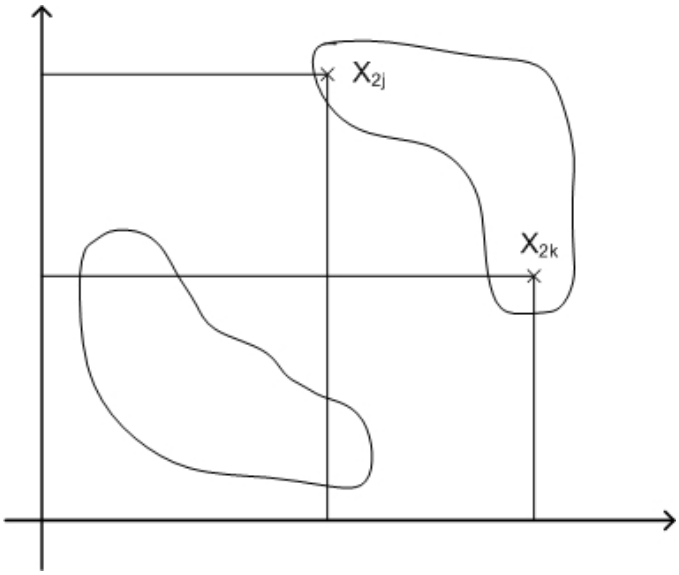

Figure 1.3. Overlapped projected data

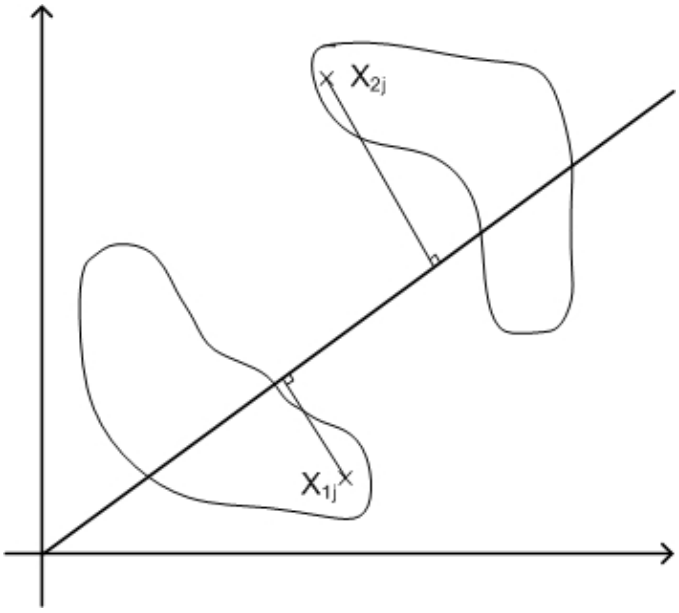

Figure 1.4. Projection towards to an arbitrary direction e

Even though in both of these situations we deal with two−dimentional data, the number of directions we had to use were two and three, respectively. So there is no clear cut rule for selecting a number of directions that we need to use in the formulation of our vector statistic. However, it would be reasonable to select at least one direction more than the dimension of the data in the samples.

CHAPTER 2
FORMULATION OF A MULTIPLE DIRECTION RANK STATISTIC

At this point let us recall some facts about the usual univariate two−sample rank statistics that are the building blocks of our multivariate statistics in (1.16). In fact before we formulate the proposing method for the multivariate problem, let us describe the univariate problem in independent notations to illustrate the basic idea within our setting. So assume that $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ are iid from $f : \mathbb{R} \longrightarrow [0, \infty)$ and $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ are iid from $g : \mathbb{R} \longrightarrow [0, \infty)$, and they are mutually independent random samples. Let $n = n_1 + n_2$.

Consider the pooled sample $Y_1, Y_2, \ldots, Y_n$ of these two samples $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ and $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ and let R_i denote the rank of Y_i for each $i = 1, 2, \ldots, n$.

Let us define a score function $\mathcal{J} : (0, 1) \longmapsto \mathbb{R}$ such that

$$\int_0^1 \mathcal{J}(t)dt = 0, \qquad \int_0^1 \mathcal{J}^2(t)dt < \infty. \tag{2.1}$$

Now we introduce regression coefficients $c_1, c_2, \ldots, c_n$ which satisfy

$$\sum_{i=1}^{n} c_i = 0, \quad \sum_{i=1}^{n} c_i^2 = 1, \quad \max_{1\le i\le n} c_i^2 \ \to 0 \quad as \ \ n \to \infty. \tag{2.2}$$

(Here we mean $c_1 = c_{n1}, c_2 = c_{n2}, \ldots, c_n = c_{nn}$, but for convenience we drop the first subscript.)

Define

$$\xi_i = F(Y_i) \tag{2.3}$$

for each $i = 1, 2, \ldots, n$. Here F is the cumulative distribution function of f and g under H_0. Note that they are iid Uniform$(0, 1)$ random variables.

Next we define the rank statistic

$$T_n = \sum_{i=1}^{n} c_i \mathcal{J}\left(\frac{R_i}{n+1}\right). \tag{2.4}$$

Recall that we wish to test the null hypothesis

$$H_0 : f = g.$$

Under this null hypothesis, asymptotic normality of T_n can be proven. To prove this result, we need the following very useful lemma.

Lemma 1. *Under the null hypothesis,*

$$\mathbb{E}\left[\sum_{i=1}^{n} c_i \left\{\mathcal{J}\left(\frac{R_i}{n+1}\right) - \mathcal{J}\left(\xi_i\right)\right\}\right]^2 \longrightarrow 0 \quad as \;\; n \longrightarrow \infty. \tag{2.5}$$

Proof. Let

$$D_i = \left\{\mathcal{J}\left(\frac{R_i}{n+1}\right) - \mathcal{J}\left(\xi_i\right)\right\} \tag{2.6}$$

for each $i = 1, 2, \ldots, n$.

Note that due to the symmetry involved, the vector $(D_1, D_2, \ldots, D_n)$ has exchangeable components under the null hypothesis. This means in particular that

$$\mathbb{E}D_i^2 = \mathbb{E}D_1^2 \;\; \forall i, \qquad \text{and} \qquad \mathbb{E}D_iD_j = \mathbb{E}D_1D_2, \;\; \forall i \neq j. \tag{2.7}$$

We now obtain,

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} c_iD_i\right)^2 &= \mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{n} c_iD_ic_jD_j \\
&= \sum_{1\leq i\neq j\leq n} c_ic_j\mathbb{E}D_iD_j + \sum_{i=1}^{n} c_i^2\mathbb{E}D_i^2 \\
&= \mathbb{E}D_1D_2\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_j - \sum_{i=1}^{n} c_i^2\right) + \mathbb{E}D_1^2\left(\sum_{i=1}^{n} c_i^2\right) \\
&= \mathbb{E}D_1D_2(0-1) + \mathbb{E}D_1^2(1) \\
&= \mathbb{E}D_1^2 - \mathbb{E}D_1D_2 \\
&\leq \mathbb{E}D_1^2 + |\mathbb{E}D_1D_2|
\end{aligned}$$

Using the Cauchy−Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{n} c_i D_i\right)^2 &\leq \mathbb{E}D_1^2 + \sqrt{\mathbb{E}D_1^2\mathbb{E}D_2^2} \\ &= 2\mathbb{E}D_1^2.\end{aligned}$$

Hence it remains to show that $\mathbb{E}D_1^2 \longrightarrow 0 \quad \text{as } n \longrightarrow \infty$, i.e.,
$\mathbb{E}\left\{\mathcal{J}\left(\frac{R_1}{n+1}\right) - \mathcal{J}\left(\xi_1\right)\right\}^2 \longrightarrow 0 \ \text{ as } n \longrightarrow \infty.$

