A STATISTICAL TEST SPECTRUM - FROM ROBUST TO POWERFUL

BY SOMNATH MUKHERJEE

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics

Written under the direction of

Prof. Kesar Singh

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2009

ABSTRACT OF THE DISSERTATION

A Statistical Test Spectrum - From Robust To Powerful

by Somnath Mukherjee

Dissertation Director: Prof. Kesar Singh

The concept of Scale Curve provides a graphical tool for analysis of multivariate data, with a broad range of statistical applications. Recent research in variants of Scale Curves have shown great promise, as they can be easily adapted to build robust non-parametric testing procedures under various scenarios, while preserving good power, and retaining the crucial virtues of easy computation and simple graphical representation.

This thesis investigates the properties of one such variant of Scale Curves, named the Determinant Scale Curve (*dsc*). It is shown that the *dsc* can be used to devise non-parametric exact tests for location of multivariate data with a special property (stated in next paragraph), under both one sample and multi-sample setups. Similar ideas are extended to tackle problems in linear regression, where the *dsc* is used to build tests for significance of the slope parameter.

For all the problems discussed, the *dsc*'s actually provide a whole spectrum of tests. The tests at the rightmost end of the spectrum are shown to be Pitman equivalent to the benchmark most powerful tests for the given problem. As one moves towards the other end, the corresponding tests become progressively more and more robust, i.e. insensitive to outliers. Simulation results show that this robustification does not come with a serious loss of power under most situations.

Applications of the *dsc* as an exploratory tool are also discussed. It is shown to be useful for investigating tail properties of a data distributions and identifying presence of linearity in multivariate data. The results are very encouraging, and suggest wider applicability of similar techniques.

Acknowledgements

I extend my sincerest gratitude to Prof. Kesar Singh, for providing me careful guidance through this research. To me, he is more like a father figure - someone I have always looked up to, whenever in doubt.

I thank the members of the defence committee for their careful reading of the thesis.

Words are not enough to express the influence that my classmates at the Indian Statistical Institute have had on me. They are an incredible group of people, who have never failed to inspire me. I owe a lot to them.

But more than anything else, what made this thesis possible are the many sacrifices made by my parents and my wife. I am, and will remain forever indebted to them.

Dedication

To my parents, my sister, and my wife.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The notion of scale curves was introduced by Liu, Parelius and Singh in their paper [Liu, Parelius, Singh 1999]. A scale curve describes how scale evolves from the center for a multivariate dataset. For $0 \leq t \leq 1$, [Liu, Parelius, Singh 1999] defines the scale curve s(t) as the volume of $100t\%$ central data. Here, the centrality is characterized by a concept of data-depth. The properties of the curve would of course depend on the statistic used to measure the data-depth. Among the well known choices are the Half-Space depth by Tukey, the Oja depth, and the Simplicial Depth, [Liu, R. 1990]. With a suitably chosen data-depth, the scale curve can yield valuable insights into the nature of the distribution. This is especially true in the case of high-dimensional data, where visualizing the spread of the data without the help of a graphical tool like the scale curve can prove to be difficult. The theoretical properties of the scale curve are, however, difficult to ascertain, as shown by Serfling, [Serfling, R. 2002].

A significant simplification of the of the computational problems associated with the scale curves was provided by [Singh, Tyler, Zhang, Mukherjee] (to appear), which defines a variant of the scale curve, called the *Quantile Scale Curve* (*qsc*). For $0 \leq t \leq 1$ the qsc is defined as: q(t) = the $100t^{th}$ percentile of the volumes of all simplices created by the data-points as the vertices. As shown in [Singh, Tyler, Zhang, Mukherjee], the *qsc* can be useful as a tool for detecting linear and non-linear associations between groups of variables, using some proposed graphical tests. Other problems that were addressed using the *qsc* were testing for multivariate location and scale, and exploring heavy-tailedness.

This thesis will explore another variant of the scale curve, which we will call the *Determinant Scale Curve* (*dsc*), which has got some special appeal. It will be shown

that, like the *qsc*, the *dsc* can also be used for a wide range of testing purposes, while retaining the nice properties of easy computation and simple graphical interpretation. Furthermore, it will be established that the graphical tests derived at the upper end of the test-plots of the *dsc*'s are Pitman equivalent to the most powerful tests in many common scenarios. This fact makes the testing procedures developed using the *dsc* far more attractive. The graphical tests using *dsc* are shown to have optimality properties in many of the testing scenarios considered in [Singh, Tyler, Zhang, Mukherjee], as well as some additional important problems in multivariate analysis.

Before formally defining *dsc*, we must introduce a few notations and definitions.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. p-variate data, where $X_i' = \{X_{i1}, X_{i2}, \ldots, X_{ip}\}, 1 \leq i \leq n$ are $p \times 1$ row vectors.

In this presentation, $\mathbf{X}$ will denote the $p \times n$ data-matrix, whose i^{th} column is $X_i, 1 \leq i \leq n$. The mean vector is $\bar{X}' = \{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p\}$, where $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, 1 \leq j \leq p$. Also, define the sum of squares matrix $SS(\mathbf{X}) = \{s_{ij}\}_{p\times p}$ as

$$s_{ij} = \sum_{k=1}^{n} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j),\ 1 \leq i, j \leq p$$

Finally, the matrix $\mathbf{X}_{(-\mathbf{i})}$ will denote the sub-matrix of $\mathbf{X}$ with the i^{th} column X_i deleted. $\bar{X}_{(-i)}$ and $SS(\mathbf{X}_{(-\mathbf{i})})$ should be interpreted analogously to $\mathbf{X}_{(-\mathbf{i})}$.

We can now go on to define the Determinant Scale Curve and explore some of its properties.

Definition 1.1. *For a data matrix* $\mathbf{X}$ *, the Determinant Scale Curve* $d(t),\ 0 \leq t \leq 1$ *is defined as follows:*

$$d(1) = \sqrt{|SS(\mathbf{X})|},$$
$$d\left(\frac{n-1}{n}\right) = \sqrt{\min_{1\leq i\leq n} \left|SS(\mathbf{X}_{(-\mathbf{i})})\right|}.$$

Suppose, for a given sample, the above minimum is attained at $\mathbf{X}_{(-\mathbf{i})}$*. Then define,*

$$d\left(\frac{n-2}{n}\right) = \sqrt{\min_{1\leq j\leq n, j\neq i} SS([\mathbf{X}_{(-\mathbf{i})}]_{(-\mathbf{j})})}.$$

This sequential elimination process is continued to define $d(\frac{k}{n})$*,* $1 \leq k \leq n$*.*

Finally, having defined $d(.)$ *at the nodes* $\left(\frac{k}{n}\right), 0 \leq k \leq n$*, we define the rest of the*

function as lines joining the nodes: for $\frac{k}{n} \leq t \leq \frac{k+1}{n}, 0 \leq k \leq (n-1)$*, i.e.*

$$d(t) = d\left(\frac{k}{n}\right) + (nt - k)\left(d\left(\frac{k+1}{n}\right) - d\left(\frac{k}{n}\right)\right).$$

1.1 Properties of the Determinant Scale Curve (dsc)

One would like any scale curve to be monotonically non-decreasing, since, addition of points to the outer regions of a data-cloud is expected to only increase the data volume. The following results would prove that *dsc* does indeed satisfy this monotonicity property.

Lemma 1.1. *For any data matrix* $\mathbf{X}$*, we have* $|SS(\mathbf{X})| \geq \left|SS(\mathbf{X}_{(-\mathbf{i})})\right|$

Proof. Observe that by partitioning $\mathbf{X}$ into $\left(\mathbf{X}_{(-\mathbf{i})} : X_i\right)$, one can get the identity

$$|SS(\mathbf{X})| = |(\mathbf{X}_{(-\mathbf{i})} - \bar{X}1'_{(n-1)})(\mathbf{X}_{(-\mathbf{i})} - \bar{X}1'_{(n-1)})' + (X_i - \bar{X})(X_i - \bar{X})'|, \tag{1.1}$$

where 1_k is the column vector of length k with all entries equal to 1.
Now, notice that

$$\bar{X} = \frac{(n-1)\bar{X}_{(-i)} + X_i}{n} = \bar{X}_{(-i)} + \frac{X_i - \bar{X}_{(-i)}}{n}, \tag{1.2}$$

and,

$$\mathbf{X}_{(-\mathbf{i})} - \bar{X}1'_{(n-1)} = \mathbf{X}_{(-\mathbf{i})} - \bar{X}_{(-i)}1'_{(n-1)} + \frac{(X_i - \bar{X}_{(-i)})1'_{(n-1)}}{n}.$$

Therefore,

$$\begin{aligned}(\mathbf{X}_{(-\mathbf{i})} - \bar{X}1'_{(n-1)})(\mathbf{X}_{(-\mathbf{i})} - \bar{X}1'_{(n-1)})' =& (\mathbf{X}_{(-\mathbf{i})} - \bar{X}_{(-i)}1'_{(n-1)})(\mathbf{X}_{(-\mathbf{i})} - \bar{X}_{(-i)}1'_{(n-1)})' \\ &+ \frac{(n-1)(X_i - \bar{X})(X_i - \bar{X})'}{n^2}\end{aligned} \tag{1.3}$$

Also, by 1.2,

$$X_i - \bar{X} = \frac{(n-1)(X_i - \bar{X}_{(-i)})}{n}. \tag{1.4}$$

Using 1.3 and 1.4 in 1.1, we get,

$$
\begin{aligned}
|SS(\mathbf{X})| &= \left| (\mathbf{X}_{(-\mathbf{i})} - \bar{X}_{(-i)}1'_{(n-1)})(\mathbf{X}_{(-\mathbf{i})} - \bar{X}_{(-i)}1'_{(n-1)})' + \frac{(n-1)(X_i - \bar{X})(X_i - \bar{X})'}{n} \right| \\
&= \left| SS(\mathbf{X}_{(-\mathbf{i})}) + \frac{(n-1)(X_i - \bar{X})(X_i - \bar{X})'}{n} \right| \\
&= \left| SS(\mathbf{X}_{(-\mathbf{i})}) \right| \left[1 + \frac{(n-1)(X_i - \bar{X})'(SS(X_{(-i)}))^{-1}(X_i - \bar{X})}{n} \right]
\end{aligned}
$$

The last equality follows from the fact that for any $k \times k$ non-singular matrix A and $k \times 1$ vector b, we have,

$$
|A| \left| 1 + b'A^{-1}b \right| = \left| \begin{pmatrix} A & -b \\ b' & 1 \end{pmatrix} \right| = \left| A + bb' \right|
$$

Since $SS(\mathbf{X}_{(-\mathbf{i})})^{-1}$ is positive definite, and $n \geq 2$, the lemma is proved. □

The lemma immediately provides us with the following result.

Theorem 1.2. *$d(.)$ is monotonically non-decreasing.*

Proof. For any data matrix $\mathbf{X}$ and $1 \leq k \leq n$, suppose $d\left(\frac{k}{n}\right)$ is attained at the submatrix $\mathbf{U}$ of order $p \times k$ of $\mathbf{X}$. Using the lemma, we have:

$$
d^2\left(\frac{k-1}{n}\right) = \min_{1 \leq i \leq k} \left| SS(\mathbf{U}_{(-\mathbf{i})}) \right| \leq \left| SS(\mathbf{U}_{(-\mathbf{k})}) \right| \leq |SS(\mathbf{U})| = d^2\left(\frac{k}{n}\right),
$$

which proves that $d(.)$ is non-decreasing at the nodes $\left(\frac{k}{n}\right), 0 \leq k \leq n$. Since $d(.)$ is linear between any two successive nodes, the theorem is proved. □

Theorem 1.2 essentially establishes that, even though the scheme to define $d(.)$ stipulates dropping one data row at a time, one can actually drop more than one row of data at any given step. The resulting curve will be a scale curve in its own right and will inherit most of the properties of the original curve. However, dropping more and more points at any given step reduces the number of nodes, and reduces the amount of information captured in the curve, making it less effective for analytical purposes. When at least two data-rows are dropped in a step, it is possible to view the result in Lemma 1.1 as a special case of well known results in multivariate analysis. To see

this, lets partition the original data matrix $\mathbf{X}$ into $\left(\mathbf{X1}_{(\mathbf{p}\times\mathbf{n_1})}, \mathbf{X2}_{(\mathbf{p}\times\mathbf{n_2})}\right), n_1+n_2=n$, where $\mathbf{X1}_{(\mathbf{p}\times\mathbf{n_1})}$ is the part that is retained and $\mathbf{X2}_{(\mathbf{p}\times\mathbf{n_2})}$ is the part that is dropped. Now, one can consider the two partitions as data coming from two populations, and, using a standard result in multivariate analysis, one can split the variance of the data into "within-group" and "between group" sum-of-squares, i.e.,

$$SS(\mathbf{X}) = \mathbf{W} + \mathbf{B}$$

where $W = SS(\mathbf{X1}) + SS(\mathbf{X2})$ is the "within group" sum-of-squares, and B is the "between group" sum-of-squares. Therefore,

$$|SS(\mathbf{X})| = |W| \left|I + W^{-1}B\right|$$

Now, in a two-sample problem like this, we have

$$\left|I + W^{-1}B\right| = 1 + \frac{n_1 n_2}{n} h' W^{-1} h \geq 1$$

where h is the difference in the means of the two samples. The last inequality holds since W is n.n.d. Therefore,

$$\begin{aligned} |SS(\mathbf{X})| &\geq |W| \\ &= |SS(\mathbf{X1}) + SS(\mathbf{X2})| \\ &\geq |SS(\mathbf{X1})| + |SS(\mathbf{X2})| \\ &\geq |SS(\mathbf{X1})|. \end{aligned}$$

The second inequality above follows from the fact that both $SS(\mathbf{X1})$ and $SS(\mathbf{X2})$ are n.n.d. matrices.

Another desirable property of a scale curve is affine equivariance, defined as follows:

Definition 1.2. *A statistic $U(\mathbf{X})$ is affine equivariant, if, for all $p \times p$ matrices $\mathbf{A}$ and $p \times n$ matrices $\mathbf{B}$, we have*

$$U(\mathbf{AX} + \mathbf{B}) = |\mathbf{A}| U(\mathbf{X})$$

This property essentially ensures that a rotation and/or a translation of the axes does not affect the scale curve. The following theorem confirms that *dsc* does, indeed, satisfy this property.

Theorem 1.3. *For any data matrix* $\mathbf{X}$*, and* $\mathbf{Y} = \mathbf{A}\times\mathbf{X}+\mathbf{B}$ *, for* $p \times p$ *matrices* $\mathbf{A}$ *and* $p \times n$ *matrices* $\mathbf{B}$*, we have*

$$d_Y(t) = abs(|\mathbf{A}|)d_X(t), \ for\ all\ 0 \leq t \leq 1.$$

Therefore, dsc is affine equivariant.

Proof. First, notice that for any $0 \leq i \leq n$, we have

$$\mathbf{Y}_{(-\mathbf{i})} = \mathbf{A} \times \mathbf{X}_{(-\mathbf{i})} + \mathbf{B}_{(-\mathbf{i})}.$$

Also,

$$|SS(\mathbf{Y})| = |SS(\mathbf{A} \times \mathbf{X} + \mathbf{B})| = |\mathbf{A}|^2 |SS(\mathbf{X})| \implies d_Y(1) = abs(|\mathbf{A}|)d_X(1).$$

Since $|\mathbf{A}|^2 \geq 0$, we have, for $0 \leq i \neq j \leq n$,

$$\left|SS(\mathbf{X}_{(-\mathbf{i})})\right| \leq \left|SS(\mathbf{X}_{(-\mathbf{j})})\right| \implies \left|SS(\mathbf{Y}_{(-\mathbf{i})})\right| \leq \left|SS(\mathbf{Y}_{(-\mathbf{j})})\right|.$$

Therefore, if $d_X\left(\frac{n-1}{n}\right)$ is attained at the submatrix $\mathbf{X}_{(-\mathbf{i})}$, $0 \leq i \leq n$, then $d_Y\left(\frac{n-1}{n}\right)$ will be attained at $Y_{(-i)}$. Hence,

$$d_Y\left(\frac{n-1}{n}\right) = \sqrt{\left|SS(\mathbf{Y}_{(-\mathbf{i})})\right|} = abs(|\mathbf{A}|)d_X\left(\frac{n-1}{n}\right).$$

Since, the above argument can be repeated, replacing $\mathbf{Y}$ and $\mathbf{X}$ by $\mathbf{Y}_{(-\mathbf{i})}$ and $\mathbf{X}_{(-\mathbf{i})}$ respectively, we can conclude

$$d_Y\left(\frac{n-2}{n}\right) = abs(|\mathbf{A}|)d_X\left(\frac{n-2}{n}\right).$$

Continuing similarly, we have, $d_Y\left(\frac{k}{n}\right) = abs(|\mathbf{A}|)d_X\left(\frac{k}{n}\right)$ for $0 \leq k \leq n$. Since, *dsc* is defined linearly in between the nodes $\left(\frac{k}{n}\right), 0 \leq k \leq n$, the theorem is proved. □

Figure 1.1 shows the *dsc* for pairs of test scores data provided by [Mardia, Kent, Bibby 1979]. [Liu, Parelius, Singh 1999] provides the data-depth scale curves for this data. [Singh, Tyler, Zhang, Mukherjee] shows the quantile scales curves for the same data. All the three types of scale curves seem to indicate the same features of the data. The *dsc* however, seems to accentuate the difference in scale of the groups more strongly, as compared to *qsc*.

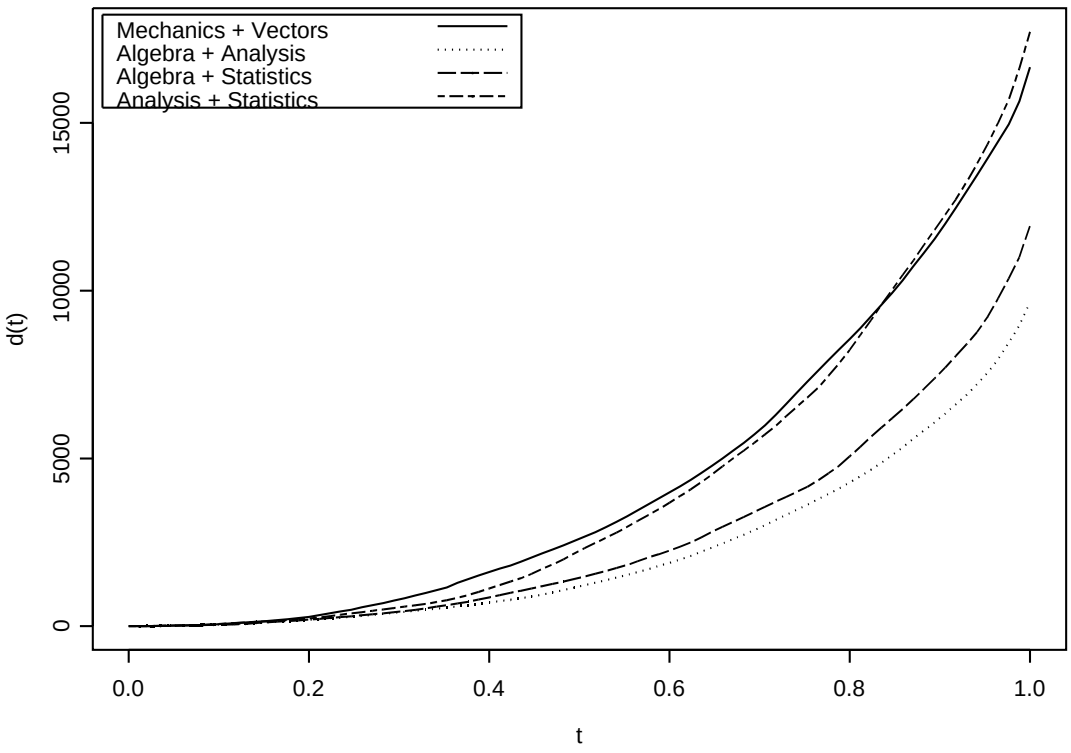


Figure 1.1: Determinant Scale Curves for Test Scores data

1.2 Tailedness

The Determinant Scale Curve is a characteristic determined by the distribution of the data. A distribution that is concentrated around its center is expected to have a flatter scale curve compared to one that is more diffused. The reason is that the far outlying points in a diffused distribution would contribute to a significant increase the scale of the data. For a concentrated distribution however, this increase in scale, even with the addition of the points that are farthest from the center, would be comparatively less pronounced.

Figure 1.2 shows a comparison of *dsc* of samples taken from some standard bivariate distributions. As expected, the scale curve for standard Cauchy distribution is

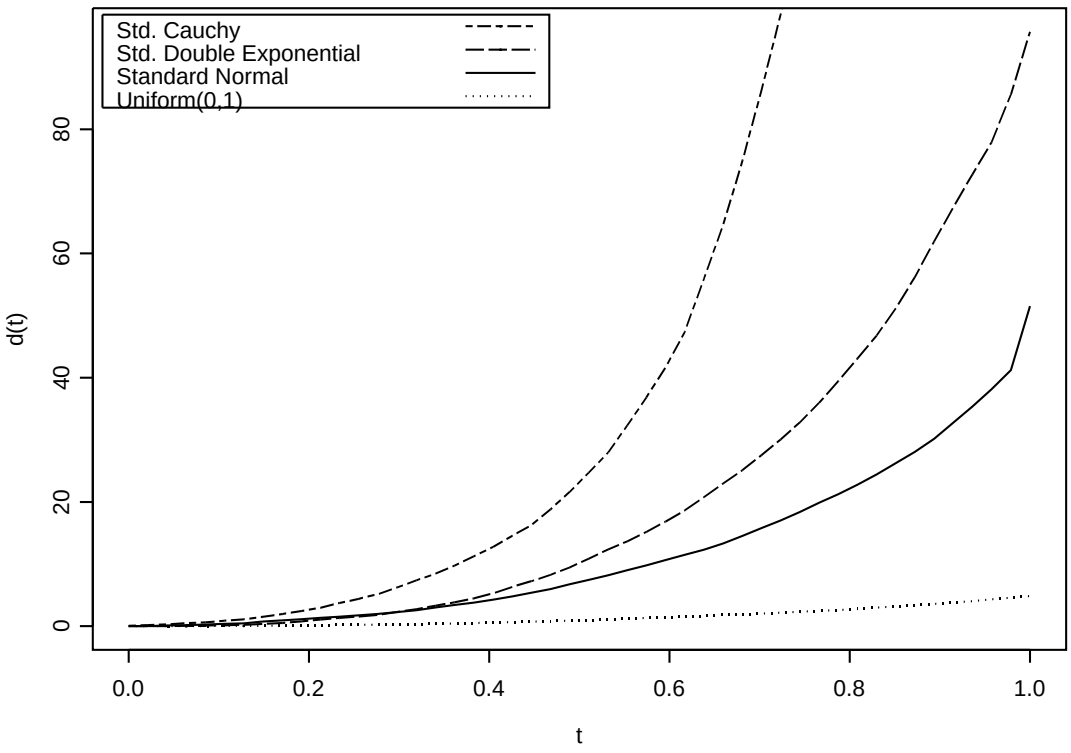


Figure 1.2: Determinant Scale curves for different distributions

much more steeper compared to that of standard Normal distribution, whereas, Double Exponential falls somewhere in between. The Uniform(0,1) distribution is a lot more concentrated than the others and have a largely flat scale curve.