Let us now confine the proof for the case where $\mathcal{J}(t) = t - \frac{1}{2}$, $0 \leq t \leq 1$, which contains the essential idea, and observe that

$$\begin{aligned}\mathbb{E}\left(\left(\frac{R_1}{n+1} - \frac{1}{2}\right) - \left(\xi_1 - \frac{1}{2}\right)\right)^2 &= \mathbb{E}\left(\frac{R_1}{n+1} - \xi_1\right)^2 \\ &= \mathbb{E}\mathbb{E}\left\{\left(\frac{R_1}{n+1} - \xi_1\right)^2 | R_1\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\left(\frac{R_1}{n+1} - \xi_1\right)^2 | R_1 = i\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\xi_{(i)} - \frac{i}{n+1}\right\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathrm{Var}\left(\xi_{(i)}\right)\right\} \\ &= \frac{1}{n}\sum_{j=1}^{n}\frac{j(n-j+1)}{(n+1)^2(n+2)} \\ &= \frac{1}{6(n+1)} \\ &= o\left(\frac{1}{n}\right) \quad as \ \ n \longrightarrow \infty.\end{aligned}$$

(Here $\xi_{(i)}$ is the i^{th} order statistic for $\xi_1, \xi_2, \ldots, \xi_n$.)

Therefore the result is true for $\mathcal{J}(t) = t - \frac{1}{2}$. We will not prove for the general

case. □

Theorem 1. *Under the above conditions*

$$T_n = \sum_{i=1}^{n} c_i \mathcal{J}\left(\frac{R_i}{n+1}\right) \xrightarrow{d} N\left(0, \sigma^2\right) \quad as \;\; n \longrightarrow \infty, \tag{2.8}$$

where

$$\sigma^2 = \int_0^1 \mathcal{J}^2(t)dt. \tag{2.9}$$

Proof. Define

$$\xi_i = F(Y_i) \qquad for \;\; i = 1, 2, \ldots, n, \tag{2.10}$$

and note that they are iid Uniform$(0, 1)$ random variables.

Let us introduce following rank statistic

$$\widetilde{T_n} = \sum_{i=1}^{n} c_i \mathcal{J}(\xi_i) \tag{2.11}$$

and observe that

$$\mathbb{E}\left(\widetilde{T_n}\right) = 0, \tag{2.12}$$

and

$$\mathrm{Var}\left(\widetilde{\mathrm{T_n}}\right) = \sigma^2 \tag{2.13}$$

from equations (2.1) and (2.9).

According to the Central Limit Theorem,

$$\widetilde{T_n} \xrightarrow{d} N\left(0, \sigma^2\right) \quad as \;\; n \longrightarrow \infty. \tag{2.14}$$

Now the proof is quite simple, being a straightforward application of Markov inequality. We have, for every $\epsilon > 0$,

$$
\begin{aligned}
P\left(\left|(T_n - \widetilde{T_n})\right| > \epsilon\right) &\leq \frac{\mathbb{E}\left[(T_n - \widetilde{T_n})\right]^2}{\epsilon^2} \\
&= \frac{1}{\epsilon^2}\,\mathbb{E}\left[\sum_{i=1}^{n} c_i \left\{\mathcal{J}\left(\frac{R_i}{n+1}\right) - \mathcal{J}\left(\xi_i\right)\right\}\right]^2 \longrightarrow 0 \;\; as \;\; n \longrightarrow \infty.
\end{aligned}
$$

The last step is due to Lemma 1. Hence $\left(T_n - \widetilde{T_n}\right) \xrightarrow{p} 0$, and

$$
T_n = \widetilde{T_n} + (T_n - \widetilde{T_n}) \xrightarrow{d} N\left(0, \sigma^2\right) \quad as \;\; n \longrightarrow \infty. \tag{2.15}
$$

Therefore the proof is complete. □

CHAPTER 3
ASYMPTOTICS UNDER H_0

Let us return to the multivariate two−sample problem, and let us specify more precisely the vector statistic $S = (S_{e_1}, S_{e_2}, \ldots, S_{e_k})^T$, which we call as a *multiple direction rank statistic* in our formulation. To specify the α^{th} component S_{e_α}, we consider the pooled sample $X_1, X_2, \ldots, X_n$ and their projections $X_i^T e_\alpha = X_{i,e_\alpha}$ (for $i = 1, 2, \ldots, n$) onto the direction e_α. Next let $R_{\alpha i}$ denote the rank of $X_i^T e_\alpha$ in the random sample $X_1^T e_\alpha, X_2^T e_\alpha, \ldots, X_n^T e_\alpha$.

In order to formulate the multivariate rank statistics we need some more notations. So define

$$F_\alpha(x) = \mathbb{P}\left(X^T e_\alpha \leq x\right), \qquad x \in \mathbb{R}. \tag{3.1}$$

The empirical analogue is

$$\hat{F}_\alpha(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}\left(X_i^T e_\alpha\right), \qquad x \in \mathbb{R}, \;\; \alpha = 1, 2, \ldots, k. \tag{3.2}$$

Furthermore, let

$$F_{\alpha\beta}(x, y) = \mathbb{P}\left(X^T e_\alpha \leq x, X^T e_\beta \leq y\right), \qquad (x, y) \in \mathbb{R}^2. \tag{3.3}$$

The empirical analogue is

$$\hat{F}_{\alpha\beta}(x, y) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}\left(X_i^T e_\alpha\right) I_{(-\infty,y]}\left(X_i^T e_\beta\right), \;\; (x, y) \in \mathbb{R}^2, \tag{3.4}$$

for $\alpha = 1, 2, \ldots, k$, and $\beta = 1, 2, \ldots, k$.

Then we can write

$$R_{\alpha i} = n\hat{F}_\alpha(X_i^T e_\alpha). \tag{3.5}$$

Now for each direction e_α, we define the multivariate rank statistic S_{e_α} for location as follows:

$$S_{e_\alpha} = \sum_{i=1}^{n} c_i \mathcal{J}\left(\frac{R_{\alpha i}}{n+1}\right) \tag{3.6}$$

Under the null hypothesis that $X_{11}, X_{12}, \ldots X_{1n_1}$ and $X_{21}, X_{22}, \ldots X_{2n_2}$ are iid, we have

$$S_{e_\alpha} = U_{e_\alpha} + o_p(1) \quad as \quad n \to \infty, \tag{3.7}$$

where

$$U_{e_\alpha} = \sum_{i=1}^{n} c_i \mathcal{J}\left(\xi_{\alpha i}\right) \tag{3.8}$$

with the $\xi_{\alpha i} = F_\alpha(X_i^T e_\alpha)$ being independent copies of

$$\xi_\alpha = F_\alpha(X^T e_\alpha) \sim \text{Uniform(0,1)}. \tag{3.9}$$

see, for instance [1].

Now we define our Multiple Direction Rank Statistic S as follows:

$$S = (S_{e_1}, S_{e_2}, \ldots, S_{e_k})^T \tag{3.10}$$

Theorem 2. *Under the null hypothesis (1.1) we have*

$$S = (S_{e_1}, S_{e_2}, \ldots, S_{e_k})^T \xrightarrow{d} Z \stackrel{d}{=} N_k(0, V), \tag{3.11}$$

where the right hand side is $k-$variate normal distribution with mean vector 0 and covariance matrix V with elements

$$V_{\alpha\beta} = \mathbb{E}\mathcal{J}(\xi_\alpha)\mathcal{J}(\xi_\beta) = \int\int \mathcal{J}\left(F_\alpha(x)\right)\mathcal{J}\left(F_\beta(y)\right)\, dF_{\alpha\beta}(x,y) \tag{3.12}$$

for $\alpha = 1, 2, \ldots, k$ *and* $\beta = 1, 2, \ldots, k$.

Proof. Let us introduce the random vector $U = (U_{e_1}, U_{e_2}, \ldots, U_{e_k})^T$ where U_{e_i} is a rank statistic as in (3.8) for each $i = 1, 2, \ldots, k$, and observe that we have

$$S = U + o_p(1) \quad as \quad n \to \infty \tag{3.13}$$

because of the result in equation (3.7).

Exploiting the Cramér−Wold device, let us choose an arbitrary $a \in \mathbb{R}^k$ and consider

$$a^T U = \sum_{i=1}^{n} c_i \sum_{\alpha=1}^{k} a_\alpha \mathcal{J}(\xi_{\alpha i}). \tag{3.14}$$

Since the c_i's satisfy the conditions in (2.2) and

$$\mathbb{E}\left\{\sum_{\alpha=1}^{k} a_\alpha \mathcal{J}(\xi_{\alpha i})\right\} = 0, \tag{3.15}$$

$$\mathrm{Var}\left\{\sum_{\alpha=1}^{k} a_\alpha \mathcal{J}(\xi_{\alpha i})\right\} = a^T V a, \tag{3.16}$$

it follows that

$$a^T U \xrightarrow{d} N_k(0, a^T V a) \quad as \;\; n \to \infty. \tag{3.17}$$

Hence according to equations (3.13), (3.14) and (3.17) we have

$$a^T S \xrightarrow{d} N_k(0, a^T V a) \quad as \;\; n \to \infty. \tag{3.18}$$

Because $a \in \mathbb{R}^k$ is arbitrary, the theorem follows. □

Now, we will present an example to elaborate about these formulations. The content of this example will be used for later sections in this thesis, especially for simulations we have run to investigate the accuracy of the proposed method.

Example 1.

Consider a two sample problem for testing the null hypothesis that all $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots X_{2n_2}$ are iid against the alternative hypothesis that two independent samples are from different densities.

We choose the regression coefficients as follows.

$$c_1 = \cdots = c_{n_1} = \sqrt{\frac{n_2}{nn_1}}, \quad c_{n_1+1} = \cdots = c_{n_1+n_2} = -\sqrt{\frac{n_1}{nn_2}} \tag{3.19}$$

It is easily verifed that the c_i's satisfy the conditions in (2.2) provided that $n_1 \wedge n_2 \to \infty$ as $n \to \infty$.

Let us use the score function

$$\mathcal{J}(t) = t - \frac{1}{2}, \quad 0 < t < 1, \tag{3.20}$$

which satisfies the conditions in (2.1).

In this case the multivariate rank statistic towards the direction e_α is

$$S_{e_\alpha} = \sqrt{\frac{n_2}{n\,n_1}} \sum_{i=1}^{n_1} \left(\frac{R_{\alpha i}}{n+1} - \frac{1}{2} \right) - \sqrt{\frac{n_1}{n\,n_2}} \sum_{i=n_1+1}^{n_1+n_2} \left(\frac{R_{\alpha i}}{n+1} - \frac{1}{2} \right). \tag{3.21}$$

Now we want to develop a test statistic $\|S\|^2$ in terms of the Multiple Direction Rank Statistic $S = (S_{e_1}, S_{e_2}, \ldots, S_{e_k})^T$ to test the null hypothsis (1.1). Here,

$$\|S\|^2 = S^T S. \tag{3.22}$$

In order to specify the decision rule, we should however know the exact distribution of $\|S\|^2$.

Theorem 3. *Under the* $H_0 : f = g$,

$$\|S\|^2 \xrightarrow{d} \|Z\|^2, \tag{3.23}$$

where $Z \stackrel{d}{=} N_k(0, V)$.

The distribution of $\|Z\|^2$ depends on the unknown parameter V. In particular, unless we know the rank of the matrix V, we cannot specify the distribution of $\|Z\|^2$.

If V were full rank then

$$V^{-\frac{1}{2}} S \xrightarrow{d} V^{-\frac{1}{2}} Z \stackrel{d}{=} N_k(0, I_k). \tag{3.24}$$

Hence,

$$\left\| V^{-\frac{1}{2}} S \right\|^2 \xrightarrow{d} \left\| V^{-\frac{1}{2}} Z \right\|^2 \stackrel{d}{=} \chi^2(k). \tag{3.25}$$

In practice, most of the time an estimate of the matrix V is a full rank matrix with rank equal to the number of directions we consider. This fact is being confirmed by most of the Matlab simulations we have run, in order to check the validity of the proposed method. The Matlab code of the simulation program for two dimensional data is shown in the Appendix.

As another justification for the non−degeneracy of the covariance matrix in a different manner, let us consider the Example 1 given that our data are projected only to directions e_α and e_β. In this situation our covariance matrix is as follows.

$$V = \begin{bmatrix} \mathbb{E}\mathcal{J}^2(\xi_\alpha) & \mathbb{E}\mathcal{J}(\xi_\alpha)\mathcal{J}(\xi_\beta) \\ \mathbb{E}\mathcal{J}(\xi_\alpha)\mathcal{J}(\xi_\beta) & \mathbb{E}\mathcal{J}^2(\xi_\beta) \end{bmatrix} \tag{3.26}$$

Recall that ξ_α and ξ_β are Uniform$(0, 1)$, and observe that

$$\mathbb{E}\mathcal{J}^2(\xi_\alpha) = \int_0^1 \left(\xi_\alpha - \frac{1}{2} \right)^2 d\xi_\alpha = \frac{1}{12}. \tag{3.27}$$

Similarly,

$$\mathbb{E}\mathcal{J}^2(\xi_\beta) = \frac{1}{12}. \tag{3.28}$$

Next notice that

$$\begin{aligned}
\mathbb{E}\mathcal{J}(\xi_\alpha)\mathcal{J}(\xi_\beta) &= \mathbb{E}\left[\left(\xi_\alpha - \frac{1}{2}\right)\left(\xi_\beta - \frac{1}{2}\right)\right] \\
&= \mathbb{E}\left[\xi_\alpha\xi_\beta - \frac{1}{2}(\xi_\alpha + \xi_\beta) + \frac{1}{4}\right] \\
&= \mathbb{E}[\xi_\alpha\xi_\beta] - \frac{1}{4} \qquad \left(\because \mathbb{E}[\xi_\alpha] = \mathbb{E}[\xi_\beta] = \frac{1}{2}\right).
\end{aligned}$$

Now if V is not a full rank matrix, we have

$$\begin{aligned}
\begin{vmatrix} \frac{1}{12} & \mathbb{E}[\xi_\alpha\xi_\beta] - \frac{1}{4} \\ \mathbb{E}[\xi_\alpha\xi_\beta] - \frac{1}{4} & \frac{1}{12} \end{vmatrix} &= 0 \\
\left(\frac{1}{12}\right)^2 - \left(\mathbb{E}[\xi_\alpha\xi_\beta] - \frac{1}{4}\right)^2 &= 0 \\
\left(\frac{1}{12} + \mathbb{E}[\xi_\alpha\xi_\beta] - \frac{1}{4}\right)\left(\frac{1}{12} - \mathbb{E}[\xi_\alpha\xi_\beta] + \frac{1}{4}\right) &= 0 \\
\left(\mathbb{E}[\xi_\alpha\xi_\beta] - \frac{1}{6}\right)\left(-\mathbb{E}[\xi_\alpha\xi_\beta] + \frac{1}{3}\right) &= 0,
\end{aligned}$$

i.e.,

$$\mathbb{E}[\xi_\alpha\xi_\beta] = \frac{1}{6} \quad \text{or} \quad \mathbb{E}[\xi_\alpha\xi_\beta] = \frac{1}{3}. \tag{3.29}$$