This above property indicates that the *dsc* can be used as a tool to measure how heavy-tailed the data-distribution is, when compared to some of the common distributions. The steeper the curve is, specially towards the end, the more heavy tailed the distribution is expected to be. One should however make a scale adjustment, to make the comparisons meaningful. Thus, we define a measure of "*tailedness*" as follows:

Definition 1.3. *The Determinant Scale Curve based Tailedness Curve is defined as:*

$$T(t) = \frac{d(t)}{d(0.5)},\ 0.5 \leq t \leq 1$$

Figure 1.3 shows the simulated Tailedness Curves for some bivariate distributions. Even though the component variances of $N(0, 10I)$ are higher than that of Standard

Double Exponential, the tails of the latter are heavier than that of the former, and the curves capture this fact nicely. The Cauchy distribution, as expected, has the heaviest tails among the ones considered.

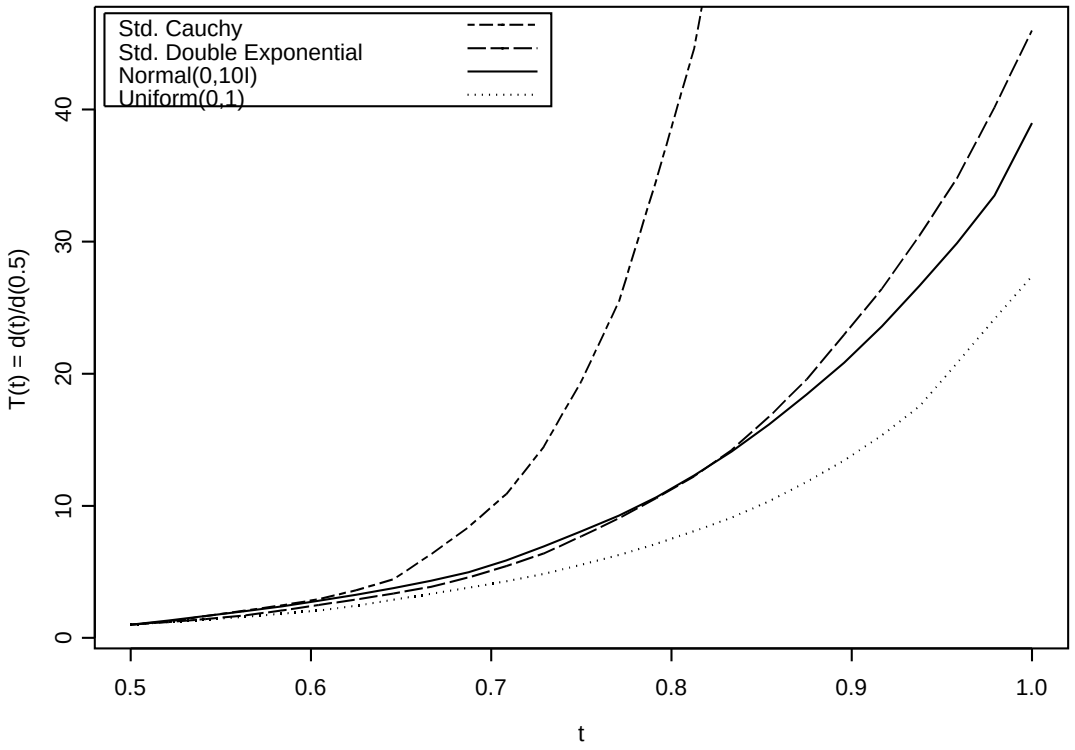


Figure 1.3: Tailedness Curves for different distributions

1.3 Applications of *dsc*

In the following chapters, we will look into a wide range of possible applications of the Determinant Scale Curve. In chapter 2, the *dsc* will be combined with a reflection scheme to develop tests for location of multivariate datasets. Chapter 3 will deal with the problems involving comparisons of means of multiple multivariate samples. A multi-sample version of the *dsc* will be introduced, and the idea will be combined with a permutation scheme to develop non-parametric tests. Finally, chapter 4 will be devoted

to investigating problems in linear regression and developing tests for the regression parameters using the *dsc*. In all these developments, an appropriate optimality property will be established for the 'end point' test corresponding to $d(1)$.

Chapter 2

Determinant Scale Curve based Test Plots for Location

The most fundamental and commonly encountered problems in statistics involve questions regarding the means of populations, or, in general, the location parameters of the underlying distributions of the populations. Often, the statistician would be interested in testing whether a population is centered around a hypothesized value. Statistical literature is rich in treatment of these problems, both in the parametric and non-parametric settings. See [Mardia, Kent, Bibby 1979] for parametric and [Hettmansperger et al 1994], [Hettmansperger et al 1997], among others, for non-parametric multivariate location testing. [Gelman et al 1995] provides detailed analysis of estimation and testing problems on the centrality parameter from a Bayesian point of view. In this chapter, we will explore a robust non-parametric technique to tackle the aforementioned problems, using the Determinant Scale Curve. It will be shown that the *dsc* will produce a testing procedure at every point of its domain. Moreover, the test at the rightmost extreme of the domain will be shown to have optimality properties which make them comparable to the most powerful tests under the standard parametric testing setting. Also, the tests get progressively more robust as one moves inwards from the right. These methods would make use of a "reflection principle", which will be discussed below.

2.1 A Reflection Principle and the One-Sample Location Problem

Let $\{X_1, X_2, \ldots, X_n\}$ be a sample on $\Re^p$, from a distribution $\mathbf{F}$, which is assumed to be symmetric around the parameter μ. Let $\{\tau_i\}_{\{i=1,\ldots,n\}}$ be i.i.d. Bernoulli random variables taking values 0 and 1, each with probability 0.5. Define the "reflected sample"

as $\{X_i{}^*\}_{\{i=1,\ldots,n\}}$, where

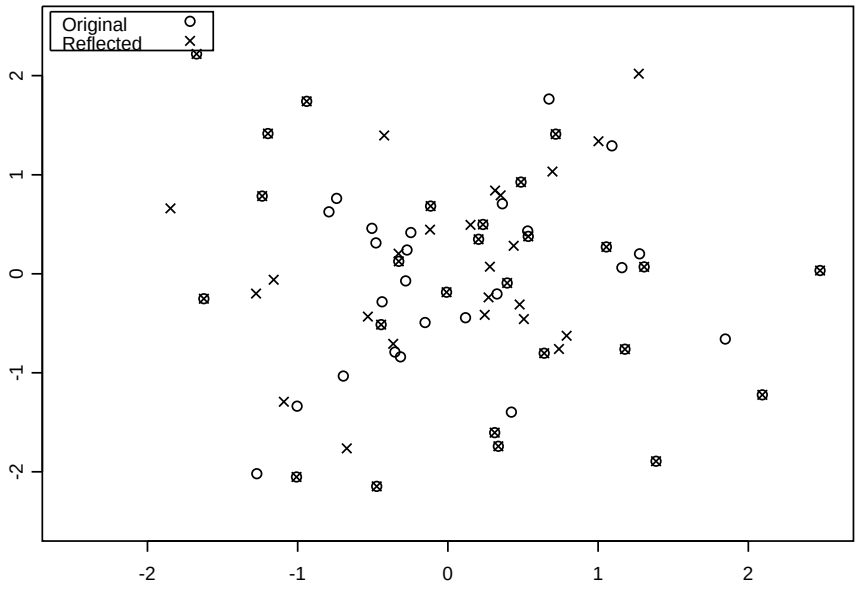
$$X_i{}^* = (2\tau_i - 1)\, X_i + 2\,(1 - \tau_i)\,\mu \tag{2.1}$$

Effectively, we are tossing an unbiased coin, and depending on which side comes up, we are either leaving the data-point untouched, or, reflecting the data-point across the assumed point of symmetry μ. We will denote the *dsc* of this reflected dataset by $d^*(t)$. If the reflection is done on or close to the true point of symmetry, the overall spread of the data will not change appreciably. Consequently, if we draw the $dsc's$ of the original and reflected data - $d(t)$ and $d^*(t)$, on the same graph, the two curves will not be significantly apart. On the other hand, if the reflections are done across a point that is away from the point of symmetry for the data, the resulting data will be much more diffused. As a result, $d^*(t)$ would much steeper, and the right tail of $d^*(t)$ would lie entirely above $d(t)$. This phenomenon is illustrated in Figure 2.1 and Figure 2.2.

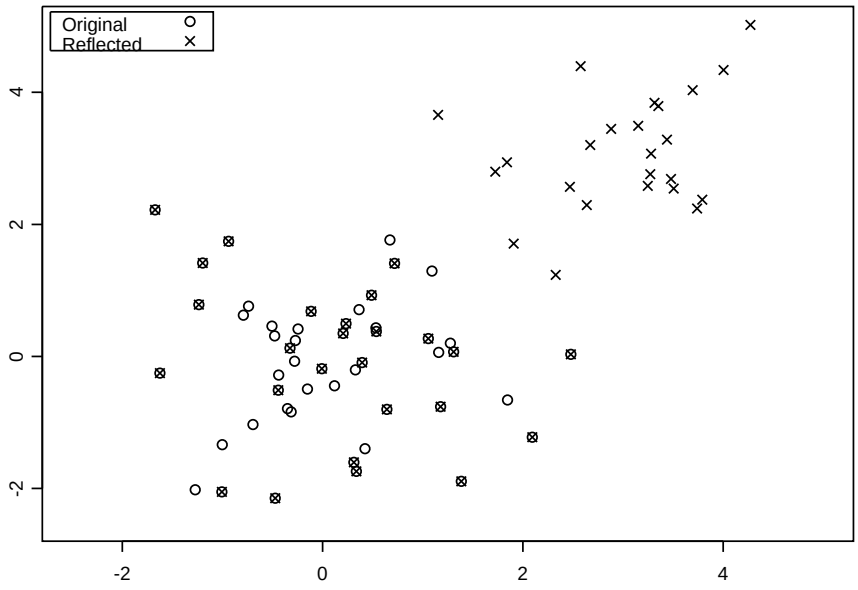
The above principle directs us towards a testing procedure for the one sample location problem in a natural way. Suppose, our interest lies in testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. One can generate a a large number of replicates of $d^*(t)$ by reflecting the data across the null value μ_0. A practical choice for the number of replicates would be 100. For the purpose of doing the test at level α, we reject the null hypothesis if $d(t)$ lies in or under the bottom $100\alpha\%$ of the $d^*(t)'s$. Note here that t can be chosen to be any fixed number within its domain $0 \leq t \leq 1$, and for every distinct value of t, we have a separate test.

Even though, there can be infinitely many choices for t, we restrict ourselves to $t = 0.25, 0.5, 0.75, 1$ only, since these values, apart from representing a natural order of dropping a quarter of the data at a time, also suffice in demonstrating the properties of the tests across the domain of t. Also, note that these are permutation type tests, and hence, by design, the tests are exact at each t.

Table 2.1 shows the power comparisons of the proposed test with the Hotelling T^2 test, which is the benchmark test under (approx.) Normality assumptions. As the numbers show, the test at $d(1)$ behaves similar to Hotelling T^2, for the Normal and Double Exponential data. However, for Cauchy, which does not have finite second

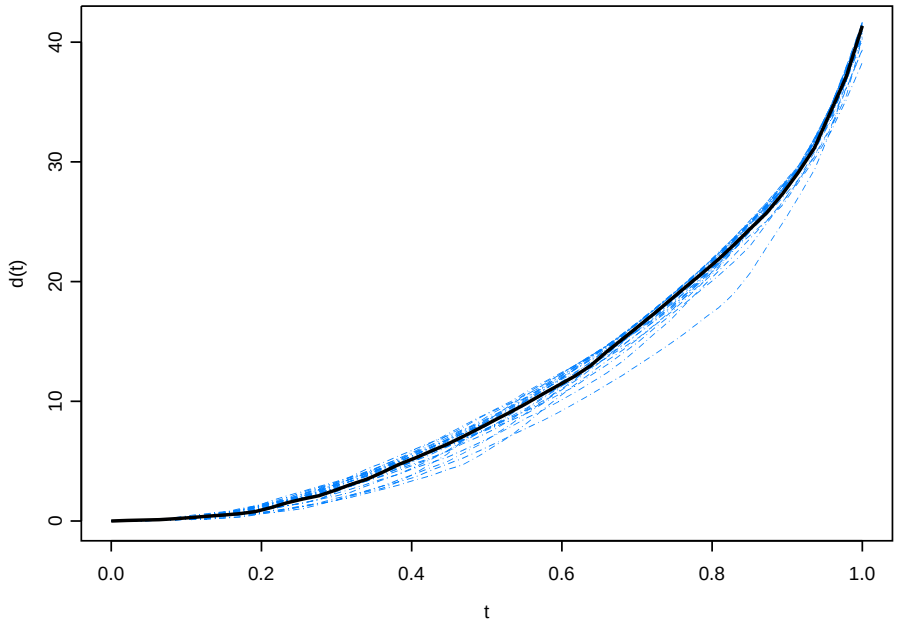


(a) Sample reflected across the origin.

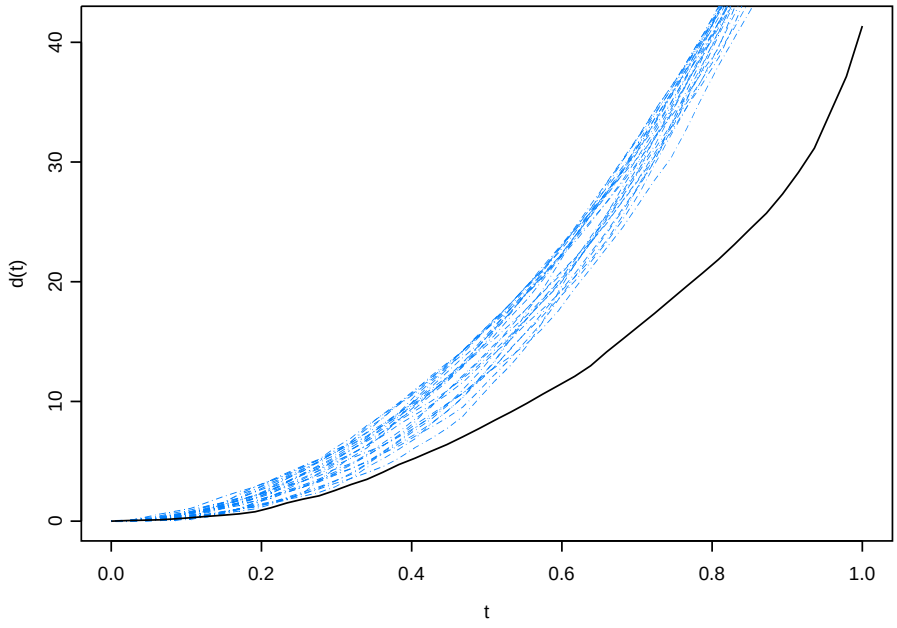


(b) Sample reflected across (1.5,1.5)

Figure 2.1: The reflection principle illustrated using a bivariate N(0,I) sample.



(a) Reflection done across the origin



(b) Reflection done across (1,1)

Figure 2.2: Comparison of $d(t)$ and 25 randomly generated $d^*(t)'s$ for a N(0,I) sample.

Mean	d(.25)	d(.5)	d(.75)	d(1)	Hotelling T^2
0,0	0.047	0.048	0.053	0.052	0.059
0.25,0.25	0.066	0.124	0.211	0.383	0.374
0.5,0.5	0.120	0.410	0.680	0.940	0.910
0.75,0.75	0.418	0.754	0.956	1.000	0.999
1,1	0.668	0.950	1.000	1.000	1.000

(a) Bivariate Normal

Mean	d(.25)	d(.5)	d(.75)	d(1)	Hotelling T^2
0,0	0.044	0.051	0.058	0.057	0.055
0.25,0.25	0.103	0.164	0.203	0.194	0.196
0.5,0.5	0.286	0.514	0.656	0.686	0.677
0.75,0.75	0.492	0.822	0.932	0.956	0.944
1,1	0.700	0.970	0.996	0.998	0.996

(b) Bivariate Double Exponential

Mean	d(.25)	d(.5)	d(.75)	d(1)	Hotelling T^2
0,0	0.052	0.047	0.056	0.048	0.013
0.25,0.25	0.098	0.126	0.132	0.077	0.035
0.5,0.5	0.100	0.330	0.330	0.170	0.087
0.75,0.75	0.460	0.700	0.670	0.280	0.157
1,1	0.640	0.870	0.850	0.380	0.288

(c) Bivariate Cauchy

Table 2.1: Power comparison for One Sample Test using samples of size 30 and $\alpha = 0.05$

moment, Hotelling T^2, expectedly, does poorly. It is in these cases, where the test at $d(0.5)$ performs better, since the test, by design, has taken care of the far outlying points, and as a result, is far more robust. In fact, even for the Double Exponential case, the test at $d(0.5)$ performs reasonably well. All this indicates to the fact that whenever we have a significant number of outliers, the test at $d(0.5)$ can be a very effective robust non-parametric test alternative to the Hotelling T^2.

We show below that the fact that the test at $d(1)$ performs similar to the Hotelling T^2 for Normal and Double Exponential populations is not a coincidence.

Let $\mathbf{X}$ and $\mathbf{X}^*$ denote the matrices with the X_i's and $X_i{}^*$'s as their columns. Define $S = \frac{1}{n}SS(\mathbf{X})$ and $S^* = \frac{1}{n}SS(\mathbf{X}^*)$.

Lemma 2.1. *Under the above reflection scheme, the following identity holds:*

$$|S|\left[1 + (\bar{X} - \mu)'S^{-1}(\bar{X} - \mu)\right] = |S^*|\left[1 + (\bar{X}^* - \mu)'S^{*-1}(\bar{X}^* - \mu)\right]. \tag{2.2}$$

In other words, for any given μ, the statistic $|S|\left[1+(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu)\right]$ is invariant under the reflection scheme, when reflected across μ.

Proof. By 2.1,

$$X_i{}^* - \mu = (2\tau_i - 1)\,(X_i - \mu)\;; i = 1, \ldots, n.$$

Since $(2\tau_i - 1)^2 \equiv 1, \forall i$, we have,

$$(\mathbf{X} - \mu\mathbf{1}')(\mathbf{X} - \mu\mathbf{1}')' = (\mathbf{X}^* - \mu\mathbf{1}')(\mathbf{X}^* - \mu\mathbf{1}')' \tag{2.3}$$

Now,

$$\begin{aligned}(\mathbf{X} - \mu\mathbf{1}')(\mathbf{X} - \mu\mathbf{1}')' =& (\mathbf{X} - \bar{X}\mathbf{1}')(\mathbf{X} - \bar{X}\mathbf{1}')' + n(\bar{X} - \mu)(\bar{X} - \mu)' \\ =& SS(\mathbf{X}) + n(\bar{X} - \mu)(\bar{X} - \mu)'\end{aligned}$$

Thus,

$$\begin{aligned}\left|\frac{1}{n}(\mathbf{X} - \mu\mathbf{1}')(\mathbf{X} - \mu\mathbf{1}')'\right| &= \left|\frac{1}{n}SS(\mathbf{X}) + (\bar{X} - \mu)(\bar{X} - \mu)'\right| \\ &= \left|S + (\bar{X} - \mu)(\bar{X} - \mu)'\right| \\ &= |S|\left[1 + (\bar{X} - \mu)'S^{-1}(\bar{X} - \mu)\right]\end{aligned}$$

Similarly,

$$\left|\frac{1}{n}(\mathbf{X}^* - \mu\mathbf{1}')(\mathbf{X}^* - \mu\mathbf{1}')'\right| = |S^*|\left[1 + (\bar{X}^* - \mu)'S^{-1}(\bar{X}^* - \mu)\right]$$

Therefore, by 2.3, we have the required identity. □

Theorem 2.2. *For testing $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ at significance level α, under the assumptions:*

(i) $E||X_1||^2 < \infty$

(ii) $\mu - \mu_0 = O_p(n^{-\frac{1}{2}})$

the permutation test using $d(1)$ has the asymptotic rejection region:

$$n(\bar{X} - \mu_0)S^{-1}(\bar{X} - \mu_0)' > {\chi^2}_{p,(1-\alpha)} \tag{2.4}$$

where S is the sample variance-covariance matrix of $\{X_i\}_{\{i=1,\ldots,n\}}$

Proof. Without loss of generality, we may assume $\mu_0 = 0$ and the variance-covariance matrix Σ of F to be identity, i.e., $\Sigma = I$, since replacing each X_i by $\Sigma^{-\frac{1}{2}}(X_i - \mu_0)$ makes 2.4 free of μ_0 and Σ. Thus by Lemma 2.1, we have:

$$|S| \left[1 + \bar{X}' S^{-1} \bar{X}\right] = |S^*| \left[1 + \bar{X}^{*\prime} S^{*-1} \bar{X}^*\right], \tag{2.5}$$

So, the rejection region of the test given by : $|S| <$ lower $100\alpha\%$ of $|S^*|$, which, by 2.5, is true iff $n\bar{X}' S^{-1} \bar{X} >$ upper $100\alpha\%$ of $n\bar{X}^{*\prime} S^{*-1} \bar{X}^*$.