Since the covariance matrix V is nonnegative definite matrix, $|V|$ should be greater than or equal to zero. Therefore according to the above calculation, $\mathbb{E}[\xi_\alpha\xi_\beta]$ can only take values between $\frac{1}{3}$ and $\frac{1}{6}$ including the two end points. Only this two end points are resulting in $|V| = 0$. Therefore it is reasonable to assume that V is a full rank matrix except for a very rare situation. The reasonability of this claim can be convinced through the following integral calculations. Notice that the possible domain for the joint density function of ξ_α and ξ_β is the unit square. First we

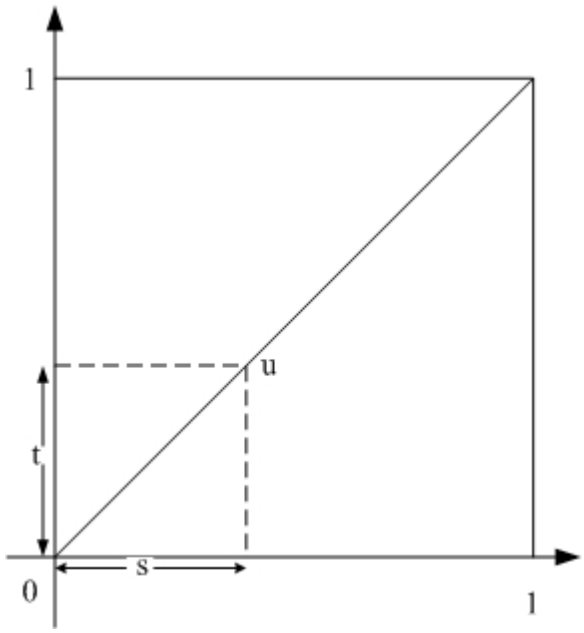

Figure 3.1. The Main Diagonal of the Unit Square

assume the domain of the joint density function of ξ_α and ξ_β is the main diagonal of the unit square. The integration on the main diagonal results in $\mathbb{E}[\xi_\alpha \xi_\beta] = \frac{1}{3}$.

$$\begin{aligned}
\mathbb{E}[\xi_\alpha \xi_\beta] &= \int\int st\, dF_{\alpha\beta}(s,t) \\
&= \frac{1}{\sqrt{2}} \int_0^{\sqrt{2}} \frac{u}{\sqrt{2}} \frac{u}{\sqrt{2}}\, du \\
&= \frac{1}{2\sqrt{2}} \left[\frac{u^3}{3}\right]_0^{\sqrt{2}} \\
&= \frac{1}{3}
\end{aligned}$$

Next we assume the domain of the joint density function of ξ_α and ξ_β is the other diagonal of the unit square. The integration on this diagonal results in $\mathbb{E}[\xi_\alpha \xi_\beta] = \frac{1}{6}$.

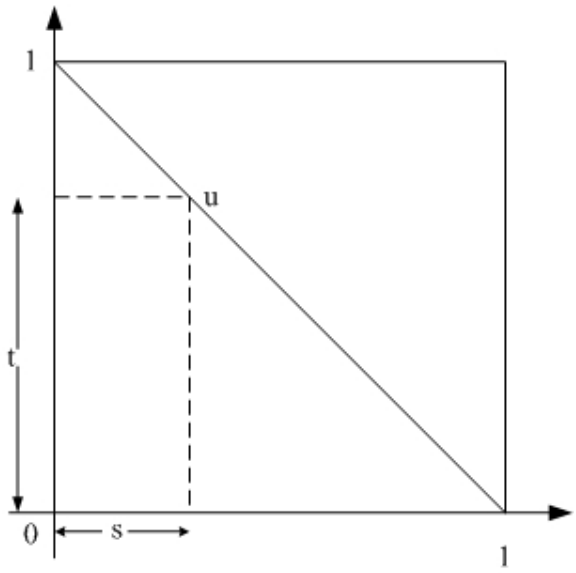

Figure 3.2. The Other Diagonal of the Unit Square

$$\begin{aligned}
\mathbb{E}\left[\xi_\alpha \xi_\beta\right] &= \int\int st\, dF_{\alpha\beta}(s,t) \\
&= \frac{1}{\sqrt{2}} \int_0^{\sqrt{2}} \frac{u}{\sqrt{2}} \left(1 - \frac{u}{\sqrt{2}}\right) du \\
&= \frac{1}{2\sqrt{2}} \left[\frac{\sqrt{2}u^2}{2} - \frac{u^3}{3}\right]_0^{\sqrt{2}} \\
&= \frac{1}{6}
\end{aligned}$$

There are situations where the estimate of the matrix V is not a full rank matrix. So let r be the rank of V, i.e.

$$\text{rank}\,(\text{V}) = \text{r} \leq \text{k}. \tag{3.30}$$

Lemma 2. *Since the matrix V is positive semidefinite, there exists an orthonormal $k \times k$ matrix O and positive numbers $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_r > 0$ $(1 \leq r \leq k)$ such that $V = O^T \Delta O$ where Δ is the $k \times k$ diagonal matrix with diagonal $(\delta_1, \delta_2, \cdots, \delta_r, 0, 0, \cdots, 0)$.*

Define $\Delta^{-1/2}$ to be the diagonal matrix with diagonal $(\delta_1^{-1/2}, \delta_2^{-1/2}, \cdots, \delta_r^{-1/2}, 0, 0, \cdots, 0)$. The following theorem is immediate from Theorem 2 and the continuous mapping theorem.

Theorem 4. *Under the null hypothesis, we have*

$$\left\|\Delta^{-1/2} OS\right\|^2 \to \left\|\Delta^{-1/2} OZ\right\|^2 \stackrel{d}{=} \chi^2(r), \text{ as } n \to \infty, \tag{3.31}$$

where $Z \stackrel{d}{=} N_k(0, V)$.

However, in practice the covariance matrix V will be unknown. For the element $V_{\alpha\beta}$ of V, however, $\sqrt{n}-$consistent estimators can be constructed by replacing the true unknown cdf's in the expression for $V_{\alpha\beta}$ (see (3.12) and (3.9)) by their emperical counterparts. This yields,

$$\hat{V}_{\alpha\beta} = \int\int \mathcal{J}\left(\frac{n}{n+1}\hat{F}_\alpha(x)\right) \mathcal{J}\left(\frac{n}{n+1}\hat{F}_\beta(y)\right) d\hat{F}_{\alpha\beta}(x, y) \tag{3.32}$$

i.e.,

$$\hat{V}_{\alpha\beta} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{J}\left(\frac{R_{\alpha i}}{n+1}\right) \mathcal{J}\left(\frac{R_{\beta i}}{n+1}\right). \tag{3.33}$$

where we assume $\mathcal{J}$ to be bounded as in (2.1). In fact the expression on the right of (3.33) is a rank statistic for testing independence of Spearman type. It is known that after proper centering and scaling with $\sqrt{n}$, these statistics are asymptotically normal under fixed "alternatives", i.e. arbitrary continuous joint cdf's $F_{\alpha\beta}$, hence $\sqrt{n}-$consistent, see [2].