Now, using standard Central Limit Theorem for triangular arrays (see e.g. Feller(Vol 2)), we have the convergence

$$\sqrt{n} S^{*-\frac{1}{2}} \bar{X}^* \to N(0, I)$$

under the distribution of the τ_i's, a.s. $\mathbf{X}$. The claim of the theorem follows. □

Corollary 2.3. *The test using $d(1)$ is Pitman equivalent to the Hotelling T^2 test.*

Proof. The result follows immediately from the above theorem. □

2.1.1 Application to Paired data

Paired data appear frequently in experimental situations where two observations are made on a given individual or experimental object, resulting in a natural pairing of values. A natural question that arises in such situations is whether the two components in the paired data are exchangeable or not, possibly upto a real-valued transformation. Thus, if $(X_{i1}, X_{i2})_{i=1\ldots,n}$ denote a paired sample of size n, and $f : \Re \to \Re$ be a known real valued function, we define:

$$X_i = X_{i1} - f(X_{i2}), \quad i = 1, \ldots, n.$$

The hypothesis of interest is H_0 : the distribution of $\{X_i\}_{\{i=1,\ldots,n\}}$ is symmetric about 0, vs. the alternative hypothesis H_1 of asymmetry.

It can be seen that the test developed in the previous section can be applied directly to $\{X_i\}_{\{i=1,\ldots,n\}}$ to test H_0, as long as the the assumptions in Theorem 2.2 hold. Thus the test using $d(1)$ will have the same nice asymptotic optimality property.

Note here that in most practical applications, the function f would be assumed to be identity, i.e. $f(x) = x, x \in \Re$. The special case where $f \equiv$ constant, yields tests for the components having symmetric distributions.

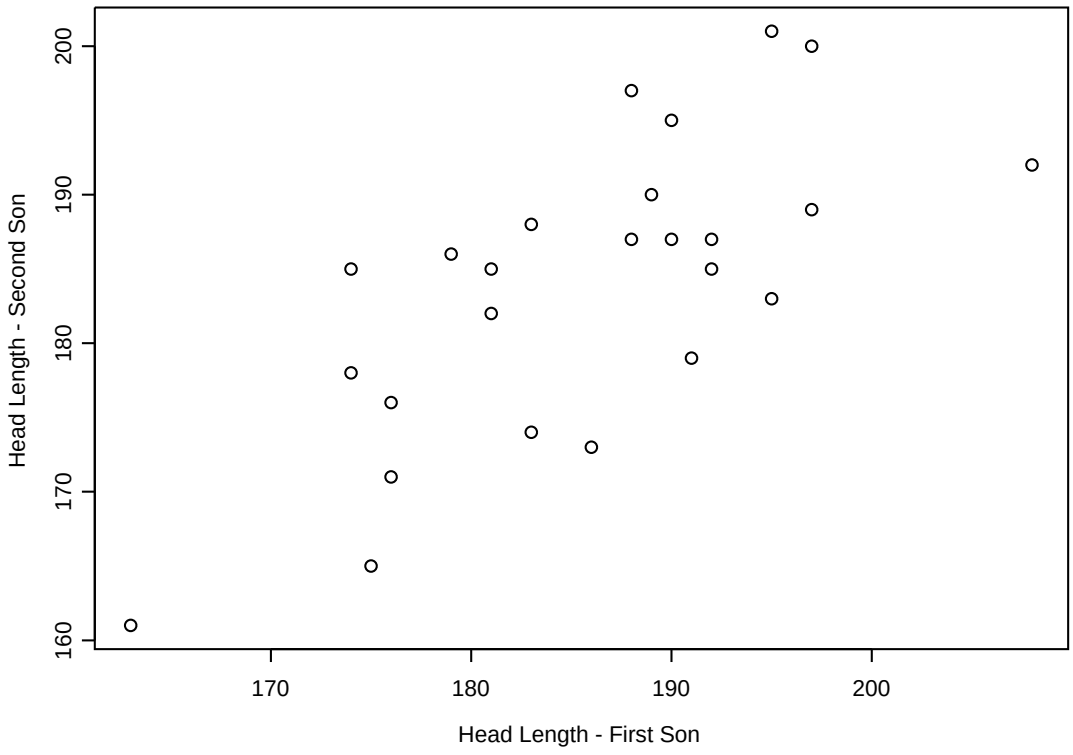
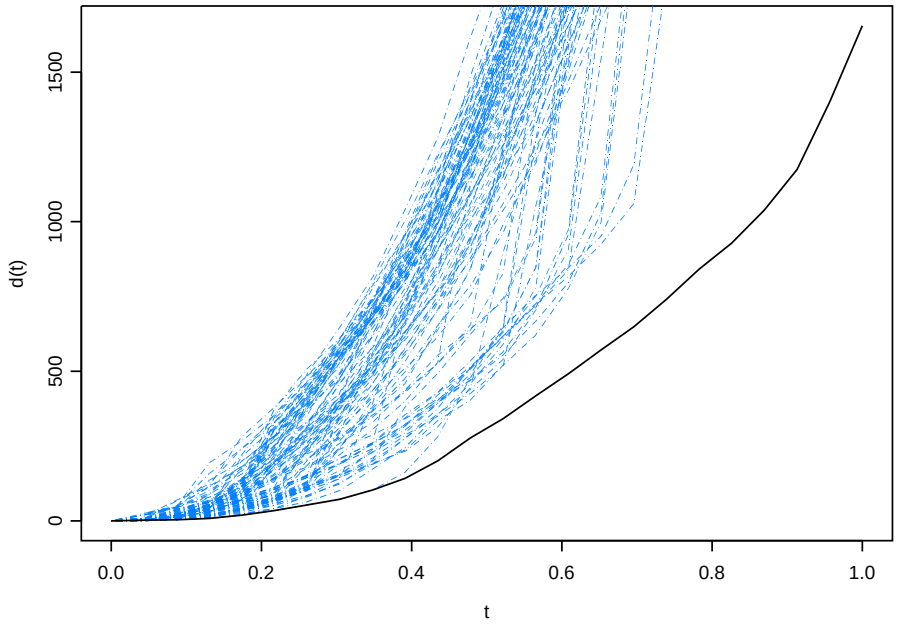


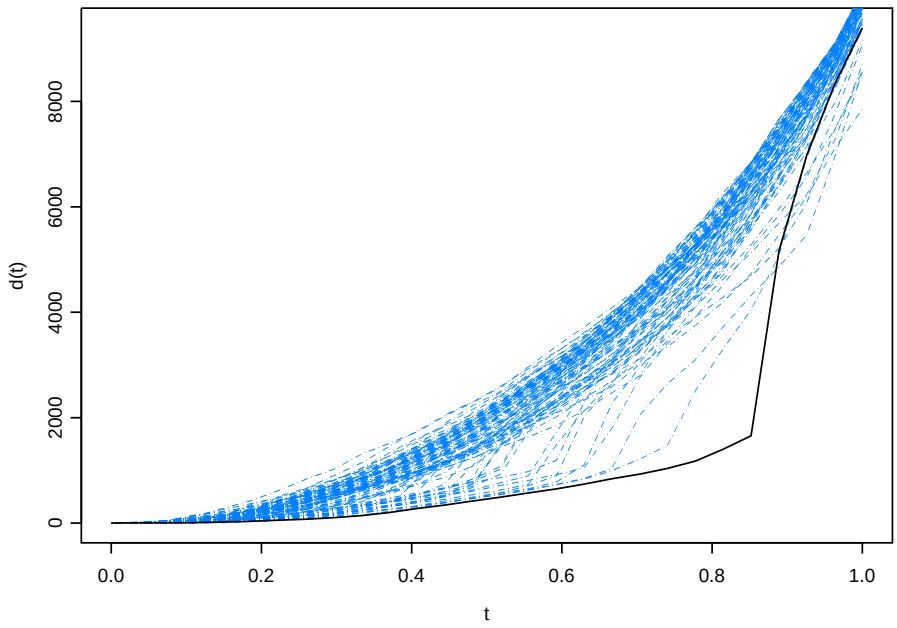
Figure 2.3: Scatterplot of Head Length Data

2.2 An Example

As discussed earlier, we have a version of the test using $d(t)$, for every $0 < t \leq 1$. The results in the previous section deal mostly with the test using $d(1)$. However, Monte Carlo simulation of powers shown in Table 2.1 exhibit clearly that the test using $d(.5)$ should be the better choice in most practical situations, where one would expect a good number of outliers. In this section, we will examine this property, using a dataset from



(a) *dsc* without outlier



(b) *dsc* with outlier

Figure 2.4: Tests using *dsc* on Head Length Data

[Frets, G. P. 1921] which shows the measurement of head length of the first and second adult sons in 25 families. For the entire dataset as well as some analysis, please refer [Mardia, Kent, Bibby 1979]. Figure 2.3 shows the scatter plot.

Let μ denote the mean of the underlying bivariate distribution. Suppose we choose to test $H_0 : \mu = (150, 150)$ vs $H_1 : \mu \neq (150, 150)$, to illustrate the robustness of the tests at the middle section of the *dsc*'s.

The mean of the sample is $(185.72, 183.84)$. [Mardia, Kent, Bibby 1979] show that the hypothesis $\mu = (182, 182)$ cannot be rejected by either the LRT or the Hotelling T^2 test. Thus one might expect the true mean to be somewhere close to the point $(182, 182)$, and the Null hypothesis here to be false. Indeed, the Hotelling T^2 test yields p-value 0 for the above test, thus rejecting H_0 at all levels. The same is reflected in Figure 2.4(a) where $d(.)$, the *dsc* of the original data lies far below the $dsc's$ of the reflected datasets, $d^*()$'s.

Now, we add four outliers to the data, each having value (50,50). The p-value for the Hotelling T^2 test now becomes 0.134. So, the test cannot reject H_0 at even level 10%. The dsc plots in this case, shown in Figure 2.4(b) , clearly exhibit the robustness property of the test at $d(0.5)$. The end point $d(1)$ lies within the band of the $d^*()$'s, and hence shows characteristics similar to the Hotelling T^2 test. The midpoint $d(0.5)$ however lies below the entire band of $d^*()$'s. So, the test at $d(0.5)$ manages to reject H_0 even in the presence of the outliers.

Chapter 3

The Multisample Multivariate Location Problem

A basic question that arises in all statistical analyses that involve two or more datasets, is whether the samples come from distributions having the same centrality parameter. The situation appears frequently in experimental setups where subjects are administered different levels of treatment, and observations obtained for the different treatment levels are compared to see whether they vary significantly from one another or not. These questions are often handled via ANOVA or MANOVA, to analyze univariate and multivariate samples respectively. See [Montgomery, D.C. 1976], [Rao, C.R. 1973], [Anderson, T.W. 1958], [Wilks, S.S. 1962] or [Mardia, Kent, Bibby 1979] for detailed discussions of these problems. For non-parametric analysis of similar problems, see [Puri, Sen 1971]. In this chapter, we will introduce a multi-sample variant of the Determinant Scale Curve and then describe a permutation scheme that will help us devise tests for the multi-sample multivariate location problems using the *dsc*'s. Similar to the results in the previous chapter, the test at the rightmost end of the multi-sample *dsc* will be shown to have optimality properties, and tests in the central region will be shown to be robust. The results will be reinforced through Monte-Carlo simulation of the powers of these tests, under different distributions.

3.1 The Multisample Determinant Scale curve

Let $\{X^{(j)}{}_i\}_{\{i=1,\ldots,n_j\}}$ denote the j-th sample of size n_j on $\Re^p$, from a distribution $\mathbf{F_j}$, which is assumed to be symmetric around the parameter μ_j, $j = 1, \ldots, k$. Our primary interest in this chapter would be to test the equality of two or more of the μ_j's.

Define the total sample size as $N = \sum_{j=1}^{k} n_j$. Let $\mathbf{X^{(j)}}$ denote the $p \times n_j$ matrix with $X^{(j)}{}_i$'s as its columns.

For each $j = 1, \ldots, k$, we define:

$$W^{(j)}(1) = SS(\mathbf{X^{(j)}}),$$
$$W^{(j)}\left(\frac{n_j - 1}{n_j}\right) = \min_{1 \le i \le n_j} SS(\mathbf{X^{(j)}}_{(-\mathbf{i})}),$$

where the minimum is taken over the determinants of the matrices.

Suppose, for a given sample, the above minimum is attained at $\mathbf{X^{(j)}}_{(-\mathbf{i})}$. Then define,

$$W^{(j)}\left(\frac{n_j - 2}{n_j}\right) = \min_{1 \le l \le n_j, l \ne i} SS([\mathbf{X^{(j)}}_{(-\mathbf{i})}]_{(-\mathbf{l})}).$$

Now, for $\frac{k}{n_j} \le t < \frac{k+1}{n_j}$, $k = 1, \ldots, n_j - 1$, define:

$$W^{(j)}(t) = W^{(j)}\left(\frac{k}{n_j}\right)$$

Now, we go on to define the Multi-sample Determinant Scale Curve.

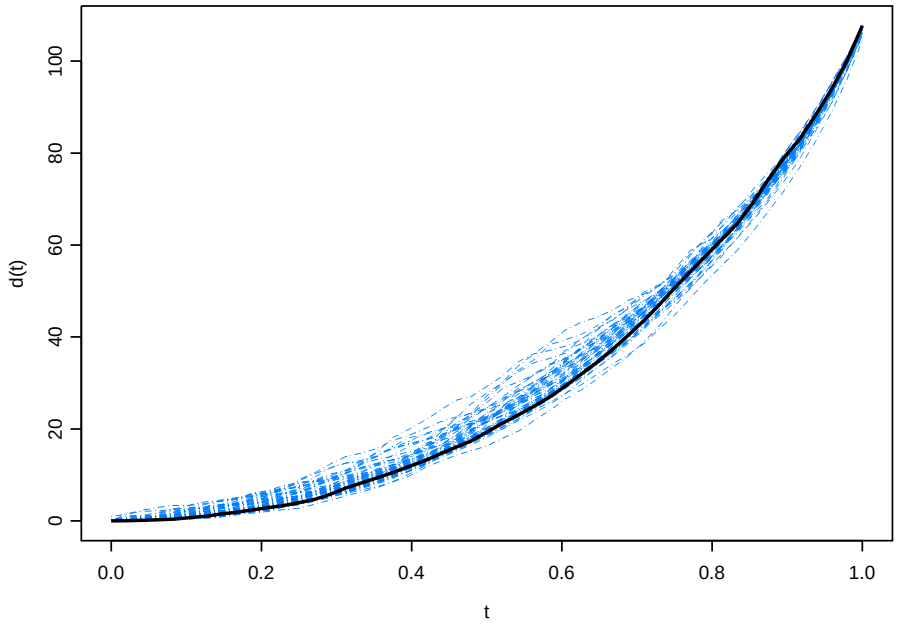
Definition 3.1. *For $0 \le t \le 1$, the multi-sample dsc $d(.)$ for k samples is defined as:*

$$d(t) = \left| \sum_{j=1}^{k} W^{(j)}(t) \right|$$

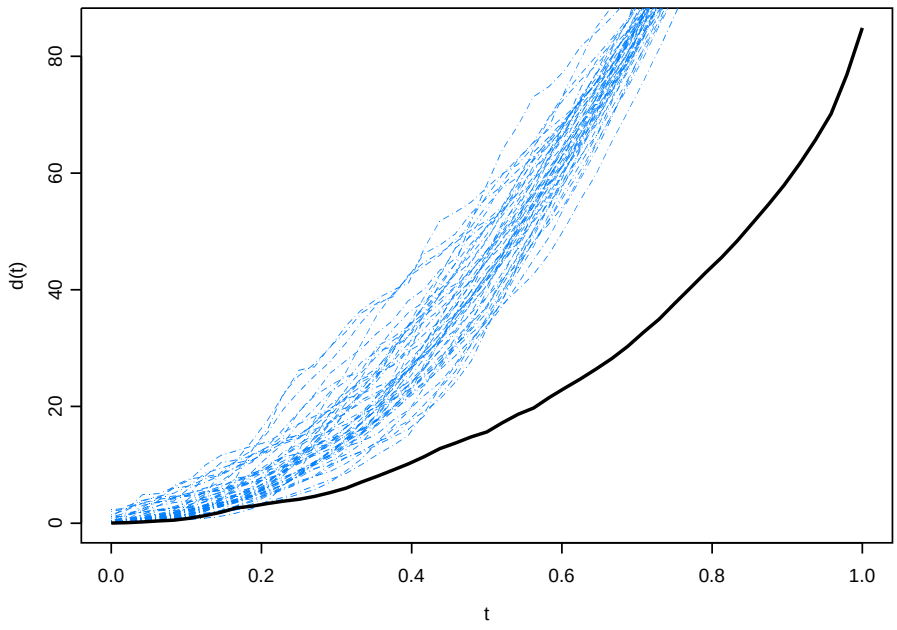
3.1.1 A Permutation Scheme and tests using the Multi-Sample *dsc*

The tests in this chapter would involve a permutation scheme. To create a permuted set of data, we first combine all the k samples into a single dataset of size N. Then, we randomly split the dataset into k partitions, the size of the j-th partition being n_j, $1 \le j \le k$. Now, one can create the multisample *dsc* using these permuted samples. We will denote the permuted samples as $\{X^{(j)*}{}_i\}_{\{i=1,\ldots,n_j\}}$ and multisample *dsc* computed on them as $d^*(.)$. If the means of the k populations are close to each other, then this process of randomization will not change the overall spread of the samples appreciably. Hence, the curve $d^*(.)$ will be close to the original curve $d(.)$. On the other hand, if at least two of the μ_j's are far apart, the permutation will make the samples significantly more diffused. As a result, $d^*(.)$ will tend to lie above $d(.)$.

One can make use of this phenomenon to devise a test for the hypothesis H_0 : All μ_j's are equal vs. $H_1 : \mu_j \ne \mu_l$ for some $1 \le j \ne l \le k$. Suppose we generate a large



(a) Two-sample *dsc* under H_0



(b) Two-sample *dsc* under H_1

Figure 3.1: Tests using multisample *dsc* on Bivariate Normal samples

number of replicates of $d^*(.)$. Now, to test at level α, we reject the null hypothesis if $d(.)$ lies in or below the bottom $100\alpha\%$ of the $d^*(.)$'s. As in the previous chapter, here also we have distinct test for every $0 < t \leq 1$, and we would restrict ourselves to the choices $t = 0.5, 0.75, 0.9, 1$. The tests are exact for every choice of t.

An illustration of this testing procedure is shown in Figure 3.1. Figure shows the *dsc*'s for two N(0,1) samples. The $d(.)$ lies within the band of $d^*(.)$, thus failing to reject the Null hypothesis at all t, for all reasonable α. When the second sample is centred around (3,3), the $d(.)$ lies well outside the band of $d^*(.)$'s. The H_0 is hence rejected. Before going into the theoretical properties of these tests, we will try to apply the method on two real life examples.

3.2 Examples

Reeve (see [Reeve, E.C.R. 1941]) provides measurements of the skulls of 13 ant-eaters, belonging to the sub-species *chapadensis*, deposited in the British Museum, from 3 different locations in South America. Table 3.2 shows the natural logarithms of the measurements. The variables are respectively- x1: the basal length excluding premaxilla, x2: the occipito-nasal length, and x3: the maximum nasal length.

Minas Graes, Brazil			Matto Grosso, Brazil			Santa Cruz, Bolivia		
x1	x2	x3	x1	x2	x3	x1	x2	x3
2.068	2.070	0.048	1.580	2.045	1.580	2.093	2.098	1.653
2.068	2.074	1.602	2.076	2.088	1.602	2.100	2.106	1.623
2.090	2.090	1.613	2.090	2.093	1.643	2.104	2.101	1.653
2.097	2.093	1.613	2.111	2.114	1.643	-	-	-
2.117	2.125	1.663	-	-	-	-	-	-
2.140	2.146	1.681	-	-	-	-	-	-

Table 3.1: Logarithms of measurements of the skulls of 13 ant-eaters, provided by Reeve(1941)

The null hypothesis to be tested is that there is no significant difference between the locations, versus the alternative of having a significant difference between at least two locations.The Wilks Lambda statistic comes to 0.772, which, when compared to the $F_{6,16}$ distribution, provides a p-value of nearly 40%. Hence, the Null hypothesis cannot be rejected at any reasonable level. The *dsc* test, shown in Figure 3.2, comes to

the same conclusion for all t.