Matrix $\hat{V}$ is also positive semidefinite and can be diagonalized by an orthonormal $k \times k$ matrix $\hat{O}$; let $\hat{\Delta}$ be the diagonal matrix with diagonal $(\hat{\delta}_1, \hat{\delta}_2, \cdots, \hat{\delta}_{\hat{r}}, 0, 0, \cdots, 0)$ for some $1 \leq \hat{r} \leq k$, where $\hat{\delta}_i$'s are eigenvalues of $\hat{\Delta}$. It follows from results in multivariate analysis (see, for instance, [3]) that $\hat{O}$ is an

$\sqrt{n}$-consistent estimator of O, and that each $\hat{\delta}_i$ is an $\sqrt{n}$-consistent estimator of δ_i; also $\hat{r} \to r$ as $n \to \infty$. This entails the following conjecture.

Conjecture 1. *Under the null hypothesis we have*

$$\left\| \hat{\Delta}^{-1/2} \hat{O} S \right\|^2 \xrightarrow{d} \chi^2(r) \text{ as } n \to \infty. \tag{3.34}$$

Since r, the number of degrees of freedom on the right in (3.34) is unknown, in practice we reject the null hypothesis of iid observations against the two-sample alternative if

$$\left\| \hat{\Delta}^{-1/2} \hat{O} S \right\|^2 > \chi^2_{1-\alpha}(\hat{r}), \tag{3.35}$$

where S is the vector of two sample statistics in (3.10), and $\chi^2_{1-\alpha}(\hat{r})$ is the quantile of order $1-\alpha$ of the chi-square distribution with $\hat{r}$ degrees of freedom.

CHAPTER 4
ASYMPTOTICS UNDER ALTERNATIVES

4.1 Contiguity

Properties of statistical procedures are typically derived locally, i.e., under the assumption that a given distribution is the true underlying distribution. To derive such properties usually a kind of local smoothness will be required, in the sense of some stability under small changes of the underlying distributions. Closeness of probability distributions can be expressed as "contiguity". The most common sufficient requirement for contiguity is the "local asymptotic normality" of the log likelihood ratio which also expresses the desired smoothness.

Now we will show the contiguity of our model by proving the asymptotic normality of the log likelihood ratio. Recall that we are dealing with two samples $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ which are in $\mathbb{R}^\nu$. These two samples are from densities f and g, respectively. Here we wish to test

$$H_o : f = g = h(.) \quad \text{vs.} \quad H_1 : f = h(.) \quad \text{and} \quad g = h\left(. - \frac{\gamma}{\sqrt{n_2}} e\right), \tag{4.1}$$

where γ is a scalar and e is a fixed unit vector in $\mathbb{R}^\nu$.

The likelihood ratio is

$$\begin{aligned} \lambda_e &= \frac{\max \mathrm{L}(e)_{model}}{\max \mathrm{L}(e)_{null\ hypothesis}} \\ &= \frac{\prod_{j=1}^{n_1} h(X_{1j}) \prod_{j=1}^{n_2} h\left(X_{2j} - \frac{\gamma}{\sqrt{n_2}} e\right)}{\prod_{j=1}^{n_1} h(X_{1j}) \prod_{j=1}^{n_2} h(X_{2j})} \\ &= \prod_{j=1}^{n_2} \frac{h\left(X_{2j} - \frac{\gamma}{\sqrt{n_2}} e\right)}{h(X_{2j})}. \end{aligned} \tag{4.2}$$

Let $\Lambda = \log \lambda_e$ and then,

$$\Lambda = \sum_{j=1}^{n_2} \left[\ln h \left(X_{2j} - \frac{\gamma}{\sqrt{n_2}} e \right) - \ln h(X_{2j}) \right]. \tag{4.3}$$

Now, letting $\log h = \ell$ and using Taylor series expansion, we obtain

$$\Lambda = \frac{-\gamma}{\sqrt{n_2}} \sum_{j=1}^{n_2} \left[\sum_{k=1}^{\nu} \frac{\partial}{\partial k} \ell(X_{2j}) e_k \right] + \frac{\gamma^2}{2n_2} \sum_{j=1}^{n_2} \left[\sum_{s=1}^{\nu} \sum_{t=1}^{\nu} e_s \frac{\partial^2}{\partial s \partial t} \ell(X_{2j}) e_t \right] + r_{n_2} \tag{4.4}$$

with $r_{n_2} \to 0$ as $n_2 \to \infty$. (Here $\frac{\partial}{\partial k}$ denotes the partial derivative with respect to k^{th} coordinate and $\frac{\partial^2}{\partial s \partial t}$ denotes the partial derivative with respect to coordinates with indexes t and s.)

Let us re-write the random variable Λ as follows

$$\Lambda = T_1 + T_2 + r_{n_2}, \tag{4.5}$$

where

$$T_1 = \frac{-\gamma}{\sqrt{n_2}} \sum_{j=1}^{n_2} \left\{ \sum_{k=1}^{\nu} \frac{\partial}{\partial k} \ell(X_{2j})\, e_k \right\}, \tag{4.6}$$

$$T_2 = \frac{\gamma^2}{2n_2} \sum_{j=1}^{n_2} \left\{ \sum_{s=1}^{\nu} \sum_{t=1}^{\nu} e_s \frac{\partial^2}{\partial s \partial t} \ell(X_{2j})\, e_t \right\}. \tag{4.7}$$

Now we will do few simple calculations before specifying the limiting distribution of Λ.

Remark: The Fisher−information matrix of the family of densities $h(. - \theta)$, $\theta \in \mathbb{R}^\nu$, is given as the $\nu \times \nu$ matrix with entries

$$\mathbb{E} \frac{\partial}{\partial j} \log h(X - \theta) \cdot \frac{\partial}{\partial k} \log h(X - \theta). \tag{4.8}$$

Here $\dfrac{\partial}{\partial j}$ and $\dfrac{\partial}{\partial k}$ denote the partial derivatives with respect to j^{th} and k^{th} coordinates of $\theta \in \mathbb{R}^\nu$. Also it should be noted that here the expectation $\mathbb{E}$ is with respect to the density $h(. - \theta)$.

We need this Fisher−information matrix at $\theta = 0$, but as the following calculations shows, the matrix doesn't depend on θ.

$$\begin{aligned}
\mathbb{E}\frac{\partial}{\partial j}\log h(X-\theta)\cdot\frac{\partial}{\partial k}\log h(X-\theta) &= \int \frac{1}{h(x-\theta)}\frac{\partial}{\partial j}\ h(x-\theta)\cdot\frac{\partial}{\partial k}\ h(x-\theta)\,dx \\
&= \int \frac{1}{h(y)}\frac{\partial}{\partial j}\ h(y)\cdot\frac{\partial}{\partial k}\ h(y)\,dy \quad (\text{letting } y = x-\theta) \\
&= \mathbb{E}\left(\frac{\partial}{\partial j}\log h(X), \frac{\partial}{\partial k}\log h(X)\right) \\
&= I_{j,k}.
\end{aligned} \tag{4.9}$$

In fact the last expectation is with respect to the density $h(.-0) = h(.)$. Therefore this Fisher−information matrix with entries $I_{j,k}$ will simply be denoted by I. From now on all the expectations and variances are with respect to the density $h(.)$.

Using the notation ∇ for the gradient, written as a column, the Fisher information matrix can be written as

$$\mathbb{E}\left(\nabla \log h(X)\right)\left(\nabla \log h(X)\right)^T = I. \tag{4.10}$$

Also note that

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2}{\partial j \partial k}\log h(X)\right] &= \int \frac{\partial}{\partial j}\left[\frac{\frac{\partial}{\partial k}h(x)}{h(x)}\right]\ h(x)\,dx \\
&= \int \frac{\left[h(x)\ \frac{\partial}{\partial j}\frac{\partial}{\partial k}h(x) - \frac{\partial}{\partial j}h(x)\ \frac{\partial}{\partial k}h(x)\right]}{[h(x)]^2}\,h(x)\,dx \\
&= \int \frac{\partial}{\partial j}\frac{\partial}{\partial k}\,h(x)\,dx - \int \frac{\frac{\partial}{\partial k}h(x)\ \frac{\partial}{\partial j}h(x)}{h(x)}\,dx \\
&= 0 - \int \frac{\frac{\partial}{\partial k}h(x)\ \frac{\partial}{\partial j}h(x)}{h(x)}\,dx \\
&= -I_{j,k}.
\end{aligned} \tag{4.11}$$

Using the notation H for the Hessian (i.e. the matrix of mixed second partial derivatives), we may express this result as

$$\mathbb{E}\left(H \log h(X)\right) = -I. \tag{4.12}$$

For simplicity, from now onwards let us write Y instead of X_{2j} in our calculations. Note that under $H_0 : f = g = h$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{\nu} \frac{\partial}{\partial k}\ell(Y)e_k\right] &= \sum_{k=1}^{\nu} e_k \mathbb{E}\frac{\partial}{\partial k}\ell(Y) \\
&= \sum_{k=1}^{\nu} e_k \int \frac{\frac{\partial}{\partial k}h(y)}{h(y)} h(y)\, dy \\
&= \sum_{k=1}^{\nu} e_k \frac{\partial}{\partial k} \underbrace{\int h(y)\, dy}_{=1} \\
&= 0.
\end{aligned} \tag{4.13}$$

This shows that $\mathbb{E}(T_1) = 0$. Also because of (4.13) we have

$$\begin{aligned}
\mathrm{Var}\left(\sum_{k=1}^{\nu} \frac{\partial}{\partial k}\ell(Y)e_k\right) &= \mathbb{E}\left[\sum_{k=1}^{\nu} \frac{\partial}{\partial k}\ell(Y)e_k\right]^2 \\
&= \mathbb{E}\left[\sum_{s=1}^{\nu}\sum_{t=1}^{\nu} e_s \frac{\partial}{\partial s}\ell(Y)\frac{\partial}{\partial t}\ell(Y)\, e_t\right] \\
&= \sum_{s=1}^{\nu}\sum_{t=1}^{\nu} e_s \mathbb{E}\left[\frac{\partial}{\partial s}\ell(Y)\frac{\partial}{\partial t}\ell(Y)\right] e_t \\
&= \sum_{s=1}^{\nu}\sum_{t=1}^{\nu} e_s \left[\int \frac{\frac{\partial}{\partial s}h(y)}{h(y)} \frac{\frac{\partial}{\partial t}h(y)}{h(y)}\, h(y)\, dy\right] e_t \\
&= \sum_{s=1}^{\nu}\sum_{t=1}^{\nu} e_s I_{st} e_t \\
&= e^T I e.
\end{aligned} \tag{4.14}$$

Now consider (again under $H_0 : f = g = h$),

$$\begin{aligned}\mathbb{E}\left[\sum_{s=1}^{\nu}\sum_{t=1}^{\nu} e_s \frac{\partial^2}{\partial s \partial t}\ell(Y) e_k\right] &= \sum_{s=1}^{\nu}\sum_{t=1}^{\nu} e_s \mathbb{E}\left[\frac{\partial^2}{\partial s \partial t}\ell(Y)\right] e_t \\ &= \sum_{s=1}^{\nu}\sum_{t=1}^{\nu} e_s \left(-I_{st}\right) e_t \qquad \text{(see (4.10))} \\ &= -e^T I e.\end{aligned} \tag{4.15}$$

Using above calculation and sticking with familiar notation again,

$$\begin{aligned}\mathrm{Var}(T_1) &= \frac{\gamma^2}{n_2}\sum_{j=1}^{n_2} \mathrm{Var}\left[\sum_{k=1}^{\nu}\frac{\partial}{\partial k}\ell(X_{2j}) e_k\right] \\ &= \frac{\gamma^2}{n_2}\sum_{j=1}^{n_2} e^T I e \\ &= \gamma^2 e^T I e,\end{aligned} \tag{4.16}$$

because the X_{2j}'s are iid. Therefore in combination with the Central Limit Theorem we obtain

$$T_1 \xrightarrow{d} N(0, \gamma^2 e^T I e). \tag{4.17}$$

Next let us observe that T_2 is an average of iid variables, where each term has expectation

$$\begin{aligned}\mathbb{E}\left[\frac{\gamma^2}{2}\sum_{s=1}^{\nu}\sum_{t=1}^{\nu} e_s \frac{\partial^2}{\partial s \partial t}\ell(X_{2j})\, e_t\right] &= \frac{\gamma^2}{2}\sum_{s=1}^{\nu}\sum_{t=1}^{\nu} e_s\, \mathbb{E}\left[\frac{\partial^2}{\partial s \partial t}\ell(X_{2j})\right] e_t \\ &= -\frac{\gamma^2}{2} e^T I e,\end{aligned} \tag{4.18}$$

because the X_{2j}'s are iid. Hence the Weak Law of Large Numbers yields

$$T_2 \xrightarrow{p} -\frac{\gamma^2}{2} e^T I e. \tag{4.19}$$

We have now proved the following theorem.

Theorem 5. *The limiting distribution of* Λ *under the null hypothesis is*

$$\Lambda \xrightarrow{d} N\left(-\frac{\gamma^2}{2}e^T Ie, \gamma^2 e^T Ie\right). \tag{4.20}$$

It should be noted that due to the special relation between mean and variance of Λ, Theorem 5 entails contiguity of our model.

4.2 Asymptotic Power

Under the alternatives described above we can determine the asymptotic power of the test described in Example 1 in Chapter 3. We will only give a brief sketch here. First of all, because

$$\sum_{i=1}^{n} \frac{R_{\alpha i}}{n+1} = \frac{1}{2}n, \tag{4.21}$$

the test statistic S_{e_α} in (3.21) becomes

$$S_{e_\alpha} = -\sqrt{\frac{n}{n_1 n_2}} \sum_{j=1}^{n_2} \left(\frac{R_{\alpha(n_1+j)}}{n+1} - \frac{1}{2}\right), \tag{4.22}$$

It follows at once that S_{e_α} equivalent to

$$S_{e_\alpha}^{'} = \frac{1}{\sqrt{n_2}} \sum_{j=1}^{n_2} \left(\frac{R_{\alpha(n_1+j)}}{n+1} - \frac{1}{2}\right), \tag{4.23}$$

where the $R_{\alpha(n_1+1)}, \ldots, R_{\alpha(n_1+n_2)}$ are the ranks of the elements of the second sample within the pooled sample.

Also recall that we have shown in (2.5) that, under H_0, we have

$$S_{e_\alpha}^{'} = T_{e_\alpha} = \frac{1}{\sqrt{n_2}} \sum_{j=n_1+1}^{n_1+n_2} \left(\xi_{\alpha j} - \frac{1}{2}\right) + o_p(1) \qquad as \quad n_2 \to \infty, \tag{4.24}$$

so that for asymptotics under the null hypothesis we may replace $S_{e_\alpha}^{'}$ with T_{e_α}.