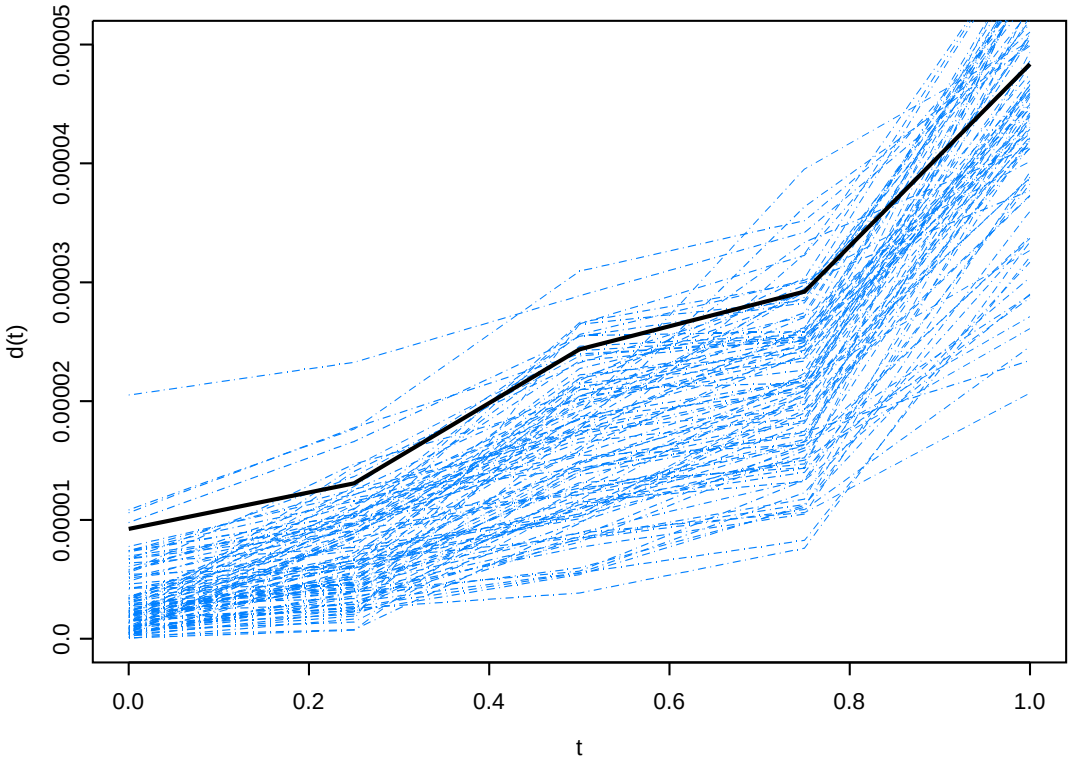


Figure 3.2: Tests using multisample *dsc* on Reeve's Ant Eater Data

As our second example, we look at R.A.Fisher's Iris Data (see [Fisher,R.A. 1936]), showing the lengths and widths of sepals and petals of three different varieties of Iris. It is well known that the means of the three varieties are significantly different. Our proposed array of tests, shown in Figure 3.3, provides a confirmation at all t.

3.3 Power Simulations

Table 3.3 shows the Monte Carlo simulations of powers for some standard bivariate distributions using samples at a time. For Normal and Double Exponential samples,

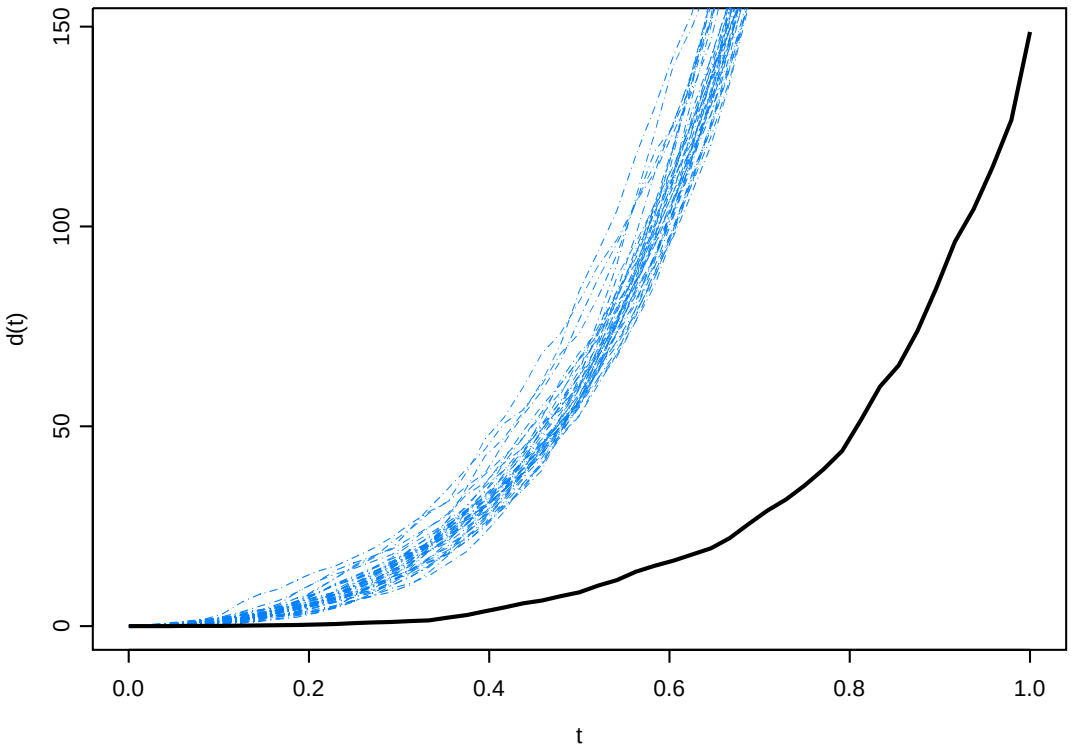


Figure 3.3: Tests using multisample *dsc* on Fisher's Iris Data

the test at $d(1)$ performs almost as well as the Wilks Lambda test, which is the most powerful test under (approx.) Normality assumptions. As in the case for the one-sample location tests, the tests in the middle region of the *dsc* show strong robustness. The Wilks Lambda test performs poorly for the Cauchy distribution, where one would expect to have a significant number of extreme observations. The test at $d(0.5)$ provides much better power in such cases. In fact, for moderate separation between the means of the populations, the power of the test at $d(0.5)$ for the Normal and Cauchy distributions is comparable, reinforcing the fact that the tests are robust in the middle zone of the *dsc*.

Diff. of Means	d(.5)	d(.75)	d(.9)	d(1)	Wilks Λ
(0,0)	0.044	0.050	0.044	0.052	0.049
(0.5,0.5)	0.058	0.100	0.172	0.214	0.247
(0.75,0.75)	0.104	0.204	0.354	0.498	0.490
(1,1)	0.180	0.330	0.580	0.782	0.740
(1.5,1.5)	0.282	0.610	0.894	0.986	0.983
(2,2)	0.450	0.806	0.990	1.000	1.000

(a) Bivariate Normal

Diff. of Means	d(.5)	d(.75)	d(.9)	d(1)	Wilks Λ
(0,0)	0.055	0.048	0.056	0.058	0.042
(0.5,0.5)	0.112	0.122	0.152	0.184	0.150
(0.75,0.75)	0.148	0.212	0.270	0.320	0.311
(1,1)	0.162	0.294	0.390	0.460	0.496
(1.5,1.5)	0.342	0.588	0.754	0.848	0.832
(2,2)	0.446	0.748	0.902	0.952	0.965

(b) Bivariate Double Exponential

Diff. of Means	d(.5)	d(.75)	d(.9)	d(1)	Wilks Λ
(0,0)	0.058	0.055	0.050	0.035	0.018
(0.5,0.5)	0.090	0.104	0.088	0.070	0.023
(0.75,0.75)	0.094	0.124	0.118	0.110	0.050
(1,1)	0.142	0.186	0.158	0.144	0.095
(1.5,1.5)	0.260	0.326	0.246	0.236	0.188
(2,2)	0.404	0.480	0.418	0.384	0.284

(c) Bivariate Cauchy

Table 3.2: Power comparison for Two Sample Test using samples of size 10 each and $\alpha = 0.05$

3.4 Optimality of the test at $t = 1$

In this section, we will go on to show some optimality properties of the permutation test at $t = 1$, which will explain why its power is comparable to that of Wilks Lambda test, for approximately Normal populations.

To that end, we must introduce some notations and prove the following lemmas.

Lemma 3.1. *Let* $\mathbf{X} = \{X_i\}_{\{i=1,\ldots,N\}}$ *be an univariate finite population made of i.i.d. observations, of size* N *with* $E(X_i) = \mu < \infty$ *, and let* $\mathbf{Y} = \{Y_i\}_{\{i=1,\ldots,n\}}$ *be a sample of size* n *drawn without replacement, such that* $\frac{n}{N} \to \lambda \in (0,1)$ *as* $N \to \infty$*. Let* $\bar{X}$ *and* $\bar{Y}$ *denote the population mean and the sample mean respectively. Then, for any* $\epsilon > 0$*,*

Mean	d(.5)	d(.75)	d(.9)	d(1)	Wilks Λ
(0,0),(0,0),(0,0)	0.044	0.043	0.055	0.053	0.048
(0,0),(0.25,0),(0,0.25)	0.068	0.095	0.164	0.227	0.250
(0,0),(0.5,0),(0,0.5)	0.154	0.345	0.558	0.792	0.792
(0,0),(1,0),(0,1)	0.545	0.927	0.992	1.000	1.000
(0,0),(0.5,0.5),(1,1)	0.338	0.706	0.949	0.996	0.998
(0,0),(1,1),(2,2)	0.881	1.000	1.000	1.000	1.000

(a) Bivariate Normal

Mean	d(.5)	d(.75)	d(.9)	d(1)	Wilks Λ
(0,0),(0,0),(0,0)	0.045	0.050	0.048	0.045	0.034
(0,0),(0.25,0),(0,0.25)	0.092	0.120	0.127	0.013	0.138
(0,0),(0.5,0),(0,0.5)	0.223	0.311	0.387	0.461	0.472
(0,0),(1,0),(0,1)	0.695	0.869	0.926	0.979	0.990
(0,0),(0.5,0.5),(1,1)	0.519	0.697	0.774	0.880	0.858
(0,0),(1,1),(2,2)	0.973	0.997	1.000	1.000	1.000

(b) Bivariate Double Exponential

Mean	d(.5)	d(.75)	d(.9)	d(1)	Wilks Λ
(0,0),(0,0),(0,0)	0.051	0.055	0.057	0.049	0.032
(0,0),(0.25,0),(0,0.25)	0.083	0.065	0.048	0.051	0.014
(0,0),(0.5,0),(0,0.5)	0.152	0.118	0.061	0.069	0.012
(0,0),(1,0),(0,1)	0.502	0.364	0.148	0.076	0.056
(0,0),(0.5,0.5),(1,1)	0.332	0.246	0.111	0.085	0.032
(0,0),(1,1),(2,2)	0.884	0.680	0.263	0.177	0.120

(c) Bivariate Cauchy

Table 3.3: Power comparison for Three Sample Test using samples of sizes 30,50 and 30 respectively and $\alpha = 0.05$

we have:

$$P\{|\bar{Y} - \bar{X}| > \epsilon\} \to 0 \; as \; N \to \infty \; a.s.[\mathbf{X}]$$

Note here that the probability is with respect to the simple random sampling without replacement, and the result is true for almost all path of $[\mathbf{X}]$.

Proof. By Tchebychev's Inequality, we have the following bound for the above probability:

$$\begin{aligned} P\{|\bar{Y} - \bar{X}| > \epsilon\} &\leq \frac{1}{\epsilon^2}\frac{N-n}{N-1}\frac{1}{nN}\left[\sum_{i=1}^{N}(X_i - \bar{X})^2\right] \\ &\sim \frac{1}{\epsilon^2}(1 - \frac{1}{\lambda})\frac{1}{\lambda}\left[\frac{\sum_{i=1}^{N}{X_i}^2}{N^2} - \frac{\bar{X}^2}{N}\right] \end{aligned}$$

Now, since $E(X_i) = \mu < \infty$, $\frac{\bar{X}^2}{N} \to 0$, and by Marcinkiewicz-Zigmund SLLN, $\frac{\sum_{i=1}^{N} {X_i}^2}{N^2} \to 0$ as $N \to \infty$. Hence the result follows. □

Define $\bar{X^2}$ and $\bar{Y^2}$ as the means of ${X_i}^2$'s and ${Y_i}^2$'s respectively. Analogously, for any real k, define $\bar{X^{(k)}}$ and $\bar{Y^{(k)}}$ as the means of ${X_i}^2 I(X_i > k)$'s and ${Y_i}^2 I(Y_i > k)$'s respectively. Then, as an immediate consequence of the previous lemma, we have the following results:

Lemma 3.2. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be as in Lemma* 3.1 *with* $E({X_1}^2) < \infty$. *Then, for any* $\epsilon > 0$, *we have:*

$$P\{|\bar{Y^2} - \bar{X^2}| > \epsilon\} \to 0 \text{ as } N \to \infty \text{ a.s.}[\mathbf{X}]$$

Lemma 3.3. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be as in Lemma* 3.2. *Then, for any fixed* k *and* $\epsilon > 0$, *we have:*

$$P\{|\bar{Y^{(k)}} - \bar{X^{(k)}}| > \epsilon\} \to 0 \text{ as } N \to \infty \text{ a.s.}[\mathbf{X}]$$

Under the setup of Lemma 3.3, define the set $G(\mathbf{Y}, k, \epsilon)$ as:

$$G(\mathbf{Y}, k, \epsilon) = \{\mathbf{Y} \subset \mathbf{X} : |\bar{Y} - \bar{X}| < \epsilon,\ |\bar{Y^2} - \bar{X^2}| < \epsilon,\ \bar{Y^{(k)}} - \bar{X^{(k)}}| < \epsilon\}$$

It follows from Lemmas 3.1, 3.2 and 3.3 that $P[G(\mathbf{Y}, k, \epsilon)] \to 1$ *as* $n, N \to \infty$ *a.s.*$[X]$, for all real k.

Theorem 3.4. *Let* $\mathbf{X} = \{X_i\}_{\{i=1,\ldots,N\}}$ *be an univariate finite population of size* N *with* $Var(X_i) = \sigma^2 < \infty$. *Let* $\{S_1, S_2, S_3\}$ *be a random disjoint partition of* $\mathbf{X}$, *such that the size of* S_i *is* n_i, $i = 1, 2, 3$; $\sum_{i=1}^{3} n_i = N$. *Let the mean of the partition* S_i *be* $\bar{X}_i$, $i = 1, 2, 3$. *Also assume* $\frac{n_i}{N} = \lambda_i \to \Delta_i \in (0, 1)$, $i = 1, 2, 3$. *Then we have the Central Limit Theorem:*

$$\sqrt{N}(\bar{X}_1 - \bar{X}, \bar{X}_2 - \bar{X}, \bar{X}_3 - \bar{X})' \Rightarrow N(\mathbf{0}, \Sigma),\ a.s[\mathbf{X}],$$

where $\Sigma = ((\sigma_{i,j}))$ *is given by* $\sigma_{i,i} = \sigma^2(\frac{1}{\Delta_i} - 1)$ *and* $\sigma_{i,j} = -\sigma^2$, $i, j = 1, 2, 3, i \neq j$.

Proof. Note that without loss of generality, $\sigma^2 = 1$ and $E(X_i) = 0$.

Suffices to show, for any real vector $\mathbf{a} = \{a_1, a_2, a_3\}$,

$$T_N = \sqrt{N} \sum_{i=1}^{3} a_i(\bar{X}_i - \bar{X}) \Rightarrow N(\mathbf{0}, \mathbf{a}'\Sigma\mathbf{a}),\ a.s.[\mathbf{X}],$$

Note that $\sum_{i=1}^{3} \lambda_i \bar{X}_i = \bar{X}$. Hence, $\bar{X}_2 = -\frac{\lambda_1}{\lambda_2}\bar{X}_1 - \frac{\lambda_3}{\lambda_2}\bar{X}_3 - \frac{\bar{X}}{\lambda_2}$.

Hence, we can eliminate $\bar{X}_2$ from the above expression. Let

$$\begin{aligned} T_N{}' &= \sqrt{N}\sum_{i=1}^{3} a_i \bar{X}_i \\ &= \sqrt{N}(a_1 - \frac{a_2\lambda_1}{\lambda_2})\bar{X}_1 + \sqrt{N}(a_3 - \frac{a_2\lambda_3}{\lambda_2})\bar{X}_3 + \sqrt{N}\frac{a_2\bar{X}}{\lambda_2} \\ &= \sqrt{N}\frac{a_2\bar{X}}{\lambda_2} + \sqrt{N}(c_1\bar{X}_1 + c_2\bar{X}_3), \end{aligned}$$

where $c_1 = a_1 - \frac{a_2\lambda_1}{\lambda_2}$ and $c_2 = a_3 - \frac{a_2\lambda_3}{\lambda_2}$.

Now, we can restrict ourselves to the cases when S_3 belongs to the "good" set $G(S_3, k, \epsilon)$, as follows:

For $i = \sqrt{(-1)}$ and real t,

$$\begin{aligned} E(e^{itT_N{}'}) &= e^{\sqrt{N}\frac{a_2\bar{X}}{\lambda_2}} E e^{it\sqrt{N}(c_1\bar{X}_1 + c_2\bar{X}_3)} \\ &= e^{\sqrt{N}\frac{a_2\bar{X}}{\lambda_2}} E e^{it\sqrt{N}(c_1\bar{X}_1 + c_2\bar{X}_3)} I(S_3 \in G(S_3, k, \epsilon)) + \xi, \end{aligned}$$

where $\xi = E(e^{itT_N{}'} I(S_3 \notin G(S_3, k, \epsilon))) \leq 1 - P(G(S_3, k, \epsilon)) \to 0$ as $N \to \infty$, for any fixed k, since $|e^{iu}| \leq 1$ for all real u.

Now,

$$\begin{aligned} E(e^{itT_N{}'}) &= e^{it\sqrt{N}\frac{a_2\bar{X}}{\lambda_2}} E\left[e^{it\sqrt{N}(c_1\bar{X}_1 + c_2\bar{X}_3)} I(S_3 \in G(S_3, k, \epsilon))\right] + \xi \\ &= e^{it\sqrt{N}\frac{a_2\bar{X}}{\lambda_2}} E\left[e^{it\sqrt{N}(c_2\bar{X}_3)} I(S_3 \in G(S_3, k, \epsilon)) E[e^{it\sqrt{N}c_1\bar{X}_1} I(S_3 \in G(S_3, k, \epsilon))|S_3]\right] + \xi \end{aligned}$$

Given $S_3 \in G(S_3, k, \epsilon)$, we have:

$$\sum_{S_1 \cup S_2} {X_j}^2 = (1 - \lambda_3)\sum_{j=1}^{N} {X_j}^2 + O(\epsilon)N$$

$$\sum_{S_1 \cup S_2} {X_j}^2 I(X_j > k) = (1 - \lambda_3)\sum_{j=1}^{N} {X_j}^2 I(X_j > k) + O(\epsilon)N.$$

Therefore, given $S_3 \in G(S_3, k, \epsilon)$,

$$\sup_{S_3} \frac{\sum_{S_1 \cup S_2} {X_j}^2 I(X_j > k)}{\sum_{S_1 \cup S_2} {X_j}^2} = \frac{\frac{1-\lambda_3}{N}\sum_{j=1}^{N} {X_j}^2 I(X_j > k) + O(\epsilon)}{\frac{1-\lambda_3}{N}\sum_{j=1}^{N} {X_j}^2 + O(\epsilon)}, \tag{3.1}$$

which can be made arbitrarily small by choosing a large enough k a.s.$[\mathbf{X}]$, since $E({X_1}^2) < \infty$. We can therefore use Erdos-Renyi's CLT (see [Erdös, P. and Rényi, A. (1959)]) in

conjunction with the uniformity of 3.1 over all $S_3 \in G(S_3, k, \epsilon)$ to claim:

$$E\left[e^{it\sqrt{N}c_1\bar{X}_1} I(S_3 \in G(S_3,k,\epsilon))|S_3\right] = e^{\sqrt{N}itc_1 E(\bar{X}_1 I(S_3 \in G(S_3,k,\epsilon))|S_3)} \times e^{-\frac{t^2 {c_1}^2}{2} Var(\sqrt{N}\bar{X}_1 I(S_3 \in G(S_3,k,\epsilon))|S3)+o(1)}$$

Now,

$$\begin{aligned} E(\bar{X}_1|S_3 \in G(S_3,k,\epsilon)) &= \frac{N\bar{X} - n_3\bar{X}_3}{N-n_3} \\ &= \frac{1}{1-\lambda_3}\bar{X} - \frac{\lambda_3}{1-\lambda_3}\bar{X}_3 + o(\epsilon) \end{aligned}$$

as $N \to \infty$.

Also,

$$\begin{aligned} &Var(\sqrt{N}\bar{X}_1|\bar{X}_3, S_3 \in G(S_3,k,\epsilon)) \\ &= \frac{(N-n_3)-n_1}{(N-n_3)-1} \frac{1}{\left(\frac{n_1}{N}\right)} \left[\frac{\sum_{S_1 \cup S_2} {X_j}^2}{N-n_3} - \left(\frac{N\bar{X} - n_3\bar{X}_3}{N-n_3}\right)^2\right] + o(\epsilon) \end{aligned}$$

Now, given $S_3 \in G(S_3, k, \epsilon))$, straightforward calculations show that

$$\left(\frac{N\bar{X} - n_3\bar{X}_3}{N-n_3}\right)^2 = \bar{X}^2 + o(1) + O(\epsilon)$$

and,

$$\frac{\sum_{S_1 \cup S_2} {X_j}^2}{N-n_3} = \frac{\sum_{j=1}^{N} {X_j}^2}{N} + O(\epsilon) = 1 + o(1) + O(\epsilon).$$

Thus, given $S_3 \in G(S_3, k, \epsilon))$,

$$Var(\sqrt{N}\bar{X}_1|\bar{X}_3, S_3 \in G(S_3,k,\epsilon)) = \frac{\lambda_2}{(1-\lambda_3)\lambda_1}(1+o(1)) + O(\epsilon). \tag{3.2}$$

Thus,

$$E(e^{itT_{N'}}) = e^{it\sqrt{N}\left(\frac{a_2}{\lambda_2} + \frac{c_1}{1-\lambda_3}\right)\bar{X} - \frac{t^2}{2}\frac{\lambda_2 {c_1}^2}{\lambda_1(1-\lambda_3)} + o(1) + O(\epsilon)} E\left[e^{it\sqrt{N}(c_2 - \frac{c_1\lambda_3}{1-\lambda_3})\bar{X}_3} I(S_3 \in G(S_3,k,\epsilon))\right] + \xi$$

Now,

$$E(\bar{X}_3) = \bar{X}$$

and

$$Var(\sqrt{N}\bar{X}_3) = \frac{N-n_3}{N-1}\frac{N}{n_3}(1+o(1)) = \left(\frac{1}{\lambda_3} - 1\right)(1+o(1)) \to \sigma_{33}$$

as $N \to \infty$.

Therefore,

$$E(e^{itT_N{}'}) = e^{it\sqrt{N}\left(\frac{a_2}{\lambda_2}+\frac{c_1}{1-\lambda_3}+c_2-\frac{c_1\lambda_3}{1-\lambda_3}\right)\bar{X}-\frac{t^2}{2}\{\frac{\lambda_2 {c_1}^2}{\lambda_1(1-\lambda_3)}+(c_2-\frac{c_1\lambda_3}{1-\lambda_3})^2\sigma_{33}\}+o(1)+O(\epsilon)}+\xi$$

by using the Erdos-Renyi CLT once more.

Now, using the definition of c_1 and c_2, a straightforward simplification yields

$$\begin{aligned}\frac{a_2}{\lambda_2}+\frac{c_1}{1-\lambda_3}+c_2-\frac{c_1\lambda_3}{1-\lambda_3} &= c_1+c_2+\frac{a_2}{\lambda_2}\\ &= (a_1-\frac{a_2\lambda_1}{\lambda_2})+(a_3-\frac{a_2\lambda_3}{\lambda_2})+\frac{a_2}{\lambda_2}\\ &= a_1+a_3+\frac{1-\lambda_1-\lambda_3}{\lambda_2}a_2=\sum_{j=1}^{3}a_i.\end{aligned}$$

Also, using the definition of $\Sigma=((\sigma_{ij}))$, $1\le i,j\le 3$, we have

$$\begin{aligned}\frac{\lambda_3}{1-\lambda_3}\to\frac{\Delta_3}{1-\Delta_3}&=-\frac{\sigma_{13}}{\sigma_{33}}\\ \frac{\lambda_2}{\lambda_1(1-\lambda_3)}\to\frac{\Delta_2}{\Delta_1(1-\Delta_3)}&=\sigma_{11}-\frac{{\sigma_{13}}^2}{\sigma_{33}}\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\lambda_2{c_1}^2}{\lambda_1(1-\lambda_3)}+(c_2-\frac{c_1\lambda_3}{1-\lambda_3})^2\sigma_{33} &\sim {c_1}^2\left(\sigma_{11}-\frac{{\sigma_{13}}^2}{\sigma_{33}}\right)+\left(c_2+c_1\frac{\sigma_{13}}{\sigma_{33}}\right)^2\sigma_{33}\\ &= {c_1}^2\sigma_{11}+{c_2}^2\sigma_{33}+2c_1c_2\sigma_{13}\\ &= Var(\sqrt{N}(c_1\bar{X}_1+c_2\bar{X}_3))\\ &= Var(\sqrt{N}\sum_{i=1}^{3}a_i\bar{X}_i)\\ &= \mathbf{a}'\Sigma\mathbf{a}\end{aligned}$$

Hence,

$$E(e^{itT_N})\to e^{-\frac{t^2}{2}\mathbf{a}'\Sigma\mathbf{a}}$$

as $N\to\infty$, for all real vector $\mathbf{a}$.

Hence proved. □

The same result could possibly have been proved using the Wald-Wolfowitz-Noether CLT, for sampling without replacement from a finite population. For details see [Hajek, J 1961] and [Noether, G.E. 1949]. We will, however, need to establish some

extensions to the above result, and the Wald-Wolfowitz-Noether CLT does not lend itself easily to that purpose.

The above theorem can be immediately extended to more than three partitions as well as to the the multivariate case. Extension to four partitions from three would follow by applying an added level of conditioning on the fourth partition S_4 to the above proof. Extensions to greater number of partitions would follow analogously. The result for the multivariate case is similar and is presented below.

Theorem 3.5. *Let $\mathbf{X} = \{X_i\}_{\{i=1,\ldots,N\}}$ be an p-variate finite population of size N such that $Var(X_i) = \Gamma$ is positive-definite. Let $\{S_1, S_2, S_3\}$ be a random disjoint partition of $\mathbf{X}$, such that the size of S_i is n_i, $i = 1, 2, 3$; $\sum_{i=1}^{3} n_i = N$. Let the mean of the partition S_i be $\bar{X}_i = \{\bar{X}_{i1}, \ldots, \bar{X}_{ip}\}$, $i = 1, 2, 3$. Also assume $\frac{n_i}{N} = \lambda_i \to \Delta_i \in (0, 1)$, $i = 1, 2, 3$. Then we have the Central Limit Theorem:*

$$\sqrt{N}\left((\bar{X}_1 - \bar{X})', (\bar{X}_2 - \bar{X})', (\bar{X}_3 - \bar{X})'\right)' \Rightarrow N(\mathbf{0}, \Sigma \otimes \Gamma),\ a.s[\mathbf{X}],$$

where $\Sigma = ((\sigma_{i,j}))$ is given by $\sigma_{i,i} = (\frac{1}{\Delta_i} - 1)$ and $\sigma_{i,j} = -1$, $i, j = 1, 2, 3, i \neq j$, and the symbol $\otimes$ indicate Kronecker product.

Proof. The proof will follow exactly on the lines of the corresponding proof for the univariate case. Hence, only the outlines for the critical steps will be indicated. For simplicity of presentation, we will assume $p = 2$.

First, note that without loss of generality, we can assume $\Gamma = I$.

Also, note that $G(S_3, k, \epsilon)$ will be redefined analogously as:

$$G(S_3, k, \epsilon) = \{S_3 \subset \mathbf{X} : |\bar{X}_{3j} - \bar{X}_j| < \epsilon,\ |\bar{X^2}_{3j} - \bar{X^2}_j| < \epsilon,\ |\bar{X^{(k)}}_{3j} - \bar{X^{(k)}}_j| < \epsilon, \forall\ j = 1, 2,$$
$$\&\ |\frac{1}{n_3}\sum_{S_3} X_{j1}X_{j2} - \frac{1}{N}\sum_{j=1}^{N} X_{j1}X_{j2}| < \epsilon\},$$

and, as before, $P[G(S_3, k, \epsilon)] \to 1$ as $N \to \infty$ for all k, ϵ.

Now, since $\Gamma = I$, given $S_3 \in G(S_3, k, \epsilon)$,

$$\frac{\sum_{S_1 \cup S_2} X_{j1}X_{j2}}{N - n_3} = \frac{\sum_{j=1}^{N} X_{j1}X_{j2}}{N} + O(\epsilon) = O(\epsilon).$$

Therefore, if $Z = c_1\bar{X_{11}} + c_2\bar{X_{12}}$ for some constants c_1, c_2, given $S_3 \in G(S_3, k, \epsilon)$,

$$\sup_{S_3} \frac{\sum_{S_1 \cup S_2} Z^2 I(Z > k)}{\sum_{S_1 \cup S_2} Z^2} = \frac{\frac{1-\lambda_3}{N}\sum_{j=1}^{N} Z^2 I(Z > k) + O(\epsilon)}{\frac{1-\lambda_3}{N}\sum_{j=1}^{N} Z^2 + O(\epsilon)}$$

can be made arbitrarily small by choosing k to be large enough.

So, the Erdos-Renyi CLT can be applied to claim:

$$E\left[e^{it\sqrt{N}Z} | S_3 \in G(S_3, k, \epsilon)\right] = e^{\sqrt{N}itE(Z|S_3 \in G(S_3,k,\epsilon)) - \frac{t^2}{2}Var(\sqrt{N}Z|S_3 \in G(S_3,k,\epsilon))}$$

Also, 3.2 can be revised to:

$$Var(\sqrt{N}Z|S_3 \in G(S_3, k, \epsilon)) \to \frac{\lambda_2}{(1-\lambda_3)\lambda_1}\left({c_1}^2 + {c_2}^2 + o(1)\right) + O(\epsilon). \tag{3.3}$$

The rest of the proof follows by imitating the proof of Theorem 3.4. □

As in the univariate case, this result can also be extended to $k > 3$ many partitions.

The above results lead us to the theorem:

Theorem 3.6. *Under the setup of Theorem* 3.5 *and k partitions, we have*

$$\left(\sqrt{n_1}(\bar{X}_1 - \bar{X})', \sqrt{n_2}(\bar{X}_2 - \bar{X})', \ldots, \sqrt{n_k}(\bar{X}_k - \bar{X})'\right)' \Rightarrow N(\mathbf{0}, \Sigma_1 \otimes \Gamma),\ a.s[\mathbf{X}],$$

where $\Sigma_1 = ((\sigma_{i,j}))$ *is given by* $\sigma_{i,i} = 1 - \Delta_i$ *and* $\sigma_{i,j} = -\Delta_i\Delta_j$, $i, j = 1, 2, 3, i \neq j$, *and the symbol* $\otimes$ *indicate Kronecker product.*

Proof. The proof is immediate from Theorem 3.5 and the fact that $\frac{n_1}{N} \to \Delta_i$ for $i = 1, \ldots, k$. □

Using standard terminology from MANOVA, we define the "Total Sum of Squares" (T), the "Between Group Sum of Squares" (B) and the "Within Group Sum of Squares" (W) as follows:

$$\begin{aligned}
T &= SS(\mathbf{X}) \\
B &= \sum_{j=1}^{k} n_j(\overline{X^{(j)}} - \bar{X})(\overline{X^{(j)}} - \bar{X})' \\
W &= \sum_{j=1}^{k}\sum_{i=1}^{n_j} ({X_i}^{(j)} - \overline{X^{(j)}})({X_i}^{(j)} - \overline{X^{(j)}})'
\end{aligned}$$

where $\mathbf{X} = \{\mathbf{X^{(1)}} : \mathbf{X^{(2)}} : \ldots : \mathbf{X^{(k)}}\}$.

The same would be defined on the permuted data $\mathbf{X}^* = \{\mathbf{X^{(1)}}^* : \mathbf{X^{(2)}}^* : \ldots : \mathbf{X^{(k)}}^*\}$ as:

$$B^* = \sum_{j=1}^{k} n_j(\overline{X^{(j)*}} - \bar{X})(\overline{X^{(j)*}} - \bar{X})'$$
$$W^* = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_i{}^{(j)*} - \overline{X^{(j)*}})(X_i{}^{(j)*} - \overline{X^{(j)*}})'$$

It is well known that $T = W^* + B^*$. Since $\frac{1}{N}T = \frac{1}{N}(W^* + B^*) \to \Gamma$ a.s. $[\mathbf{X}]$, W^* can also be expressed as:

$$W^* = \sum_{j=1}^{k} n_j[\frac{T}{n} - (\overline{X^{(j)*}} - \bar{X})(\overline{X^{(j)*}} - \bar{X})']$$
$$\to \sum_{j=1}^{k} n_j[\Gamma - (\overline{X^{(j)*}} - \bar{X})(\overline{X^{(j)*}} - \bar{X})']$$

Hence, the asymptotic distribution of both W^* and B^* are dictated by the asymptotic joint distribution of the vector $\{\sqrt{n_i}(\overline{X^{(j)*}} - \bar{X}), i = 1, \ldots, k\}$.

Now, under the standard MANOVA assumption that $\mathbf{X}$ is a data matrix created from i.i.d. observations from a multivariate Normal distribution, ie, when $\mathbf{X^{(j)}}_i \sim N_p(0, \Gamma), \forall i = 1, \ldots, n_j; j = 1, \ldots, k$, both W and B follow Wishart distributions, and the ratio $\frac{|W|}{|T|}$ follows the Wilks Lambda distribution:

$$\begin{aligned} B &\sim Wishart(\Gamma, k-1) \\ W &\sim Wishart(\Gamma, n-k) \\ \frac{|W|}{|T|} &\sim \Lambda(p, N-k, k-1) \end{aligned} \tag{3.4}$$

Applying Theorem 3.6 to our permutation scheme, we can claim that the asymptotic joint distribution of $\{\sqrt{n_i}(\overline{X^{(j)*}} - \bar{X}), i = 1, \ldots, k\}$ under the permutation scheme is exactly the same as seen under the above mentioned MANOVA setup, i.e., when $\mathbf{X}$ is a data matrix of i.i.d. multivariate Normal observations. As a result, we can claim that the asymptotic permutation distribution of W^*, B^* and $\frac{|W^*|}{|T^*|}$ will be exactly the same as corresponding terms in 3.4.

This fact leads us to the most important result of the section.

Theorem 3.7. *Under the assumptions:*

(i) $Var(\mathbf{X}) = \Gamma$ *is p.d.*

(ii) $\frac{n_i}{N} \to \Delta_i \in (0,1)\ \forall i = 1, \ldots, k,$

the permutation test for testing $H_0 : \mu_1 = \ldots = \mu_k$ *vs.* $H_1 : \mu_i \neq \mu_j$ *for some* $i \neq j$*, at* $d(1)$ *is Pitman equivalent to the Wilks Lambda test.*

Proof. First, note that T is invariant under the permutation scheme. Hence, for the permutation test, comparing $d(1)$ with $d^*(1)$, is equivalent to comparing $\frac{d(1)}{|T|}$ with $\frac{d^*(1)}{|T|}$.

Now, since $\frac{d(1)}{|T|}$ is precisely the Wilks Lambda statistic, by Corollary 3.4 and the discussion following it, the asymptotic permutation distribution of $\frac{d^*(1)}{|T|} \to \frac{d^*(1)}{|N\Gamma|}$, under H_0, is the same as the Wilks Lambda distribution $\Lambda(p, N-k, k-1)$. Thus, at $100\alpha\%$ level of significance, the rejection region of the test, given by: $d(1) <$ lower $100\alpha\%$ of $d^*(1)$, is true iff $\frac{d(1)}{|T|} <$ lower $100\alpha\%$ of $\Lambda(p, N-k, k-1)$ distribution.

For local alternatives, $H_1 : \mu = \frac{c_i}{\sqrt{n_i}}, c_i \in \Re,\ 1 \leq i \leq k$, note that the CLT in Theorem 3.6 still holds, and as a result, the asymptotic distribution of B under the permutation scheme remains the same as under the standard Normal data-matrix setup in MANOVA. Hence the asymptotic distribution of $\frac{d^*(1)}{|T|} = \frac{|W^*|}{|T^*|} = \frac{|T-B^*|}{|T|} \to |I - \frac{\Gamma^{-1}B^*}{n}|$ is the same as in the Normal data-matrix case.

This proves the result. □

The above result explains why the Monte-Carlo estimates of the power of the test at $d(1)$ came out to be close to that of the Wilks Lambda test. One should however note a crucial difference between the two cases. Under normal data-matrix setup, any increase in $|T|$ under the alternative is fueled by an increase in $|B|$, while $|W|$ remains unchanged. On the other hand, under the permutation scheme, it is $|W|$ that gets inflated under the alternative, which causes $|T|$ to be larger. This difference, however, does to affect the asymptotic Pitman equivalence of the two tests.

Chapter 4

Applications to Linear Regression

Regression analysis is a statistical methodology, that tries to establish a relationship between two or more quantitative variables, so that the response or the outcome can be predicted using the remaining variables. It is one of the oldest disciplines within statistics, with the earliest research dating back to the nineteenth century with the works of Gauss ([Gauss, C.F. 1809]) and Legendre ([Legendre, A.M. 1809]), who together contributed to the initial developments of the theory of least squares. Linear regression models provide the simplest regression setups, where the response is modeled as a linear combination of the predictor variables.

The appeal of these methods lie in their conceptual simplicity, while retaining applicability. With time, linear regression has come to be one of the most widely used statistical tools for multifactor data - its application ranging from business administration and economics to social, health and biological sciences.

The theory available today is very rich, with significant developments being done in Bayesian, non-parametric and robust regression techniques. Literature on the subject is abundant, a few examples being [Kutner, Nachtsheim, Neter 1987], [Rao, C.R. 1973], [Draper, N.R., Smith H. 1966], [Hardle, W. 1990], [Ryan, T.P. 1997], [Fox, John 1997] and [Stapleton, J.H. 1995].

We describe below the simple linear regression model, and go on to introduce testing procedures based on the Determinant Scale curve which can be applied to the model.

4.1 The Linear Regression Model

Suppose we have a dataset consisting of the response variable $\mathbf{V}' = \{V_1, V_2, \ldots, V_n\}$ and explanatory variables $\{U_1, U_2, \ldots, U_n\}$ where $U_i' = \{U_{i1}, U_{i2}, \ldots, U_{ip}\}, 1 \leq i \leq n$

are $p \times 1$ i.i.d. p-variate row vectors.

Let $\mathbf{U}$ denote the $p \times n$ data-matrix, whose i^{th} column is $U_i, 1 \leq i \leq n$. Note here that $\mathbf{U}$ is non-random.

In this chapter, we will be dealing with the the multiple linear regression model, given by:

$$\mathbf{V} = \beta_{int}\mathbf{1_n} + \mathbf{U}'\beta + \epsilon \tag{4.1}$$

where $\beta = \{\beta_1, \beta_2, \ldots, \beta_p\}$ and β_{int} are the regression coefficients and ϵ is the "error" vector, such that $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2 G$ where G is a known non-singular matrix, and $\epsilon_i \perp \epsilon_j, 1 \leq i \neq j \leq n$. Without loss of generality, we may assume $G = I$.

Note that the parameter of interest here is β.

We define the variables as $Y_i = V_i - \bar{V}$ and $X_{ij} = U_{ij} - \bar{U_j}, 1 \leq i \leq n, 1 \leq j \leq p$, so that $\mathbf{Y}$ and X_i's are now centered.

As in previous chapters, $\mathbf{X}$ will denote the $p \times n$ data-matrix, whose i^{th} column is $X_i, 1 \leq i \leq n$. However, unlike previous chapters, $\mathbf{X}$ here will be considered to be non-random.

The most important hypothesis that one would typically be interested in testing for the above model is $H_0 : \beta = \beta_{\mathbf{0}}$ vs. $H_1 : \beta \neq \beta_{\mathbf{0}}$, where $\beta_{\mathbf{0}}$ is known. In the next section, we will investigate ways to apply testing procedures based on determinant scale curves to this problem.

4.2 Tests for β

For testing $H_0 : \beta = \beta_{\mathbf{0}}$ vs. $H_1 : \beta \neq \beta_{\mathbf{0}}$, we first note that without loss of generality, we may assume $\beta_{\mathbf{0}} = \mathbf{0}$, since, otherwise, we may redefine $\mathbf{V}$ as $\mathbf{V} = \mathbf{V} - \mathbf{U}'\beta_{\mathbf{0}}$.

To obtain the test, we shall employ a permutation scheme on the centered variables as follows:

Define $d(.)$ as the *dsc* for the combined dataset $(\mathbf{V} : \mathbf{U}')$. By definition, $d(.)$ is also the *dsc* of $(\mathbf{Y} : \mathbf{X}')$. Now, we randomly permute the vector $\mathbf{V}$ to create $\mathbf{V}^*$. Define $d^*(.)$ as the *dsc* of the matrix $(\mathbf{V}^* : \mathbf{U}')$, which again is same as the *dsc* of $(\mathbf{Y}^* : \mathbf{X}')$, where $\mathbf{Y}^* = \mathbf{V}^* - \bar{V}^*$.

Under the null hypothesis, we have $\mathbf{V} = \beta_{int}\mathbf{1_n} + \epsilon$, so that $d()$ is the *dsc* of $(\epsilon : \mathbf{X}')$ and $d^*()$ is the *dsc* with of $(\epsilon^* : \mathbf{X}')$, where ϵ^* is the permutation of ϵ. Since ϵ is a vector of i.i.d. "errors", this permutation will not change the $d()$ appreciably, and so, $d()$ and $d^*()$ will remain close to each other.

Now, under any point in the alternative, say $\beta = \beta_{\mathbf{1}}$, $d()$ = the *dsc* of $(\mathbf{Y} : \mathbf{X}') = dsc$ of $(\epsilon + \mathbf{X}'\beta_{\mathbf{1}} : \mathbf{X}) = dsc$ of $(\epsilon : \mathbf{X}')$ will be the same as under the null. However, under the permutation, $d^*()$ =*dsc* of $\left(\epsilon^* + \mathbf{X}^{*'}\beta_{\mathbf{1}} : \mathbf{X}'\right)$ will tend have larger values. One way to see it is to note that while computing $SS\left(\epsilon^* + \mathbf{X}^{*'}\beta_{\mathbf{1}} : \mathbf{X}'\right)$, the permutation effectively inflates the diagonal entry corresponding to the variance of ϵ^* by the sample variance of $\mathbf{X}^{*'}\beta_{\mathbf{1}}$, while the off-diagonal entries do not change significantly, since the covariances between $\epsilon^*, \mathbf{X}^{*'}\beta_{\mathbf{1}}$ and the columns of $\mathbf{X}'$ remain small. Hence, $d^*()$ will tend to lie above $d()$.

A more geometric way of visualizing the same effect is to note the fact that $|SS(\epsilon : \mathbf{X}')|$ is a measure of the volume of the data. Under null, ϵ is distributed evenly around 0, for the entire range of the X's. Hence, the permutation does not cause any appreciable change to the volume of the scatter of $(\epsilon : \mathbf{X}')$. However, under any alternative: $\beta = \beta_{\mathbf{1}}$, $d^*()$ essentially measures the volume of $\left(\epsilon^* + \mathbf{X}^{*'}\beta_{\mathbf{1}} : \mathbf{X}'\right)$. $\epsilon^* + \mathbf{X}^{*'}\beta_{\mathbf{1}}$ is no longer centered around the same value across the domain of $\mathbf{X}'$, and the permutation destroys the relationship between $\epsilon + \mathbf{X}'\beta_{\mathbf{1}}$ and the X's, thus inflating the volume of the scatter.

Equivalently, one might say that under the null, the data scatter of $\{\mathbf{X}, \mathbf{Y}\}$ approximately lies in a lower dimensional subspace of $\Re^{p+1}$, and thus, the volume, as computed using $|SS(\epsilon : \mathbf{X}')|$ remains low. The permutation process does not inflate this volume, because of the reasons described in the previous paragraph. However, under the alternative, the permutation disturbs the linear relationship between $\mathbf{X}$ and $\mathbf{Y}$, so that the scatter of $\{\mathbf{X}^*, \mathbf{Y}\}$ no longer lies in a lower dimensional subspace. This inflation of volume shows up in the permuted scale curves being higher than the scale curve of the un-permuted data.