In order to find the asymptotic distribution of $(S_{e_1}^{'}, \ldots, S_{e_k}^{'})$ under the local alternatives in (4.1), according to LeCam's third lemma (see, for instance, [6]) it suffices to show that

$$\left(S_{e_1}^{'}, \ldots, S_{e_k}^{'}, \Lambda\right)^T \overset{d(0)}{\to} N_{k+1}\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right), \quad as \quad n_1 \wedge n_2 \to \infty, \quad (4.25)$$

under the null hypothesis, for some $\mu \in \mathbb{R}^k$, $\tau \in \mathbb{R}^k$, a $k \times k$ covariance matrix Σ, and $\sigma^2 > 0$. The aforementioned lemma then entails that

$$\left(S_{e_1}^{'}, \ldots, S_{e_k}^{'}\right)^T \overset{d(1)}{\to} N_k\left(\mu + \tau, \Sigma\right), \quad as \quad n_1 \wedge n_2 \to \infty, \quad (4.26)$$

under the local alternatives. Also (4.25) entails that, under the null hypothesis, Λ tends to a $N\left(-\frac{1}{2}\sigma^2, \sigma^2\right)$ distribution, so that this result includes Theorem 5.

The proof is almost immediate from the multivariate Central Limit Theorem for iid sequences, because we can replace the $S_{e_\alpha}^{'}$ with the T_{e_α}, and because Λ is also a sum of iid random variables (all are based on the same iid sequence of vectors in the second sample) apart from a term of lower order.

Once (4.26) has been established, the limiting distribution of the actual test statistic, which is based on

$$\sum_{\alpha=1}^{k}\left(S_{e_\alpha}^{'}\right)^2, \quad (4.27)$$

can be derived and will be of a noncentral chi−square type. Hence the asymptotic power of the test can be determined.

CHAPTER 5
SIMULATIONS

After our formulation of the theory, we decided to develop a Matlab simulation program (the Matlab code is listed in the appendix) to demonstrate our results for two−dimensional data. The main objective behind running these type of simulations was to investigate the validity of our results.

Simulations were done for bivariate normal random samples with sizes 100 and 1000, and using 2 directions (the two main axes) and 3 directions (including the direction which has 45° angle with the x axis apart from two main axes). The first samples which we have created as the control smaple in our simulation, was randomly generated from the distribution $N_2((2,2)^T, I_2)$. The second sample was generated using a bivariate normal population which has the same covariance structure, but with a small shift in the mean compared to the mean in the first sample. After running our simulation for these randomly generated samples 100 times, we calculated the estimated power of our test for 2 and 3 directions, respectively. The results of the simulation for random samples with size 100 is shown in the Table 5.1. In Table 5.2 we have listed the simulation results for sample size equal to 1000.

The results of these two simulations show that our testing procedure is capable of identifying the location difference in the distributions of these random samples reasonably well, even for samples with size 100. In fact, Table 5.1 shows that we can achieve the asymptotic power of the test up to the satisfactory level, when there is a noticably difference in the distributions. Especially when we throw in one additional direction it always improves the asymptotic power of the test. For example, the test for two bivariate normal samples with mean $(2,2)^T$ and $(2.3,2.3)^T$ has achieved the asymptotic power .75 for 2 directions and .85 for 3 directions. Anyway, according to the Table 5.2, it can be seen for larger samples like the samples with size 1000, our testing procedure produce phenomenal results by identifying even the smaller shifts of the distributions of these random samples. For example, the tests for two bivariate random samples with mean $(2,2)^T$ and $(2.1,2.1)^T$ have reached the

Table 5.1. Estimated power calculations for sample with size 100

Mean of Sample 1	Mean of Sample 2	No. of Directions	No. of Tests	No. of Rejections	Estimated Power
$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}2.05\\2.05\end{pmatrix}$	2	100	5	.05
		3	100	22	.22
$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}2.1\\2.1\end{pmatrix}$	2	100	14	.14
		3	100	25	.25
$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}2.2\\2.2\end{pmatrix}$	2	100	34	.34
		3	100	56	.56
$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}2.3\\2.3\end{pmatrix}$	2	100	75	.75
		3	100	85	.85

asymptotic power .76 for 2 directions and .87 for 3 directions, which may be the maximum they can be for a practical problem. Therefore, argubaly these simulations provides strong evidence for the accuracy of our testing procedure.

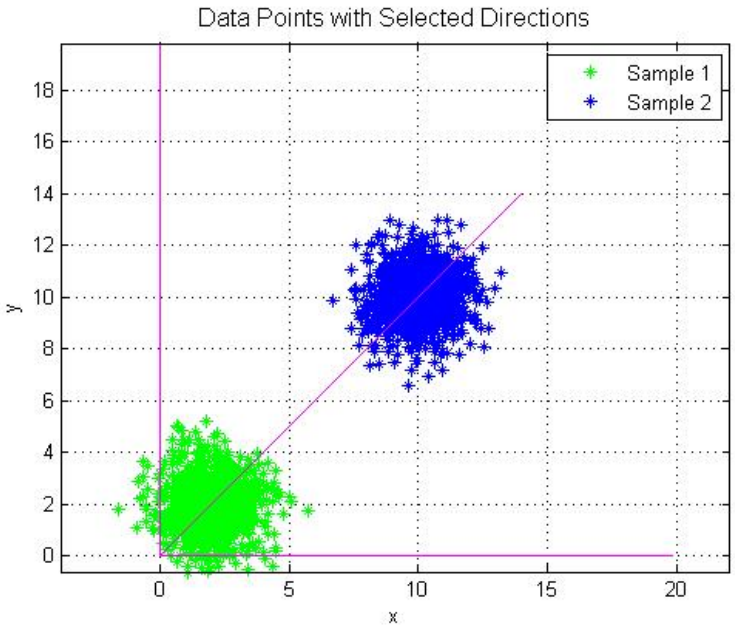

Figure 5.1. Two samples from different distibutions

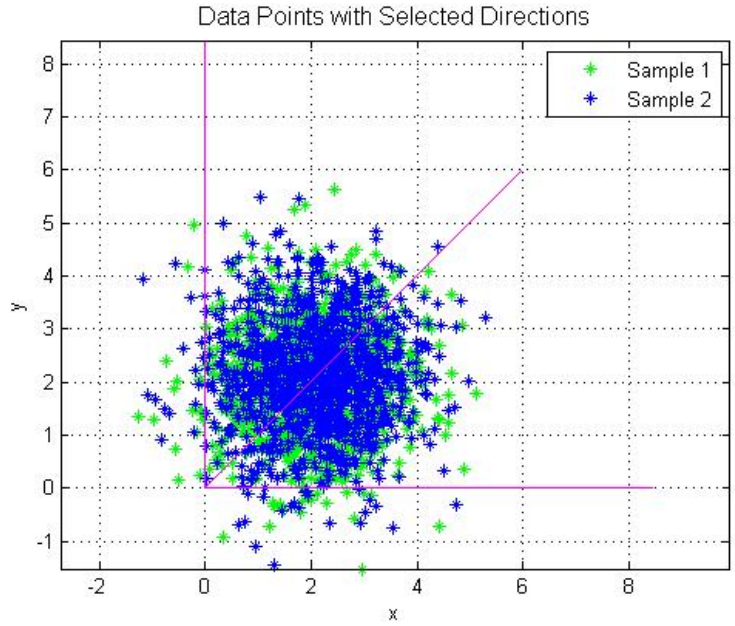

Figure 5.2. Two samples from identical distibutions

Table 5.2. Estimated power calculations for sample with size 1000

Mean of Sample 1	Mean of Sample 2	No. of Directions	No. of Tests	No. of Rejections	Estimated Power
$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}2.05\\2.05\end{pmatrix}$	2	100	27	.27
		3	100	25	.25
$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}2.1\\2.1\end{pmatrix}$	2	100	76	.76
		3	100	87	.87
$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}2.2\\2.2\end{pmatrix}$	2	100	99	.99
		3	100	100	1
$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}2.3\\2.3\end{pmatrix}$	2	100	100	1
		3	100	100	1

CHAPTER 6
CONCLUSIONS AND FUTURE WORKS

Even though our method gives us the desired result as a powerful procedure to test whether two multivariate sample data come from the same underlying distribution, there is room for further improvement. In particular, although we were proposing to use a finite number of directions to formulate a Multiple Direction Rank Statistic, we didn't provide criteria for selection of these directions. Also the rank of the covariance matrix V requires further investigation. Furthermore, my brief sketch to determine the asymptotic power of the test using contiguity needs to be completed. In other words, the construction of a comprehensive theory still requires some more work. In the future, I would like to investigate these areas.