This phenomenon can be utilized to devise a permutation test at any given level α. To perform the test, first we create a large number of replicates of $d^*()$. Now, we

reject H_0 if $d()$ lies in or under the bottom $100\alpha\%$ of the $d^*()$'s. Note that, like in previous chapters, this procedure leads to a distinct test for every $0 \leq t \leq 1$. However, as before, we would restrict ourselves to the specific values $t = 0.5, 0.75.0.9, 1$ as they would suffice in demonstrating the properties and usefulness of the tests.

Figures 4.1(a) and 4.1(b) provide illustration of the tests under the null and alternative hypotheses.

Observations and results in the subsequent sections will show that the above test for $t = 1$ is Pitman equivalent to the parametric F-test, which is the most powerful test for the above hypotheses, under (approx.) Normality assumptions on the error (ϵ) terms. Also, the *dsc* tests for $0.5 \leq t < 1$ will be shown to have significant robustness properties, being resilient to the presence of outliers. Hence, the tests will prove to be extremely useful in cases where there are outliers in the data, or when the distribution of ϵ is believed to be more diffused than the Normal distribution.

Before exploring these theoretical properties of the tests, we present some power simulations of the tests.

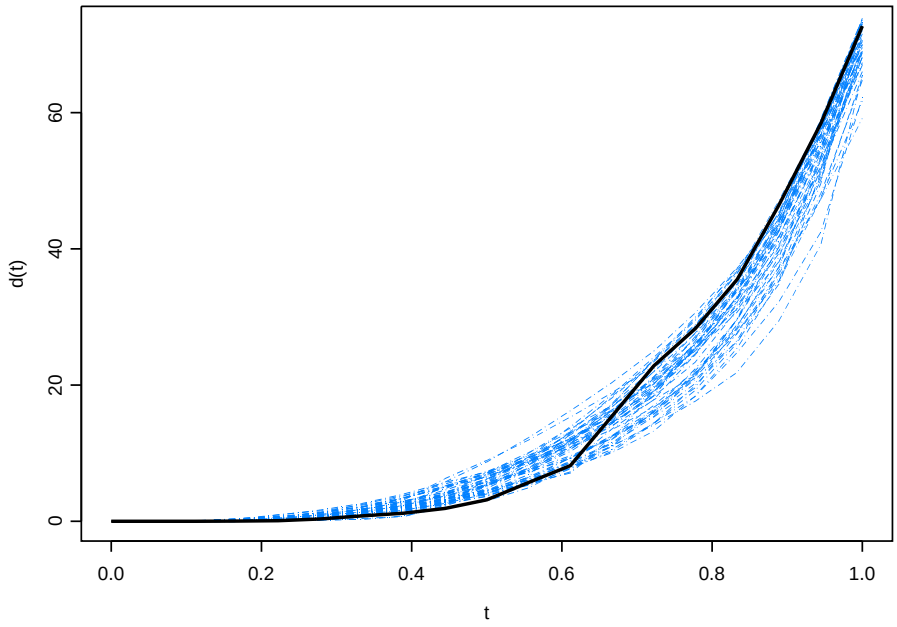
4.3 Power Simulations

We simulated the power of the *dsc* tests using a sample of size 20, on the model in equation 4.1, using

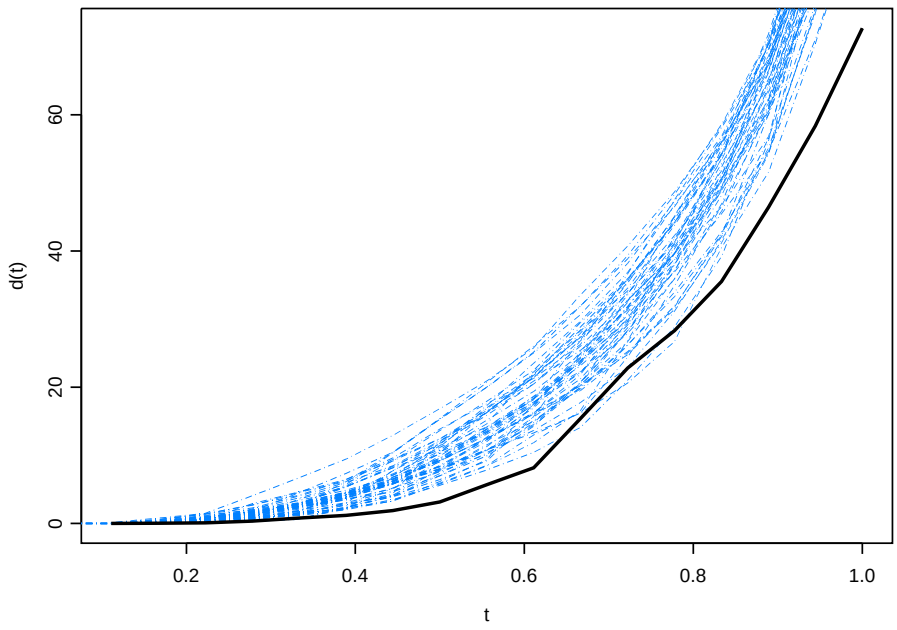
$$\mathbf{U_{1}}_i = i$$

$$\mathbf{U_{2}}_i = 1 + log(i),\ 1 \leq i \leq 20$$

Table 4.1 shows the powers compared to the F-test, using Standard Normal, Double Exponential and Cauchy as the error distributions. The column "F-test" shows the power of the F-test, using cut-off points corresponding to the Normal distribution. For Double Exponential and Cauchy distributions, these numbers are expected to be slightly inflated because of the thicker tails compared to Normal distribution. The fact becomes evident from the simulated figures corresponding to $\beta = (0, 0)$ in tables 4.1(b) and 4.1(c), where, the numbers for both Double Exponential and Cauchy are much higher than the expected number 0.05. We therefore calculated the 95^{th} percentile of



(a) Under $H_0 : \beta = (\mathbf{0}, \mathbf{0})$.



(b) Under $H_1 : \beta = (\mathbf{1}, \mathbf{1})$.

Figure 4.1: The *dsc* tests illustrated using a bivariate $N(\beta, I)$ sample of size 20.

(β_1, β_2)	d(.5)	d(.75)	d(.9)	d(1)	F-test
(0,0)	0.034	0.048	0.059	0.053	0.058
(0.0,0.4)	0.076	0.117	0.157	0.203	0.206
(0.0,0.6)	0.115	0.183	0.259	0.386	0.416
(0.0,0.8)	0.147	0.286	0.448	0.611	0.674
(0.0,1.0)	0.224	0.472	0.710	0.836	0.858
(0.2,0.0)	0.507	0.856	0.964	0.992	0.992
(0.2,0.2)	0.601	0.935	0.989	0.997	0.998
(0.2,0.4)	0.718	0.972	0.998	0.999	1.000

(a) Bivariate Normal

(β_1, β_2)	d(.5)	d(.75)	d(.9)	d(1)	F-test	Adj. F-test
(0,0)	0.047	0.053	0.052	0.053	0.074	0.044
(0.2,0.0)	0.371	0.557	0.710	0.831	0.888	0.870
(0.2,0.2)	0.445	0.675	0.794	0.901	0.932	0.912
(0.2,0.4)	0.511	0.769	0.890	0.955	0.968	0.956
(0.4,0.0)	0.879	0.983	0.998	1.000	0.998	1.000
(0.4,0.4)	0.913	0.993	1.000	0.999	1.000	1.000
(0.0,0.2)	0.066	0.062	0.068	0.070	0.080	0.076
(0.0,0.4)	0.047	0.063	0.070	0.114	0.132	0.126
(0.0,0.8)	0.129	0.179	0.243	0.347	0.424	0.368
(0.0,1.0)	0.158	0.260	0.374	0.551	0.568	0.530

(b) Bivariate Double Exponential

(β_1, β_2)	d(.5)	d(.75)	d(.9)	d(1)	F-test	Adj. F-test
(0,0)	0.052	0.051	0.058	0.053	0.076	0.060
(0.2,0.2)	0.212	0.215	0.203	0.196	0.212	0.218
(0.4,0.4)	0.544	0.625	0.584	0.449	0.520	0.440
(0.8,0.8)	0.874	0.909	0.869	0.704	0.744	0.678
(1.0,1.0)	0.940	0.959	0.921	0.757	0.754	0.730
(0.0,0.4)	0.046	0.061	0.050	0.066	0.060	0.062
(0.0,0.8)	0.073	0.092	0.081	0.089	0.106	0.070
(0.0,1.0)	0.100	0.105	0.092	0.096	0.108	0.102
(0.4,0.0)	0.460	0.533	0.479	0.400	0.446	0.394
(0.8,0.0)	0.828	0.882	0.829	0.659	0.726	0.656
(1.0,0.0)	0.923	0.952	0.917	0.740	0.768	0.744

(c) Bivariate Cauchy

Table 4.1: Power comparison of *dsc* tests with F-test using samples of size 10 each and $\alpha = 0.05$

the F-statistic under these two distributions using 5000 Monte-Carlo simulations, before using them to recalculate the power figures. These revised numbers are shown under the column "Adj. F-test", and are the ones that should be compared to those of the permutation test.

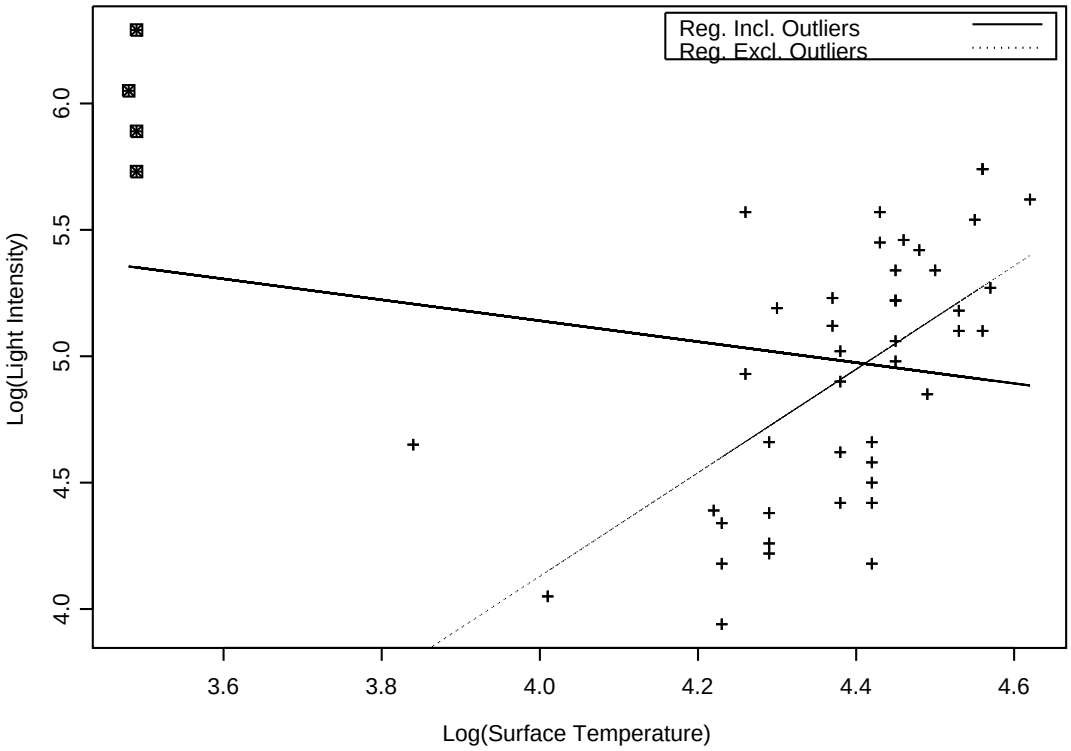


Figure 4.2: Scatter plot and fitted lines for CYG OB1 data.

The simulations show the powers for the test at $d(1)$ to be comparable to the F-test, for all the choices of the error distributions. Also, the test at $d(0.5)$ shows significant improvement over the F-test for the Cauchy case, reinforcing the fact that the tests in the middle zone of the *dsc*'s are significantly robust.

4.4 Example

[Humphreys, R.M. 1978] reported the light intensities and surface temperatures for the star cluster CYG OB1. The modified dataset, with the variables in logarithmic scale, appeared in [Rousseeuw, Leroy 1987]. The dataset is interesting in the fact that there seems to be a significant positive correlation between the variables. However, there are four outliers: data-points corresponding to four giant stars, that do not conform to the characteristics of the rest of the cluster. A scatter-plot of the data, along with the outliers, are shown in Figure 4.2.

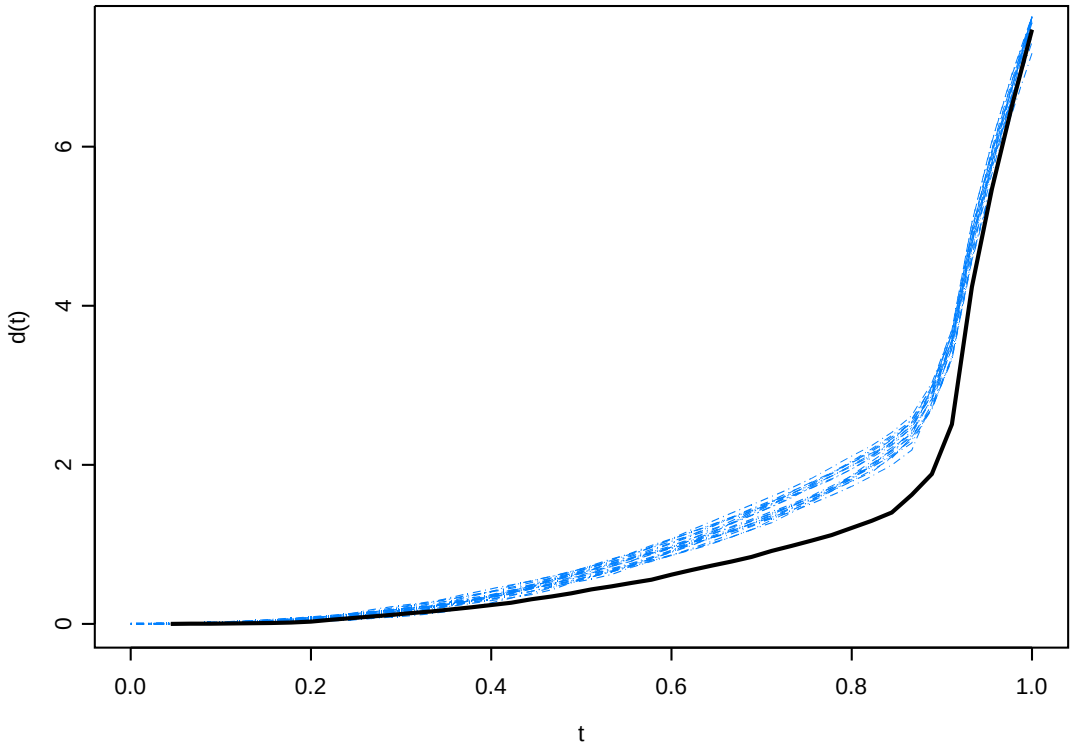


Figure 4.3: *dsc* tests on CYG OB1 data.

Interest lies in exploring the effect of the surface temperature on the light intensity, and a linear model can be applied on the logarithm of the variables to analyze the

effect. However, presence of these four outliers skews the analysis significantly. In fact, with the outliers included, the p-value of the F-test for $H_0 : \beta = \mathbf{0}$ vs. $H_1 : \beta \neq \mathbf{0}$ turns out to be 0.15, failing to reject the null hypothesis at all reasonable levels. However, the same analysis done with the outliers omitted, yields a p-value 0, thus rejecting the null at all level of significance. The solid line in Figure 4.2 shows the fitted regression line on the entire data, while the dotted line is the fitted line with the outliers omitted.

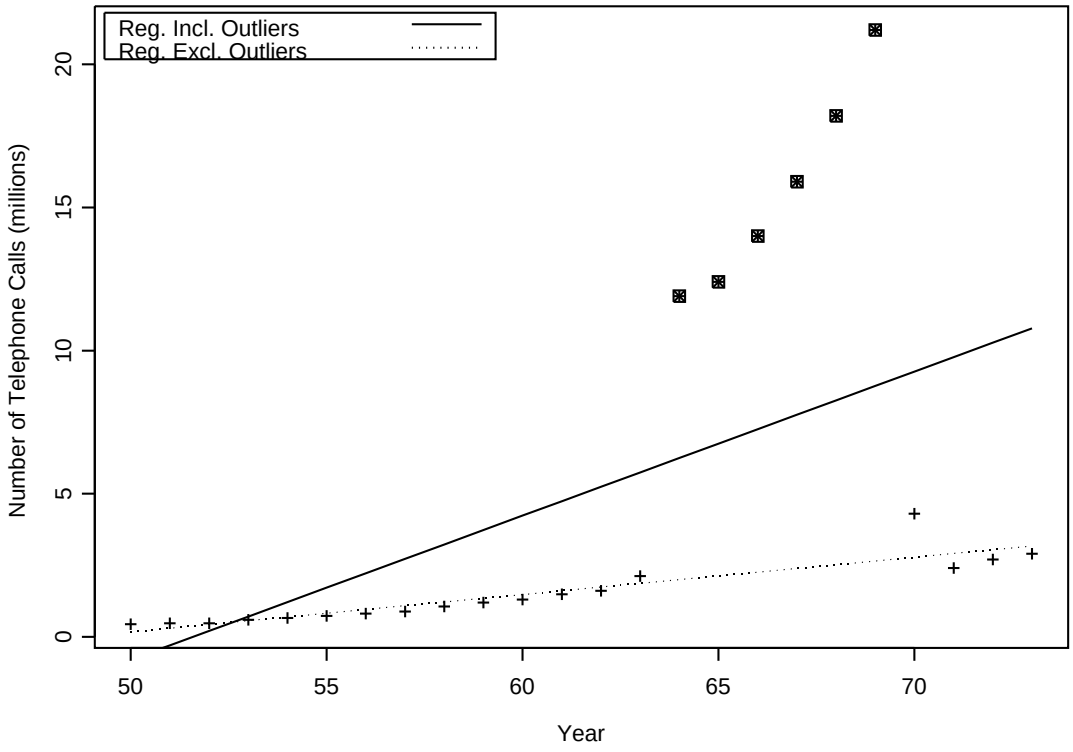


Figure 4.4: Scatter plot and fitted lines for Belgian telephone calls data.

Applying the test using *dsc* on the dataset, illustrated in Figure 4.3, shows that the null is rejected in the entire middle zone of the *dsc*. This once again demonstrates the robustness of the tests. However, at $t = 1$, the null hypothesis could not be rejected, showing the similarity of the test using $d(1)$ with the F-test.

As a second example, we look at the data set provided by the Belgian Statistical

Survey, describing the number of international phone calls from Belgium in years 1950 − 1973. The data has been analyzed by [Rousseeuw, Leroy 1987]. The scatter plot shown in Figure 4.4 shows clear outliers corresponding to the years 1964 − 1969 , when a different measurement system was used and instead of the number of phone calls, the total number of minutes of these calls were reported. The linear regression fit with the outliers included has a slope 0.5, and is heavily affected by the outliers, as shown in Figure 4.4. The regression fit without the outliers however, has a much lower slope, and seems to fit well with the rest of the data.

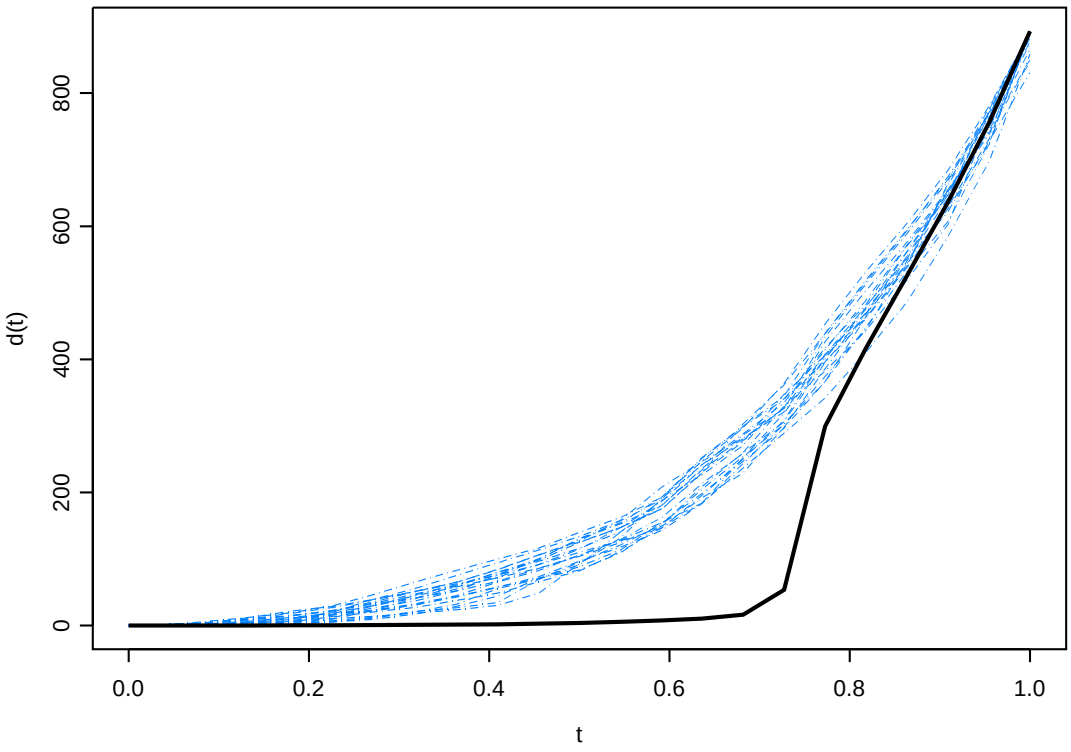


Figure 4.5: *dsc* tests on Belgian telephone calls data.