BIBLIOGRAPHY

[1] Hájek, P., and Šidák, Z. (1967). "Theory of rank tests", 1st edition, *New York, Academic Press*.

[2] Ruymgaart, F.H., Shorack, C.G., and van Zwet, W.R. (1972). "Asymptotic normality of nonparametric tests for independence", *Ann. Math. Statistics.* **43**, 1122−1135.

[3] Watson, G.S. (1983). "Statistics on Spheres", 1st edition, *Wiley, New York*.

[4] Casella, G., and Berger, R.L. (2001). "Statistical inference", 2nd edition, *Duxbury Press*.

[5] Johnson, R.A., and Wichern, D.W. (2006). "Applied multivariate statistical analysis", 5th edition. *Pearson, Prentice Hall*.

[6] van der Vaart, A.W. (1998). "Asymptotic Statistics", 1st edition, *Cambridge*.

[7] Buhrman, H., and Ruymgaart, F.H. (1981). "An application of linearization in nonparametric multivariate analysis", *Sankhyā A.* **43**, 52−66.

APPENDIX

This is the Matlab code for simulation program we used to check the validity of our method for two dimensional data.

The Main Program

```
clc
clear
close all

n = 100 ;
mean1 = 2 ;
sd = 1 ;

proj_angles = [0 90] ;

mean2 = 2.3 ;

disp('Test Statistic       Chi2 Value       Decision')
disp('==============       ==========       ========')

reject_count = 0 ;
for i = 1:100
    [test_value, chi2value] =
              statCode(proj_angles,n,mean1,mean2,sd,0);
    if chi2value < test_value
        reject_count = reject_count + 1 ;
        decision = 'Reject' ;
    else
        decision = 'Accept' ;
    end
    disp([num2str(test_value) [' ']
    num2str(chi2value) ['']  decision])
end

disp([['Reject Count  = '] num2str(reject_count)] )
```

The Sub Program

```
function [test_value, chi2value] =
                  statCode(proj_angles,n,mean1,mean2,sd,graph_ops)

if nargin == 0
    proj_angles = [0 45 90];  % Projection angles in degrees
    n = 1000 ; % Number of data points
    mean1 = 2 ; % Mean of the first data set
    mean2 = 2 ; % Mean of the second data set
    sd = 1 ; % Standard deviation (same for both data sets)
    graph_ops = 1 ; % 1 to show graph, 0 to hide graph
end

proj_angles = proj_angles * pi / 180 ;
            % Convert projection directons to radians

    num_dir = length(proj_angles) ;
            % Number of directions used

% Read and plot the data (Data in 2 columns)

        data1 =mean1 +  sd*randn(n,2) ;
        data2 =mean2 +  sd*randn(n,2) ;

     if graph_ops == 1
           figure(1) ; % Open new figure window for plotting
        %Plot the data
         plot(data1(:,1),data1(:,2),'g*',data2(:,1),data2(:,2),'b*');
        axis 'equal' ; % Set equal scaling in x and y axes
        grid on ; % Switch on the grig lines of the plot
        legend('Sample 1','Sample 2') ; % Add legend

         % Add title
        title('Data Points with Selected Directions','fontsize',12);
        xlabel('x') ; % Label of the x-axis
        ylabel('y') ; % Label of the y-axis
     end

    t = max(abs([data1(:) ; data2(:)])) * 1.5 ;
```

```
                        % Length of the direction line segment

% Projection of data to arbitrary directions
    i =1 ; % Initialize counter
    for theta = proj_angles
        direction(i,:) = [cos(theta) , sin(theta)] ;
                          % Compute projection directions
        i=i+1 ; % Increment counter

        if graph_ops == 1
            figure(1) ; % Set plot handle to figure 1
            hold on ;
            plot([0 t*cos(theta)],[0 t*sin(theta)],'m') ;
                            % Plot the directions
        end
    end
    if graph_ops == 1
        hold off ;
    end

    data1_proj  = direction*data1' ; % Projection of data set 1
    data2_proj  = direction*data2' ; % Projection of data set 2

    data_proj = [data1_proj , data2_proj] ;
                        % Merge the projected data

    data_proj_sorted = sort(data_proj')' ;
                        % Sort the data into ascending order

    [r,c] = size(data_proj_sorted) ;
                        % Get the size of the merged data set

    rank_matrix = zeros(size(data_proj_sorted)) ;
                        % Initialize the rank matrix

    % Compute the ranks using the empirical analogue
    for i = 1:r
        for j = 1:c
            rank_matrix(i,j) =
              sum((data_proj_sorted(i,:) <= data_proj(i,j))');
```

```
        end
    end

% Display the rank matrix
    rank_matrix = rank_matrix ;

% Determining regression coefficients
    data1_size = length(data1) ; % Size of sample 1
    data2_size = length(data2) ; % Size of sample 2
    total_size = data1_size + data2_size ; % Total size
    data1_constant = ones(1,data1_size)
                   * sqrt(data2_size/(total_size*data1_size)) ;

    data2_constant =
    -1*ones(1,data2_size)* sqrt(data1_size/(total_size*data2_size));

% Display regression coefficients
    constant_vector = [data1_constant data2_constant] ;

% Rank matrix
    % Score functions
    score_functions = (rank_matrix / (total_size + 1)) - 0.5 ;

    % Multiple direction rank statistic
    direction_stat = constant_vector * score_functions' ;

    % Estimated covariance matrix by empirical analogue
    v_matrix = (score_functions * score_functions' )/(total_size);

    % Rank of the V-matrix
    rank_of_v = rank(v_matrix) ;

    % Eigenvalues and eigenvectors
    [eig_vec , eig_val] = eig(v_matrix) ;
    o_matrix = eig_vec ;
    delta_matrix = eig_val ;

    % To find the generalized inverse ...
        zero_thresh = 0.0025 - eps ;
        unzero_mask = delta_matrix <= zero_thresh ;
```

```
        zero_mask = delta_matrix > zero_thresh ;

    inverse_delta_matrix =
          (1./(unzero_mask + delta_matrix)) .* zero_mask ;

%    disp(['Test value using ' num2str(num_dir) ' dirctions'])
        test_value =
        direction_stat * o_matrix' *
             inverse_delta_matrix * o_matrix * direction_stat';

    chi2value = chi2inv(0.95,rank(inverse_delta_matrix)) ;

    if nargout == 0
        chi2value = chi2value
        test_value = test_value
    end
```

PERMISSION TO COPY

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Texas Tech University or Texas Tech University Health Sciences Center, I agree that the Library and my major department shall make it freely available for research purposes. Permission to copy this thesis for scholarly purposes may be granted by the Director of the Library or my major professor. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my further written permission and that any user may be liable for copyright infringement.

Agree (Permission is granted.)

Unawatuna Gamage Asiri Gunathilaka	July 12, 2007
Student Signature	Date

Disagree (Permission is not granted.)

Student Signature	Date