This distorting effect of the outliers is shown in hypothesis tests as well. The standard F-test for testing $H_0 : \beta = 0.5$ vs $H_1 : \beta \neq 0.5$, has a p-value of 0.98 in presence of the outliers, while the same test has p-value 0 when the outliers are

removed.

The *dsc* tests capture the effect of the outliers very nicely. Figure 4.5 shows the *dsc* test for the above hypotheses. The test at $t = 1$ cannot reject H_0, just like the F-test. The tests in the middle region of the *dsc* are, however, much more resistant to the outliers, and hence, reject the null.

We now go on to some results that would explain the similarity between the test using $d(1)$ and the F-test.

4.5 Optimality of *dsc* test at $t = 1$

We begin the section with a lemma.

Lemma 4.1. *For the regression setup in equation* 4.1*, the least squares estimate* $\hat{\beta}_{\mathbf{LS}}$ *of* β *is given by:*

$$\hat{\beta}_{\mathbf{LS}} = (\mathbf{XX'})^{-1}\mathbf{XY}$$

Proof. The least square estimates are obtained as the solution to the normal equations:

$$\begin{pmatrix} \mathbf{1_n}' \\ \mathbf{U} \end{pmatrix} \begin{pmatrix} \mathbf{1_n U'} \end{pmatrix} \begin{pmatrix} \beta_{int} \\ \beta \end{pmatrix} = \begin{pmatrix} \mathbf{1_n}' \\ \mathbf{U} \end{pmatrix} \mathbf{V} \tag{4.2}$$

Define $\bar{\mathbf{U}}' = \{\bar{U}_1, \bar{U}_2, \ldots, \bar{U}_p\}$. Define $\mathbf{L} = \begin{pmatrix} -\bar{\mathbf{U}} : \mathbf{I_p} \end{pmatrix}$, where $\mathbf{I_p}$ is the identity matrix of order p. Now, note that

$$\mathbf{L} \begin{pmatrix} \mathbf{1_n}' \\ \mathbf{U} \end{pmatrix} = \mathbf{X}$$

Thus,

$$\mathbf{L} \begin{pmatrix} \mathbf{1_n}' \\ \mathbf{U} \end{pmatrix} \begin{pmatrix} \mathbf{1_n} : \mathbf{U'} \end{pmatrix} = \mathbf{X} \begin{pmatrix} \mathbf{1_n} : \mathbf{U'} \end{pmatrix} = \begin{pmatrix} \mathbf{0_p} : \mathbf{XX'} \end{pmatrix}$$

and

$$\mathbf{L} \begin{pmatrix} \mathbf{1_n}' \\ \mathbf{U} \end{pmatrix} \mathbf{V} = \mathbf{XV} = \mathbf{XY}$$

Therefore, pre-multiplying both sides of 4.2 by $\mathbf{L}$, we get

$$\mathbf{XX'}\beta = \mathbf{XY}$$

which yields the solution $\hat{\beta}_{\mathbf{LS}} = (\mathbf{XX'})^{-1}\mathbf{XY}$. Hence proved. □

The F-statistic used to test $H_0 : \beta = \beta_{\mathbf{0}}$ vs. $H_1 : \beta \neq \beta_{\mathbf{0}}$, is given by

$$F^*{}_\beta = \frac{(\hat{\beta}_{\mathbf{LS}} - \beta_{\mathbf{0}})'(\mathbf{XX}')^{-1}(\hat{\beta}_{\mathbf{LS}} - \beta_{\mathbf{0}})}{pMSE} \tag{4.3}$$

where $MSE = \frac{||\mathbf{Y} - \hat{\beta}_{\mathbf{int}}\mathbf{1_n} - \mathbf{X}'\hat{\beta}_{\mathbf{LS}}||^2}{n-p-1}$, $\hat{\beta}_{int}$ being the least squares estimate of β_{int}. The rejection region at level α is given by: $F^*{}_\beta > 100(1-\alpha)$ percentile of $F_{p,n-p-1}$ distribution.

Note here that $MSE = \frac{||\mathbf{Y} - \hat{\beta}_{\mathbf{int}}\mathbf{1_n} - \mathbf{X}'\hat{\beta}_{\mathbf{LS}}||^2}{n-p-1} \to \sigma^2$ a.s. $[\epsilon]$.

Hence, $F^*{}_\beta \simeq \frac{1}{p\sigma^2}(\hat{\beta}_{\mathbf{LS}} - \beta_{\mathbf{0}})'(\mathbf{XX}')^{-1}(\hat{\beta}_{\mathbf{LS}} - \beta_{\mathbf{0}})$ for large n.

Therefore the test is asymptotically equivalent the test that rejects H_0 at level α when $\frac{1}{\sigma^2}(\hat{\beta}_{\mathbf{LS}} - \beta_{\mathbf{0}})'(\mathbf{XX}')^{-1}(\hat{\beta}_{\mathbf{LS}} - \beta_{\mathbf{0}}) > 100(1-\alpha)$ percentile of $\chi^2{}_p$ distribution. This important fact guides us to the most important result of this section.

Define $\mathbf{e} = \mathbf{Y} - \mathbf{X}'\beta_{\mathbf{0}}$. Notice here that since $\mathbf{Y}$ and $\mathbf{X}$ are centered, so is $\mathbf{e}$.

Now, observe that

$$\begin{aligned} d(1) &= |SS(\mathbf{Y} : \mathbf{X}')| \\ &= |SS(\mathbf{Y} - \mathbf{X}'\beta_{\mathbf{0}} : \mathbf{X}')| \\ &= \left| \begin{pmatrix} \mathbf{e}'\mathbf{e} & \mathbf{e}'\mathbf{X}' \\ \mathbf{Xe} & \mathbf{XX}' \end{pmatrix} \right| \\ &= |\mathbf{XX}'| \left[\mathbf{e}'\mathbf{e} - \mathbf{e}'\mathbf{X}'(\mathbf{XX}')^{-\mathbf{1}}\mathbf{Xe}\right] \end{aligned}$$

Similarly, $d^*(1) = |\mathbf{XX}'| \left[\mathbf{e}^{*'}\mathbf{e}^* - \mathbf{e}^{*'}\mathbf{X}'(\mathbf{XX}')^{-\mathbf{1}}\mathbf{Xe}^*\right]$, where $\mathbf{e}^* = \mathbf{Y}^* - \mathbf{X}^{*'}\beta_{\mathbf{0}}$.

Since, $\mathbf{e}^{*'}\mathbf{e}^* = \mathbf{e}'\mathbf{e}$, therefore $d(1)$ will be in or under the lowest $\alpha\%$ of $d^*(1)$'s, if and only if $\frac{1}{\sigma^2}\mathbf{e}'\mathbf{X}'(\mathbf{XX}')^{-\mathbf{1}}\mathbf{Xe} \geq 100(1-\alpha)$th percentile of the permutation distribution of $\frac{1}{\sigma^2}\mathbf{e}^{*'}\mathbf{X}'(\mathbf{XX}')^{-\mathbf{1}}\mathbf{Xe}^*$.

Now, a straightforward simplification yields

$$\begin{aligned} \frac{1}{\sigma^2}\mathbf{e}'\mathbf{X}'(\mathbf{XX}')^{-\mathbf{1}}\mathbf{Xe} &= \frac{1}{\sigma^2}(\mathbf{Y} - \mathbf{X}'\beta_{\mathbf{0}})'\mathbf{X}'(\mathbf{XX}')^{-\mathbf{1}}\mathbf{X}(\mathbf{Y} - \mathbf{X}'\beta_{\mathbf{0}}) \\ &= \frac{1}{\sigma^2}((\mathbf{XX}')^{-\mathbf{1}}\mathbf{X}'\mathbf{Y} - \beta_{\mathbf{0}})'(\mathbf{XX}')((\mathbf{XX}')^{-\mathbf{1}}\mathbf{X}'\mathbf{Y} - \beta_{\mathbf{0}}) \\ &= \frac{1}{\sigma^2}(\hat{\beta}_{\mathbf{LS}} - \beta_{\mathbf{0}})'(\mathbf{XX}')(\hat{\beta}_{\mathbf{LS}} - \beta_{\mathbf{0}}) \end{aligned} \tag{4.4}$$

for large n.

To proceed further, we would need to establish the asymptotic permutation distribution of $\frac{1}{\sigma^2}\mathbf{e}^{*'}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\mathbf{e}^*$. But before tackling the most general case, we will begin with the simpler case of $p = 1$.

Lemma 4.2. *Using the notation above, for $p = 1$, and under the assumptions:*

(i) $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} {X_{i1}}^r \to O(1)$, $r = 3, 4, 5, \ldots$,

(ii) $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} {X_{i1}}^2 = \gamma^2 > 0$,

(iii) $\max_{\{1\le i\le n\}} \frac{{\epsilon_i}^2}{n} \to 0$ *a.s. as* $n \to \infty$,

(iv) $\beta - \beta_0 = O_p(n^{-\frac{1}{2}})$,

the asymptotic permutation distribution of $\frac{1}{\sigma^2}\mathbf{e}^{*'}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\mathbf{e}^*$ *is* ${\chi^2}_1$ *a.s. [X].*

Note that condition (iii) above may be replaced by the slightly stronger condition: $(iiia) E({\epsilon_i}^2) < \infty$. Proof of the fact that $(iiia)$ implies (iii) is available in literature (see for example [Singh, Xie 2003], Appendix A).

Proof. Recall, without loss of generality, $\beta_{\mathbf{0}} = 0$. Hence, $\mathbf{e}^* = \mathbf{Y}^* = \mathbf{X}^{*'}\beta + \epsilon^*$.

Also, since $p = 1$, $\mathbf{X} = \{X_{11}, X_{21}, \ldots, X_{n,1}\}$. For simplicity of notation, we'll denote $X = \{X_1, X_2, \ldots, X_n\} = \{X_{11}, X_{21}, \ldots, X_{n,1}\}$.

Thus,

$$\begin{aligned}\mathbf{e}^{*'}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\mathbf{e}^* &= \mathbf{Y}^{*'}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\mathbf{Y}^* \\ &= \left(\frac{\sum_{i=1}^{n} X_i Y_i^*}{\sqrt{\sum_{i=1}^{n} {X_i}^2}}\right)^2 \\ &= \left(\sum_{i=1}^{n} a_i Y_i^*\right)^2\end{aligned}$$

where $a_i = \frac{X_i}{\sqrt{\sum_{i=1}^{n} {X_i}^2}}$, $i = 1, 2, \ldots, n$. Thus, it suffices to show that the asymptotic permutation distribution of $\sum_{i=1}^{n} a_i Y_i^*$ is $N(0, \sigma^2)$ a.s. $[X]$.

By assumptions (i) and (ii), the a_i's trivially satisfy the Wald-Wolfowitz condition, since

$$\frac{\frac{1}{n}\sum_{i=1}^{n} {a_i}^r}{\left(\frac{1}{n}\sum_{i=1}^{n} {a_i}^2\right)^{r/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} {X_i}^r}{\left(\frac{1}{n}\sum_{i=1}^{n} {X_i}^2\right)^{r/2}} = O(1),\ r = 3, 4, 5, \ldots.$$

Also, the Y_i's satisfy the Noether condition:

$$\begin{aligned}\max_{\{1\le i\le n\}} \frac{{Y_i}^2}{n} &= \max_{\{1\le i\le n\}} \frac{(\epsilon_i + X_i\beta)^2}{n} \\ &\le \max_{\{1\le i\le n\}} \frac{{\epsilon_i}^2}{n} + \beta^2 \max_{\{1\le i\le n\}} \frac{{X_i}^2}{n} + 2\beta \max_{\{1\le i\le n\}} \frac{\epsilon_i X_i}{n} \\ &\to 0\end{aligned}$$

as $n \to \infty$, since by assumption (iii), $\max_{\{1\le i\le n\}} \frac{{\epsilon_i}^2}{n} \to 0$; by assumption (ii) and (iv),

$$\beta^2 \max_{\{1\le i\le n\}} \frac{{X_i}^2}{n} \le O_p(n^{-1}) \sum_{i=1}^{n} \frac{{X_i}^2}{n} \to 0$$

and

$$\beta \max_{\{1\le i\le n\}} \frac{\epsilon_i X_i}{n} \le O_p(n^{-\frac{1}{2}}) \sum_{i=1}^{n} \frac{|\epsilon_i X_i|}{n} \le O_p(n^{-\frac{1}{2}}) \left(\sum_{i=1}^{n} \frac{{\epsilon_i}^2}{n}\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} \frac{{X_i}^2}{n}\right)^{\frac{1}{2}} \to 0.$$

Hence, using the Wald-Wolfowitz-Noether CLT for sampling without replacement from a finite population (for details see [Hajek, J 1961] and [Noether, G.E. 1949]), we can claim the asymptotic convergence in distibution a.s.$[X]$:

$$\sum_{i=1}^{n} a_i {Y_i}^* \Rightarrow N(\theta, \tau^2)$$

where

$$\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (a_i - \bar{a}) \sum_{i=1}^{n} ({Y_i}^* - \bar{Y}) = 0$$

since $\bar{X}$ is 0; and

$$\begin{aligned}\tau^2 &= \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} (a_i - \bar{a})^2 \sum_{i=1}^{n} ({Y_i}^* - \bar{Y})^2 \\ &= \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} ({\epsilon_i}^* - \bar{\epsilon} + \beta {X_i}^*)^2 \\ &= \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} ({\epsilon_i}^* - \bar{\epsilon})^2 + \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} (\beta {X_i}^*)^2 + \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} \beta {X_i}^* ({\epsilon_i}^* - \bar{\epsilon}) \\ &= \sigma^2 + \beta^2\gamma^2\end{aligned}$$

since

$$\lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} \beta {X_i}^* ({\epsilon_i}^* - \bar{\epsilon}) = \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} \beta X_i (\epsilon_i - \bar{\epsilon}) = 0$$

by SLLN a.s. $[X]$.

Now, by assumption (iv), $\beta = O_p(n^{-\frac{1}{2}})$. Therefore

$$\tau^2 = \sigma^2 + \beta^2\gamma^2 = \sigma^2 + \gamma^2 O_p(n^{-1}) \to \sigma^2$$

Thus, the asymptotic permutation distribution of $\sum_{i=1}^{n} a_i Y_i^*$ is $N(0, \sigma^2)$ a.s. $[X]$. Hence proved.

□

The general version of the above lemma, with $p \geq 1$ follows analogously.

Lemma 4.3. *Using the notation used above, and under the assumptions:*

(i) $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} {X_{ij}}^r = {\gamma_j}^{(r)}$, $j = 1, 2, \ldots, p$, $r = 3, 4, 5, \ldots$,

(ii) $\lim_{n\to\infty} \frac{1}{n}\mathbf{X}\mathbf{X}' = \Gamma$ *id p.d.,*

(iii) $\max_{\{1\leq i\leq n\}} \frac{{\epsilon_i}^2}{n} \to 0$ *a.s. as* $n \to \infty$,

(iv) $\beta - \beta_0 = O_p(n^{-\frac{1}{2}})$,

the asymptotic permutation distribution of $\frac{1}{\sigma^2}\mathbf{e}^{*'}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\mathbf{e}^*$ *is* ${\chi^2}_p$.

The proof is similar to that of Lemma 4.2, but for a little cumbersome notation, and is thus moved to the appendix. Just like Lemma 4.2, condition (iii) above may be replaced by the slightly stronger condition: $(iiia) E({\epsilon_i}^2) < \infty$.

Using the above lemma, we have the immediate theorem:

Theorem 4.4. *Under the assumptions of Lemma* 4.3*, the dsc test using* $d(1)$ *is Pitman equivalent to the F-test using the statistic in* 4.3*.*

Proof. The theorem follows from Lemma 4.3 and equation 4.4. □

The above theorem explains the similarity in the power of the *dsc* tests at $d(1)$ and the F-tests, as shown in Table 4.1.

4.6 Exploring Linearity in Multivariate datasets

The techniques developed in the previous sections can be readily extended to problems where one would wish to explore linearities in a multivariate data. Such situations occur regularly in multiple linear regression setups, where linear relationship between the covariates lead to problems like multicollinearity.

The standard parametric test and Likelihood Ratio test (LRT) for exploring linear relationships involve the Pearson's correlation coefficient, or its multivariate analog, and utilize its asymptotic distributional properties under minor distributional assumptions on the data.

One of the most popular statistics used to quantify the linearity in the data is the Variance Inflation Factor (VIF). The VIF of a given variable is essentially a monotone increasing function of the multiple correlation coefficient R^2 achieved under a linear regression setup with the (standardized) variable in question being the response and the remaining (standardized) variables being the covariates. The higher the VIF, the greater is the degree of linear dependence. It can be shown that the idea is closely related to the LRT. For detailed discussion on the matter, see [Muirhead, Robb J. 1982].

The above approach can be easily replicated using the techniques developed in the previous sections. The test to explore whether a variable is a linear function of another group of variables would simply require us to identify them as $\mathbf{V}$ and $\mathbf{U}$ respectively in Equation 4.1. Simple modifications of Lemma 4.3 and Theorem 4.3 will establish that the *dsc* test for the slope parameter being zero is Pitman equivalent to the LRT for $H_0 : R^2 = 0$ vs. $H_1 : R^2 \neq 0$.

We however would explore a more holistic approach, where the aim is to check the existence of (one or more) linear relationship(s) among a group of variables, without explicitly identifying the variables that are linearly related.

To that end, let $\{X_i\}_{\{i=1,\ldots,n\}}$ denote a multivariate sample of size n on $\Re^p$, and let $\mathbf{X}$ denote the $p \times n$ matrix with X_i's as its columns.

To motivate the methods, we refer to a well known result:

Lemma 4.5. *Let* $S = SS(\mathbf{X}) = ((S))_{i,j}, 1 \leq i, j \leq p$. *Then* $|S| \leq \prod_{i=1}^{p} s_{i,i},\ 1 \leq i \leq p$.

See [Rao, Bhimasankaram 1992] for a quick proof. The inequality above is strict when S is p.d.

4.6.1 Permutation Scheme and test

We shall modify the permutation scheme in Section 4.2 as follows:

Let $d()$ denote the *dsc* of $\mathbf{X}$. Now we permute all the rows of $\mathbf{X}$ randomly to create $\mathbf{X}^*$, and generate the *dsc* $d^*()$ of the permuted data. If there were any existing linear relationship within (any subset of) the data, this process will completely destroy it. As a result, the off-diagonal elements of $SS(\mathbf{X}^*)$ would be small in absolute value. Lemma 4.5 indicates that this would inflate *dsc* of the permuted data. Thus, $d^*()$ would tend to lie above $d()$.

In the absence of any linearity, this permutation will not have any appreciable effect and thus $d()$ and $d^*()$ tend to lie close to each other.

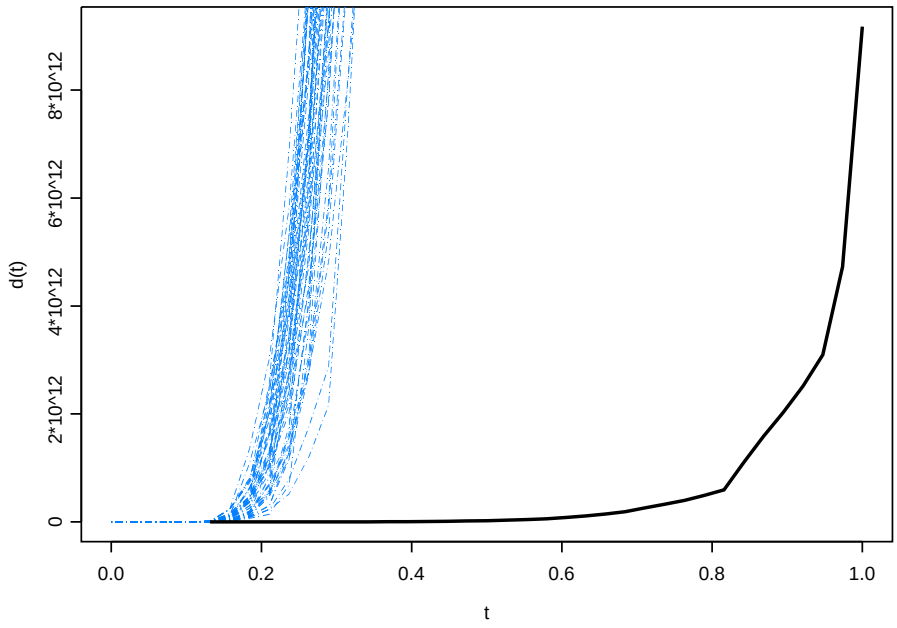
From a more geometric standpoint, one might say, that in presence of linearity, the data actually lies in a lower dimensional subspace, and thus has a smaller volume, as captured by $d()$. The permutation process essentially inflates the data to a higher dimension, causing $d^*()$ to be higher than $d()$.

This fact can be utilized to create permutation tests, similar to the ones we have described before, to test H_0 : *No Linearity* vs H_1 : *Linearity*. One can generate multiple copies of $d^*()$, using the permutation scheme, and create a band representing the distribution of the *dsc* of the permuted data. For any $0 < t \leq 1$, reject H_0 at level α if $d()$ lies in or under the bottom $100\alpha\%$ of the band of $d^*()$'s. As in all previous cases, this yields a distinct test for every $0 < t \leq 1$.

We will explore the performance of the test using an example.

4.6.2 Example

[Montgomery et al 2007] provides data from an experiment on jet turbines. The explanatory variables that are reported are respectively - X_1: primary speed of rotation,



(a) Using all variables

Figure 4.6: Tests for linearity using multisample *dsc* on Jet Turbine Data

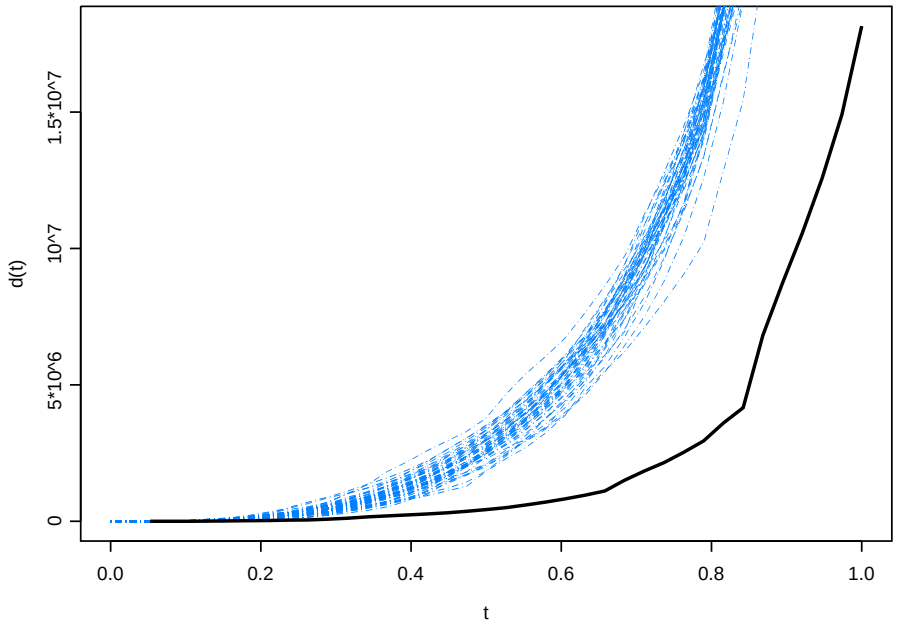
X_2: secondary speed of rotation, X_3: fuel flow rate, X_4: pressure, X_5: exhaust temperature and X_6: ambient temperature at time of test.

We use the *dsc* test to check whether there exists any linearity among these six variables.

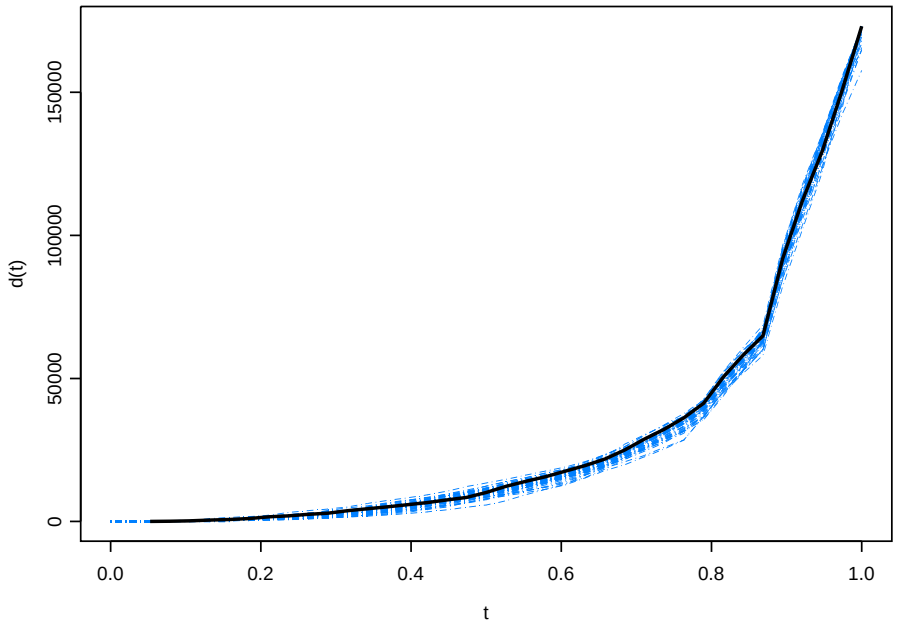
When all the six variables are considered, VIF calculations yield $VIF(X_1) = 289, VIF(X_3) = 168, VIF(X_4) = 220$, the rest being under 100. The high VIF's indicate strong linear relations involving these variables. The same is reflected in Figure 4.6(a), which emphatically rejects the hypothesis of non-linearity for all $0 < t \leq 1$.

Dropping X_1, X_3 and X_4 yields $VIF(X_2) = 15.3, VIF(X_5) = 16.8, VIF(X_6) = 2.3$, indicating that linearity still exists, albeit at a lesser degree. Figure 4.6(b) validates the fact.

Finally, dropping X_5 yields $VIF(X_2) = VIF(X_6) = 1$, indicating that there is no linear relationship between the two remaining variables. Figure 4.6(c) confirms it, with $d()$ lying entirely within the band of $d^*()$'s for all $0 < t \leq 1$.



(b) Using X_2, X_5 and X_6



(c) Using X_2 and X_6

Figure 4.6: Tests for linearity using multisample *dsc* on Jet Turbine Data (contd.)

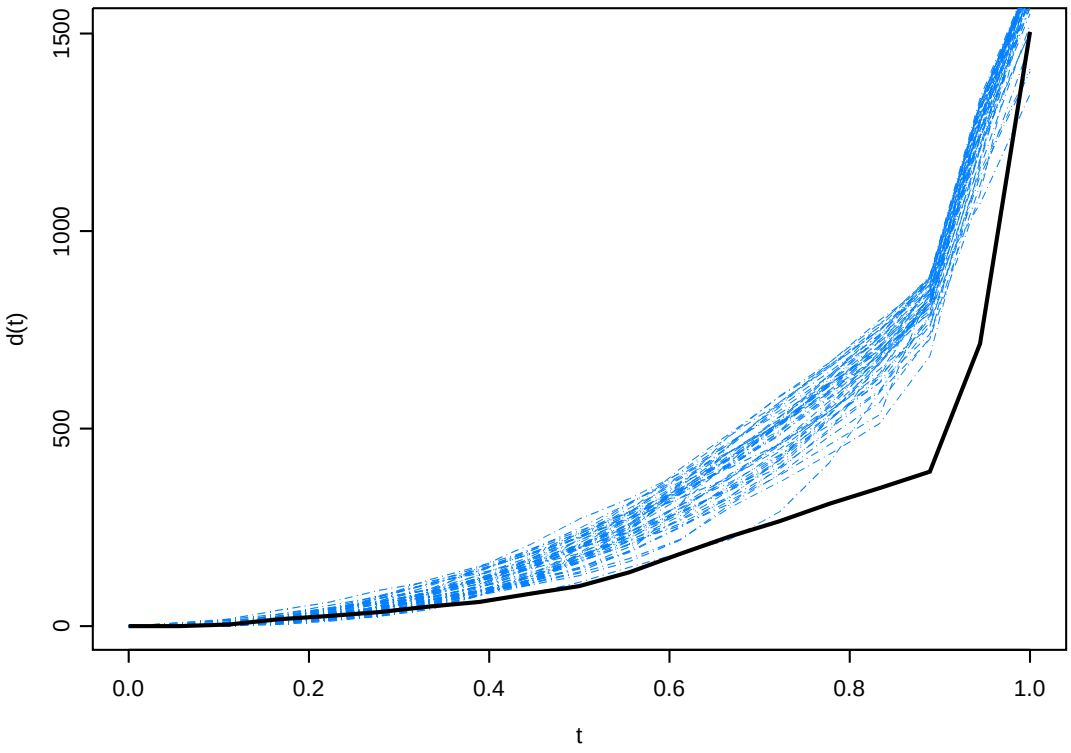


Figure 4.6: *dsc* tests on simulated Normal Data with outliers.

4.6.3 Robustness

Due to the inherent robustness of the *dsc* in its middle section, the permutation tests in these middle ranges are largely unaffected by presence of outliers. The same property is exhibited by the above test for existence of linearity.

A simple simulation exercise brings out the fact nicely, as is shown in Figure 4.6. The bivariate data used were simulated as:

$$X_{1_i} = i,\ 1 \leq i \leq 20$$

$$X_{2_i} = X_{1_i} + \epsilon_i,\ 2 \leq i \leq 19,$$

$$X_{2_1} = X_{2_{20}} = 20$$

where $\epsilon_i \sim N(0, 1)$ are independent.

By construction, X_1 and X_2 exhibit strong linear relationship, except for the points (X_{1_1}, X_{2_1}) and $(X_{1_{20}}, X_{2_{20}})$, which serve as outliers. In spite of the strong linearity, the two outliers bring down the VIF to 2.2. The same vulnerability to outliers is seen at the rightmost extreme of the *dsc*, and the test at $t = 1$ fails to reject the hypothesis of non-linearity. However, the tst performs much better in the middle sections, where $d()$ lies entirely below the band of $d^*()$'s, thus rejecting the hypothesis of non-existence of linearity.

Chapter 5

Appendix

5.1 Proof of Lemma 4.3

Proof. Recall, without loss of generality, $\beta_{\mathbf{0}} = 0$. Hence, $\mathbf{e}^* = \mathbf{Y}^* = \mathbf{X}^{*'}\beta + \epsilon^*$.

Therefore,

$$\begin{aligned} \mathbf{e}^{*'}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\mathbf{e}^* &= \mathbf{Y}^{*'}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\mathbf{Y}^* \\ &= ||(\mathbf{X}\mathbf{X}')^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{X}\mathbf{Y}^*||^2 \end{aligned} \tag{5.1}$$

Therefore, it suffices to prove that the asymptotic permutation distribution of $(\mathbf{X}\mathbf{X}')^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{X}\mathbf{Y}^*$ is $N_p(0, \sigma^2 I)$.

Let $l \in \Re^p$ and $||l||^2 = 1$. Will show that the asymptotic permutation distribution of $l'(\mathbf{X}\mathbf{X}')^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{X}\mathbf{Y}^*$ is $N(0, \sigma^2)$.

To that end, define $u' = l'(\mathbf{X}\mathbf{X}')^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{X}$. Note that $u'\mathbf{1_n} = 0$, since $\mathbf{X}\mathbf{1_n} = 0$, and $||u||^2 = l'(\mathbf{X}\mathbf{X}')^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{X}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-\frac{\mathbf{1}}{\mathbf{2}}}l = ||l||^2 = 1$.

Also, by assumption (ii), $u' \simeq \frac{1}{\sqrt{n}} l'\mathbf{\Gamma}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{X}$ as $n \to \infty$.

Denote, $v' = \frac{1}{\sqrt{n}} l'\mathbf{\Gamma}^{-\frac{\mathbf{1}}{\mathbf{2}}}$. Thus, for all $r = 3, 4, 5, ...,$ and large n, we have

$$\begin{aligned} \left(\sum_{i=1}^{n} {u_i}^r\right)^{\frac{1}{r}} &\simeq \left[\sum_{i=1}^{n}\left(\sum_{j=1}^{p} v_j X_{ij}\right)^r\right]^{\frac{1}{r}} \\ &\leq \sum_{j=1}^{p} v_j \left(\sum_{i=1}^{n} {X_{ij}}^r\right)^{\frac{1}{r}} \\ &\simeq \sum_{j=1}^{p} v_j \left(n\gamma_j^{(r)}\right)^{\frac{1}{r}} \\ &= n^{\frac{1}{r}-\frac{1}{2}} l'\mathbf{\Gamma}^{-\frac{\mathbf{1}}{\mathbf{2}}} (\gamma^{(\mathbf{r})})^{\frac{\mathbf{1}}{\mathbf{r}}} \end{aligned} \tag{5.2}$$

where $\gamma^{(\mathbf{r})'} = \{\gamma_1^{(r)}, \gamma_2^{(r)}, \ldots, \gamma_p^{(r)}\}$. The inequality above follows using Minkowski's inequality.

Therefore, for all $r = 3, 4, 5, ...,$

$$\lim_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n} {u_i}^r}{\left[\frac{1}{n}\sum_{i=1}^{n} {u_i}^2\right]^{\frac{r}{2}}} \leq \lim_{n\to\infty} n^{-1+\frac{r}{2}} \left(n^{\frac{1}{r}-\frac{1}{2}} l'\mathbf{\Gamma}^{-\frac{1}{2}} (\gamma^{(\mathbf{r})})^{\frac{1}{\mathbf{r}}}\right)^r = \left(l'\mathbf{\Gamma}^{-\frac{1}{2}} (\gamma^{(\mathbf{r})})^{\frac{1}{\mathbf{r}}}\right)^r = O_p(1) \tag{5.3}$$

and hence u satisfies the Wald-Wolfowitz condition.

Also, note that

$$\begin{aligned}\max_{\{1\leq i\leq n\}} \frac{{Y_i}^2}{n} &= \max_{\{1\leq i\leq n\}} \frac{(\epsilon_i + \sum_{j=1}^{p} X_{ij}\beta_j)^2}{n} \\ &\leq \max_{\{1\leq i\leq n\}} \frac{{\epsilon_i}^2}{n} + \max_{\{1\leq i\leq n\}} \frac{(\sum_{j=1}^{p} X_{ij}\beta_j)^2}{n} + 2\max_{\{1\leq i\leq n\}} \frac{\epsilon_i \sum_{j=1}^{p} X_{ij}\beta_j}{n}\end{aligned}$$

As $n \to \infty$, by assumption (iii), $\max_{\{1\leq i\leq n\}} \frac{{\epsilon_i}^2}{n} \to 0$.

By assumption (ii) and (iv),

$$\max_{\{1\leq i\leq n\}} \frac{(\sum_{j=1}^{p} X_{ij}\beta_j)^2}{n} \leq \sum_{i=1}^{n} \frac{(\sum_{j=1}^{p} X_{ij}\beta_j)^2}{n} \to \beta'\mathbf{\Gamma}\beta = o_p(1)$$

Also, by assumption (iv), can assume $\beta' = \frac{1}{\sqrt{n}}(c_1, c_2, \ldots, c_n), c_i \in \Re$

Thus,

$$\begin{aligned}\max_{\{1\leq i\leq n\}} \frac{\epsilon_i \sum_{j=1}^{p} X_{ij}\beta_j}{n} &\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{|\epsilon_i| \sum_{j=1}^{p} |X_{ij}c_j|}{n} \\ &\leq \left(\max_{\{1\leq i\leq n\}} \frac{|\epsilon_i|}{\sqrt{n}}\right) \left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{p} |X_{ij}c_j|^2}{n}\right)^{\frac{1}{2}} \\ &\to 0.\end{aligned}$$

Hence the Y_i's satisfy the Noether condition:

$$\max_{\{1\leq i\leq n\}} \frac{{Y_i}^2}{n} \to 0$$

Hence, using the Wald-Wolfowitz-Noether CLT for sampling without replacement from a finite population (for details see [Hajek, J 1961] and [Noether, G.E. 1949]), we can claim the asymptotic convergence in distibution a.s.$[X]$:

$$\sum_{i=1}^{n} u_i {Y_i}^* \Rightarrow N(\theta, \tau^2)$$

where

$$\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u}) \sum_{i=1}^{n} ({Y_i}^* - \bar{Y}) = 0$$

since $u'\mathbf{1_n} = 0$; and

$$\begin{aligned}
\tau^2 &= \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})^2 \sum_{i=1}^{n} (Y_i{}^* - \bar{Y})^2 \\
&= \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} (\epsilon_i{}^* - \bar{\epsilon} + \sum_{j=1}^{p} X_{ij}^* \beta_j)^2 \\
&= \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} (\epsilon_i{}^* - \bar{\epsilon})^2 + \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} (\sum_{j=1}^{p} X_{ij}^* \beta_j)^2 \\
&+ \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} (\epsilon_i{}^* - \bar{\epsilon}) \sum_{j=1}^{p} X_{ij}^* \beta_j \\
&= \sigma^2 + \beta' \Gamma \beta = \sigma^2 + o_p(1)
\end{aligned}$$

since

$$\lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} (\epsilon_i{}^* - \bar{\epsilon}) \sum_{j=1}^{p} X_{ij}^* \beta_j = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(\epsilon_i - \bar{\epsilon}) \sum_{j=1}^{p} X_{ij} c_j}{n-1} = 0$$

by SLLN a.s. $[X]$.

Thus, the asymptotic permutation distribution of $\sum_{i=1}^{n} u_i Y_i{}^*$ is $N(0, \sigma^2)$ a.s. $[X]$. Hence proved.

□

References

[Anderson, T.W. 1958] T.W.Anderson. An Introduction to Multivariate Statistical Analysis. *Wiley, New York*, 1958.

[Draper, N.R., Smith H. 1966] N.R. Draper, H. Smith. Applied regression analysis. *New York: Wiley*, 1966.

[Erdös, P. and Rényi, A. (1959)] P. Erdös, A.Rényi. On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungarian Acad. Sci.* **4**, 4961, 1959.

[Fisher,R.A. 1936] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics* **7,II**, 179-188, 1936.

[Fox, John 1997] . John Fox. Applied Regression Analysis, Linear Models, and Related Methods. *Thousand Oaks, CA: Sage*. 1997.

[Frets, G. P. 1921] G.P. Frets. Heredity of head form in man. *Genetica*, **3**, 193-384, 1921.

[Gauss, C.F. 1809] C.F. Gauss. Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum, 1809. English translation by C. H. Davis, *Dover, New York*, reprinted 1963.

[Gelman et al 1995] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin. Bayesian Data Analysis. *Chapman & Hall, New York*, 1995.

[Hajek, J 1961] Jaroslav Hajek. Some Extensions of the Wald-Wolfowitz-Noether Theorem. *Annals of Mathematical Statistics* **32, 2**, 506-523, 1961.

[Hardle, W. 1990] W. Hardle. Applied Nonparametric Regression. *Econometric Society Monograph Series 19, Cambridge University Press*, 1990.

[Hettmansperger et al 1994] T.P. Hettmansperger, H. Oja. Affine invariant multivariate multisample sign test. *Journal of the Royal Statistical Society* **B,56**, 235-249, 1994.

[Hettmansperger et al 1997] T. P. Hettmansperger, J. Möttönen and H. Oja. Multivariate affine invariant one-sample signed-rank tests. *Journal of the American Statistical Society* **92**, 1591-1600, 1997.

[Humphreys, R.M. 1978] R.M. Humphreys. Studies of luminous stars in nearby galaxies. I. Supergiants and O stars in the Milky Way *Astrophysics Journal, Supplementary Series* **38**, 309-350, 1978.

[Kutner, Nachtsheim, Neter 1987] M.H. Kutner, C.J. Nachtsheim, J. Neter. Applied Linear Regression Models. 4^{th} Edition *McGraw-Hill/Irwin*, 2004.

[Legendre, A.M. 1809] A.M. Legendre. Nouvelles méthodes pour la détermination des orbites des comètes. *Paris* 1805.

[Liu, Parelius, Singh 1999] R. Liu, J. Parelius and K.Singh. Multivariate analysis of the data-depth: descriptive statistics and inference. *Annals of Statistics* **27**, 783-858, 1999.

[Liu, R. 1990] R. Liu. On a notion of data depth based on random simplicies. *Annals of Statistics* **18**, 405-414, 1990.

[Mardia, Kent, Bibby 1979] K.V.Mardia, J.T.Kent, J.M.Bibby. Multivariate Analysis. *Academic Press, New York*, 1979.

[Montgomery, D.C. 1976] Douglas C. Montgomery. Design and Analysis of Experiments. *John Wiley & Sons, Inc, New York*, 1982.

[Montgomery et al 2007] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. Introduction to Linear Regression Analysis. *John Wiley & Sons, Inc, New York, 4th ed*, 2007.

[Muirhead, Robb J. 1982] Robb J. Muirhead. Aspects of Multivariate Statistical Theory. *Wiley, New York*, 1971.

[Noether, G.E. 1949] G.E. Noether. On a theorem by Wald and Wolfowitz. *Annals of Mathematical Statistics* **20**, 455-558, 1949.

[Puri, Sen 1971] Madan Lal Puri, Pranab Kumar Sen. Nonparametric Methods in Multivariate Analysis. *Wiley, New York*, 1971.

[Rao, Bhimasankaram 1992] A.R. Rao, P. Bhimasankaram. Linear Algebra *Tata McGraw-Hill, New Delhi*, 1992.

[Rao, C.R. 1973] C.R. Rao. Linear Statistical Inference and its Applications *Wiley, New York*, 1973.

[Reeve, E.C.R. 1941] E.C.R. Reeve. A Statistical analsys of taxonomic differences within the genus. *Tamandua* Gray (Xanartha). *Proc. Zool. Soc. Lond.* **A,111**, 279-302, 1941.

[Rousseeuw, Leroy 1987] P.J. Rousseeuw, A.M. Leroy. Robust Regression and Outlier Detection. *John Wiley and Sons, New York*, 1987.

[Ryan, T.P. 1997] Thomas P. Ryan. Modern Regression Methods. *Wiley, New York*, 1997.

[Serfling, R. 2002] R. Serfling. A depth function and a scale curve based on spatial quantiles. In , *Statistics and Data Analysis Based on L1-Norm and Related Methods, (Y. Dodge, ed.), Boston: Birkhäuser*, 25-38, 2002

[Singh, Tyler, Zhang, Mukherjee] K. Singh, D.E. Tyler, J. Zhang and S. Mukherjee. Quantile Scale Curves. *Juurnal of Computational and Graphical Statistics*

[Singh, Xie 2003] K. Singh and M. Xie. Bootlier-plot-Bootstrap based outlier detection plot. *Sankhya*, **65**, 532-559, 2003.

[Stapleton, J.H. 1995] James H. Stapleton. Linear Statistical Models. *John Wiley and Sons, New York*, 1995.

[Wilks, S.S. 1962] S.S. Wilks. Mathematical Statistics. *Wiley, New York*, 1962.

Vita

Somnath Mukherjee

2009 Ph. D. in Statistics, Rutgers University

2002 Master of Statistics, Indian Statistical Institute, Kolkata, India

2000 Bachelor of Statistics, Indian Statistical Institute, Kolkata, India

2002-04 Business Analyst, GE Capital Financial Services, India

2004-08 Teaching Assistant, Department of Statistics, Rutgers University, NJ, USA

2009- Part-Time Lecturer, Department of Statistics, Rutgers University, NJ, USA